Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

SLU University Library

SLU ID: SLU.ua.2018.2.3.2.IÄ-1

Management of data sets, software and raw data at the Swedish University of Agricultural Sciences – a survey on research data and environmental monitoring and assessment data

Karl Pettersson, Mikaela Asplund, Olof Frank, Hanna Lindroos

SLU-bibliotekets rapportserie, 9
Uppsala 2019

Management of data sets, software and raw data at the Swedish University of Agricultural Sciences – a survey on research data and environmental monitoring and assessment data

Karl Pettersson	Swedish University of Agricultural Sciences, Data Curation Unit, karl.pettersson@slu.se
Mikaela Asplund	Swedish University of Agricultural Sciences, Data Curation Unit, mikaela.asplund@slu.se
Olof Frank	Swedish University of Agricultural Sciences, Data Curation Unit, olof.frank@slu.se
Hanna Lindroos	Swedish University of Agricultural Sciences, Data Curation Unit, hanna.lindroos@slu.se

Place of publication:	Uppsala
Year of publication:	2019
Title of series:	SLU-bibliotekets rapportserie
Part number:	9
Online publication:	https://pub.epsilon.slu.se
Bibliographic reference:	Pettersson, Karl; Asplund, Mikaela; Frank, Olof; Lindroos, Hanna (2019). *Management of data sets, software and raw data at the Swedish University of Agricultural Sciences - a survey on research data and environmental monitoring and assessment data*. Uppsala: Sveriges lantbruksuniversitet. (SLU-bibliotekets rapportserie, 9).
Keywords:	research data, environmental analysis data, data management, archival, open science, open government data, survey

Swedish University of Agricultural Sciences
SLU University Library

Contents

Sammanfattning

En enkät skickades till forskare vid Sveriges lantbruksuniversitet (SLU) med frågor rörande deras hantering av forskningsdata, exempelvis vilka datatyper och mjukvaror de använder och hur stora datavolymer de hanterar, deras kunskaper om juridiska aspekter på datahantering och deras attityder till att göra data öppet tillgängliga.

I rapporten diskuteras svaren på enkätfrågorna och deras implikationer för framtida arbete med datakurering vid universitetet.

Nyckelord: forskningsdata, miljöanalysdata, datamängder, datahantering, arkivering, öppen tillgång, öppna myndighetsdata, enkät

Abstract

A survey was sent to researchers at the Swedish University of Agricultural Sciences (SLU) with questions concerning their research data management, e.g. data types and software products used and data volumes generated, their knowledge about legal aspects of data management, and their attitudes towards making data openly available. This report discusses the survey answers and their implications for future work with data curation at the university.

Keywords: research data, environmental data, data volumes, data management, archiving, open access, open government data, survey

About the report

Mikaela Asplund compiled the answers to questions 2b and 2c, and wrote part of the background section. Olof Frank compiled the answers to questions 2a and 2d. Tomas Lundén compiled the answers to question 2e. Karl Pettersson compiled the answers to questions 1 and 3 and put together the report in its entirety. Hanna Lindroos wrote part of the background information and part of the conclusions and discussion section.

1 Background

This report was put together by members of the Data Curation Unit (DCU) at the Swedish University of Agricultural Sciences (SLU). It is based on the answers to a survey that was conducted as part of the development of Tilda, an IT system for e-archiving and publishing of research and environmental monitoring and assessment (EMA) data. SLU has been commissioned by the Swedish government to conduct environmental assessment and has done that continuously for several years. The results from the EMA are primarily used as decision basis for the authorities' to reach the national and international objectives of long-term sustainable development. As a public authority and university, it is the responsibility of SLU to archive and to readily provide access to its research and EMA data. Subsequently, SLU handles large amounts of environmental data and statistics as well as research data.

For this purpose, SLU has a support organization for EMA; the Unit for Data Management Guidance and Development (DMGD), with a mandate to develop the systematic quality work, strengthen coordination and in the long-term work towards making SLU's environmental data quality assured and available.

One of the overall objectives in the SLU strategy 2017–2020, is that "SLU's researchers have good access to research infrastructure which gives the opportunity for ground-breaking, excellent research" (SLU 2016). Supporting systematic data management, thereby promoting quality assurance of research/EMA data, is a prime ambition for the university.

To ensure uniform and clear routines and methods for data management throughout the university, SLU is developing Tilda, an IT-system solution for e-archiving and publishing research and EMA data, both in raw and processed form. In Tilda the user will register metadata and upload research/EMA data via a web interface.

The data is manually curated by support staff from the DCU and subsequently archived and made publicly available via the same system. Tilda will serve as the focal point for long-term preservation and making SLU research and EMA data visible.

During the development project and in preparation for a launch of the Tilda system, questions regarding the types, size and format of data that data producers at SLU generate and manage remained. In addition it was important to determine the attitude of the data producers and the extent of their knowledge concerning archiving and publishing of (open) data in order to rightly assess the need for information to and education of the future users of Tilda.

The DCU put together a survey (appendix A) consisting of two main parts; one with questions regarding data types, sensitive data and archiving of data, and another part concerning the attitude towards making data openly available. The survey was sent via e-mail to 1441 data producers at SLU in November 2017. Approximately 20% of the recipients answered the questions, the exact number of responses are mentioned in the respective section of the report. A third, smaller part of the survey dealt with the affiliation and the position of the respondents and the answers are described in section four of this report.

2 Question 1

The answers to questions 1b and 1c have been normalized in order to be processed and visualized through a diagram.

2.1 Question 1a: Types of data

Question Which data types do you work with?

The recipients were given 16 different possible answers describing different types of data as well as the options "I do not work with digital data" and "other data types, please specify which" which included the possibility to provide a text answer. Multiple answers were possible.

278 of the recipients answered this questions, and the distribution of the answers are found in fig. 1.

The majority (63%) of the 27 text answers were related to nucleotide sequence data or similar and does not fit into any of the available categories. This is equivalent of 6% of the total number of answers.

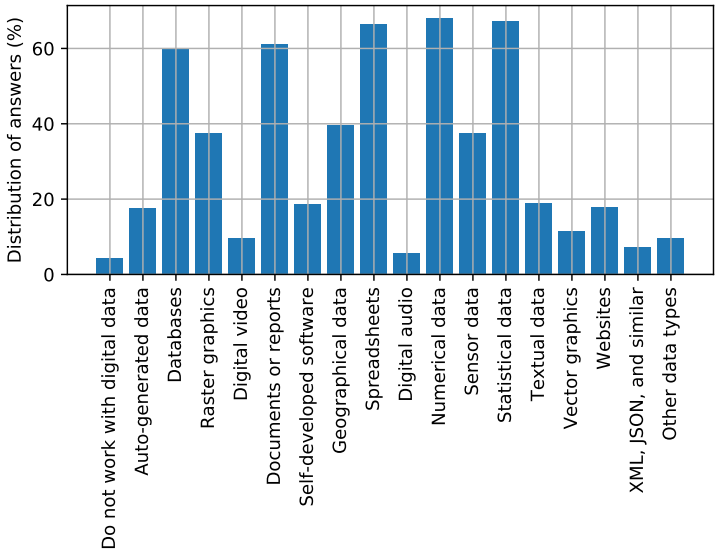

Figure 1: The distribution of answers to question 1a.

2.2 Question 1b: Software products

Question Which software products do you use when working with data?

262 recipients responded to this question, most of which have stated the use of more than one product.

After normalization of the names of the software products, we were left with a total of 194 products. A number of unspecific answers; e.g. different analysis programs, statistical programs and bioinformatics tools, were not included. Several software products have only been named by one or a few respondents, but 14 products have been mentioned by at least 10 data producers each and the distribution of these are illustrated in fig. 2. Tbl. 1 lists all 194 software products that have been mentioned in answer to question 1b. Short descriptions on the ability to manage open formats are provided for the 10 most common products from fig. 2.

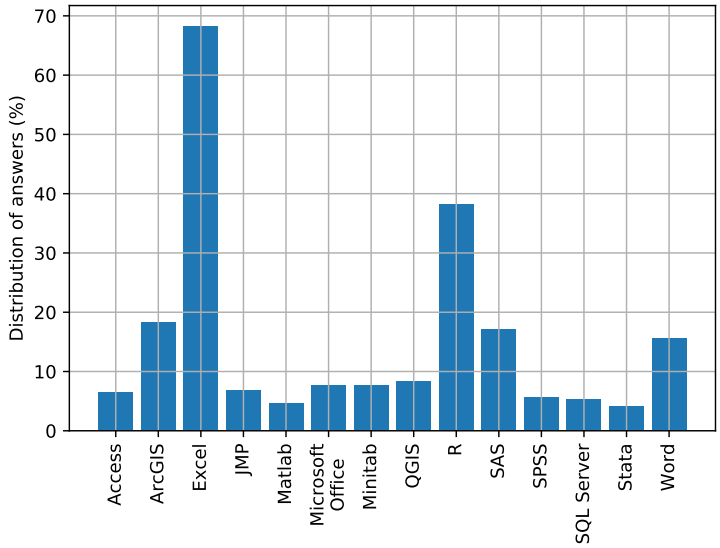

Figure 2: The proportion of respondents to question 1b that claim to use one of the software products indicated on the x-axis (only products with 10 or more answers).

Table 1: Reported software products.

Software product	Count	Relative frequency (%)	Comment
@risk	1	0	
Access	17	6	Proprietary relational database from Microsoft. Can export data to SQL or text formats.
Acrobat	4	2	
Agisoft photoscan	1	0	
Agisoft photoscan pro	1	0	
Aimms	1	0	
Alba	1	0	
Applications-master	1	0	
ArcGIS	48	18	Proprietary software for geographical information systems. Can manage data in open formats.
Arcview	2	1	
Ascii	1	0	
Asreml	2	1	
Athena	1	0	
Balancelink	1	0	

Software product	Count	Relative frequency (%)	Comment
Bash	1	0	
Bin	1	0	
Biorad	1	0	
Boris	1	0	
C	2	1	
C++	2	1	
Can-eye	1	0	
Canoco	5	2	
Ccpn	1	0	
Chemstation	1	0	
Clc genomic	1	0	
Clc genomic workbench	1	0	
Clc main	1	0	
Cloudcom-pare	1	0	
Codoncode	1	0	
Creative suite	1	0	
Ctd-sond	1	0	
Dat	1	0	
Datagraph	1	0	
Datem summit	2	1	
Delphi	1	0	
Diamond 2.1e	1	0	
Distance	1	0	
Djaview	1	0	
Dmu	1	0	
Dnastar	1	0	
Dropbox	1	0	
Ebi services	1	0	
Endnote	1	0	
Er60	1	0	
Erdas	1	0	
Erdas imagine	1	0	
Esris	1	0	
Eviews	1	0	
Exafspak	1	0	
Excel	179	68	Proprietary spreadsheet software from Microsoft. Uses an XML-based format with an open specification (XLSX) as standard format. Can also export to CSV and other delimited text files.
Fastq	1	0	
Filemaker	3	1	

Software product	Count	Relative frequency (%)	Comment
Fme	6	2	
Fortran	1	0	
Fugroviewer	1	0	
Fuji	1	0	
Fusion	3	1	
Gams	1	0	
Gap light analyzer	1	0	
Gel blot imaging	1	0	
Gnumeric	1	0	
Gnxas	1	0	
Google earth	2	1	
Google maps	1	0	
Gps plus	1	0	
Grapher	1	0	
Graphpad	1	0	
Grass	1	0	
Grassgis	1	0	
Gretl	1	0	
Gromacs	1	0	
Gsea	1	0	
Hadoop	1	0	
Heureka	2	1	
Heyex 2	1	0	
Hobo	1	0	
Horos	2	1	
Illumina	1	0	
Illustrator	7	3	
Imagej	5	2	
Iphoto	1	0	
Irfanview	1	0	
Isomap	1	0	
Jags	1	0	
Java	2	1	
JMP	18	7	Proprietary software for statistics. A scripting language (JSL) can be used to reproduce analysis where script is saved as text that can be read and edited without proprietary software. Data in the native JMP format can be imported into R. The software can also export data to text formats.
Kurvlr	1	0	
Labchart pro	1	0	
Lastools	1	0	
Latex	1	0	
Libreoffice	2	1	
Lims	1	0	

Software product	Count	Relative frequency (%)	Comment
Maestro	1	0	
Mathematica	1	0	
Matlab	12	5	Proprietary software for calculations. A script can be used to reproduce analysis, where script is saved as text and can be read and edited without proprietary software. The program can also export data to text format.
Mega	1	0	
Mendeley	1	0	
Mestrenova	1	0	
Microsoft Office	20	8	Several respondents have mentioned Microsoft Office without further specification of which program they use. We find it likely that most of the respondents who indicated Microsoft Office use Excel or Access, as these are the ones intended for data management.
Microstation	2	1	
Minitab	20	8	Proprietary software for statistics. A scripting language (MAC) can be used to reproduce analysis, where the script is saved as text and can be read and edited without proprietary software. The program can also export data to text format.
Modde	2	1	
Mysql	1	0	
N-logit	1	0	
Ncss	1	0	
Nfts	2	1	
Notepad	1	0	
Nudist	1	0	
Nvivo	8	3	
Opals	1	0	
Openbugs	1	0	
Openoffice	1	0	
Opus wire	1	0	
Origin	2	1	
Origin pro	1	0	
Osirix	1	0	
Pdf	1	0	
Perl	1	0	
Photoshop	6	2	
Plink	1	0	
Pls toolbox	1	0	
Postgres	1	0	
Powerpivot	2	1	
Powerpoint	3	1	
Premiere	1	0	
Prism	1	0	
Python	7	3	

Software product	Count	Relative frequency (%)	Comment
QGIS	22	8	Free software for geographical information systems. Manages data in open formats.
Qtmodeler	2	1	
Quick terrain modeler	1	0	
R	100	38	A programming language focused on statistics but also the main implementation of said language. Code is saved in text files. The program is free software and can export data to text formats.
Raspberry pi camera	1	0	
Rengis	1	0	
Rsem	1	0	
Rust	1	0	
SAS	45	17	Proprietary software for statistics. A scripting language can be used to reproduce analysis, where the script is saved as text and can be read and edited without proprietary software. Data in the native format can be imported to R. The program can export data to text formats.
Sharepoint	2	1	
Sigmaplot	3	1	
Simca	2	1	
Sonar5 pro	1	0	
SPSS	15	6	Proprietary software for statistics. Data in the native format can be imported to R. The program can export data to text formats.
SQL Server	14	5	Proprietary relational database from Microsoft. Can export data to SQL or text format.
Sqlite	2	1	
Stata	11	4	Proprietary software for statistics. A script can be used to reproduce analysis, where script is saved as text and can be read and edited without proprietary software. Data in the native format can be imported to R. The program can export data to text formats.
Statgraphics	1	0	
Statistica	2	1	
Statview	1	0	
Steplr	1	0	
Summit evolution	1	0	
Superbase	1	0	
Syntech	1	0	
Tapestation	1	0	
Terramod-eler	1	0	
Terrascan	2	1	
Topspin	1	0	
Trimble realworks	1	0	

Software product	Count	Relative frequency (%)	Comment
Trinity	1	0	
Ugene	1	0	
Unscrambler	2	1	
Unspecified	3	1	
Adobe (unspec.)	1	0	
Analysis (unspec.)	1	0	
Bioinfo (unspec.)	6	2	
Browser (unspec.)	2	1	
Dictaphone (unspec.)	1	0	
Dnaseq (unspec.)	2	1	
Gene (unspec.)	1	0	
Genomi-caligners (unspec.)	1	0	
Genomic-analysis (unspec.)	1	0	
Gis (unspec.)	3	1	
Graphics (unspec.)	1	0	
Instr (unspec.)	1	0	
Labinstr (unspec.)	2	1	
Microscopy (unspec.)	1	0	
Microsoft (unspec.)	2	1	
Office (unspec.)	1	0	
Public (unspec.)	1	0	
Self (unspec.)	3	1	
Sensor (unspec.)	1	0	
Spread (unspec.)	1	0	
Sql (unspec.)	5	2	

Software product	Count	Relative frequency (%)	Comment
Stat (unspec.)	5	2	
Unix (unspec.)	1	0	
Video (unspec.)	1	0	
Webapp (unspec.)	1	0	
Visualstudio	1	0	
Windows	1	0	
Windows document manager	1	0	
Windows media player	1	0	
Winstat	1	0	
Wizard	1	0	
Word	41	16	Proprietary word processor from Microsoft. Uses an XML-based format with an open specification (DOCX) as standard format. Can export text files. We expect that this is more commonly used for reports and documents than for storage or analysis of data.
Zen black	1	0	
Ztree	1	0	

2.3 Question 1c: Data formats

Question Do you save data in other formats than those that are standard in the software products specified in 1b? If so, please specify which.

Among the recipients that answered question 1b, 118 also answered this text question. Half of them answered "no". After normalization of the remaining answers we are left with 52 different data formats. The distribution of the data formats given by more than one respondent is illustrated by fig. 3. Tbl. 2 lists all 52 formats that have been mentioned in answer to question 1c. Short descriptions focussing on to what extent the formats mentioned in fig. 3 can be considered to be open formats are provided in the table.

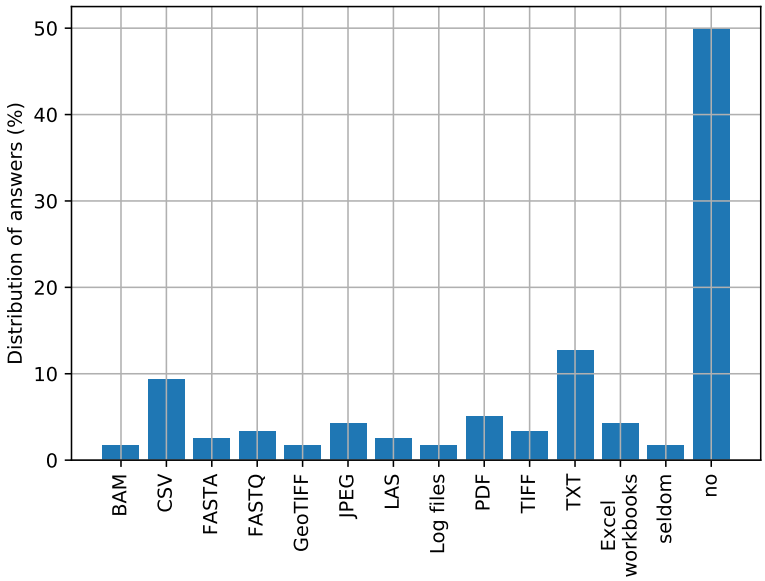

Figure 3: The distribution of the data formats named in question 1c given by more than one respondent.

Table 2: Formats named in question 1c with description of the formats mentioned in fig. 3.

Format	Count	Relative frequency (%)	Comment
BAM	2	2	A binary version of the SAM format, used by SAMtools, a suite of free tools for DNA sequencing.
BIN	1	1	
BMP	1	1	
CONFOCAL-IMAGE	1	1	
CSV	11	9	Comma-separated text files. An open format used by many applications. However, details such as decimal separator and quoting of string may vary between implementations.
CVS	1	1	
EDR	1	1	
FASTA	3	3	A text-based format for DNA and protein sequencing. Used by the free FASTA software package.
FASTQ	4	3	A text-based format used to store information about biological sequence along with its corresponding quality scores. Can be used with free software, such as FASTX-Toolkit.
FILECON-VERTER	1	1	
GENANALY-SIS	1	1	

Format	Count	Relative frequency (%)	Comment
GENSTAT	1	1	
GeoTIFF	2	2	A format for embedding geographical coordinates in TIFF files. Can be used with free software products.
GRO	1	1	
JOURNALS	1	1	
JPEG	5	4	An open format for compressed bitmap graphics.
LABBOOKS	1	1	
LARGE	1	1	
LAS	3	3	An open format for cloud data.
LaTeX	1	1	
Log files	2	2	Some respondents have mentioned log files. However, specific formats for log files may vary and have not been specified further. Formats for log files are often, but not always, text-based.
MAPINFO	1	1	
MDB	1	1	
MINITAB	1	1	
MP4	1	1	
MTS	1	1	
NETCDF	1	1	
new formats needed	1	1	
ODF	1	1	
OLAP CUBE	1	1	
PDF	6	5	The PDF format is probably most often used for things like scientific reports rather than pure research data.
PX	1	1	
R	1	1	
RAR	1	1	
RAWINSTRU-MENT	1	1	
RDS	1	1	
SAM	1	1	
SEQDATA	1	1	
SFF	1	1	
SHP	1	1	
spreadsheet	1	1	
SQL	1	1	
SQL SERVER	1	1	
TIFF	4	3	An open format for bitmap graphics.
TXT	15	13	Text files with digital data may adhere to many different conventions as regards such things as filed delimiters. In order to facilitate digital preservation, it is important that these conventions are adequately documented.
unspecified	1	1	

Format	Count	Relative frequency (%)	Comment
Excel workbooks	5	4	Native format in Microsoft Excel. There is an older, binary format (the `.XLS` extension) as well as newer XML-based format (the `.XLSX` extension). It has not been possible to differentiate between these from the answers.
XML	1	1	
XTC	1	1	
ZIP	1	1	
seldom	2	2	
no	59	50	

2.4 Question 1d: Volume of data

Question How large volumes of data do you need to publish and archive? Please select one or more of the following options.

231 answers were provided to this question and the distribution is shown in tbl. 3.

Table 3: Volumes of data for publication and archiving.

Interval	Count	Relative frequency (%)
1–500 MB	69	30
500 MB–1 GB	45	19
1–500 GB	74	32
500 GB–1 TB	23	10
More than 1 TB	43	19

39 text answers were provided in response to this question. Some of the respondents state a specific volume; i.e. “1–500 MB”, “approximately 4 TB” or “approximately 8 TB”. Others have remarked that they don’t know how much data they need to archive, some that the questions are not clearly defined, that they don’t know whether it is the total amount of data or the amount per project that is required or that they are unsure whether the question refers to raw data or working material. Especially noteworthy are the answers that state larger volumes; “many TB”, “>50 TB”, “approximately 80 TB”, “>100 TB” and “PB”.

2.5 Question 1e: Sensitive data

Question Do you work with sensitive data?

Five options were given; “I am not sure whether or not I work with sensitive data”, “No, I do not work with sensitive data”, “Yes, data which is regulated according to Swedish openness and secrecy law” (a link to the Public Access to Information and Secrecy Act was provided (SLU 2009)), “Yes, information about persons” or “Yes, other sensitive data. Please specify which kinds” where the data producers could provide a text answer. Multiple answers were possible.

To this question 270 answers were given, distributed as illustrated in fig. 4, and seen in tbl. 4.

32 text answers were provided by the data producers. In most cases, the respondents have specified the type of sensitive data that they manage; e.g. animal experiments or geographical data that can be connected to the land owner and location of protected species. One respondent was not sure what constitutes personal data.

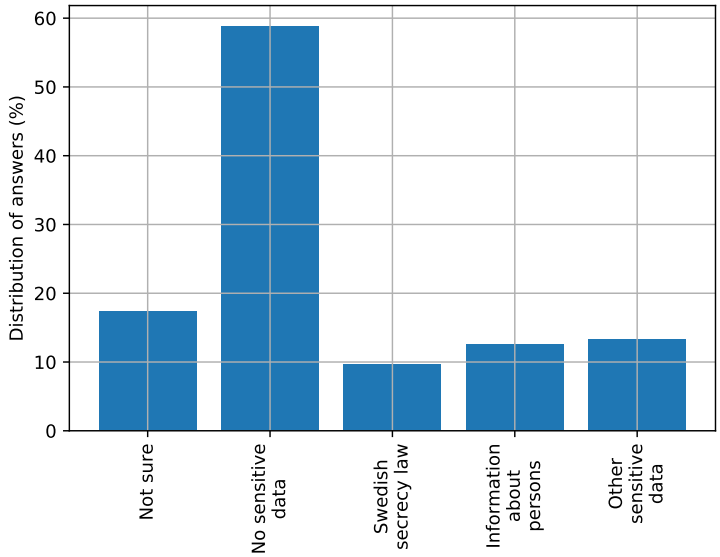

Figure 4: Management of sensitive data.

Table 4: Management of sensitive data.

Type of sensitive data	Count	Relative frequency (%)
Not sure	47	17
I do not work with sensitive data	159	59
Data which is regulated according to Swedish openness and secrecy law	26	10
Information about persons	34	13
Other sensitive data	36	13

2.6 Question 1f: Archival requirements

Question To what extent are you familiar with the archival requirements a Swedish government agency is required to fulfil, and how those requirements affect your data management?

Four options were given; “To a great extent”, “To some extent”, “To a small extent” or “Not at all”.

The 274 responses to this question were distributed as shown in fig. 5 and tbl. 5.

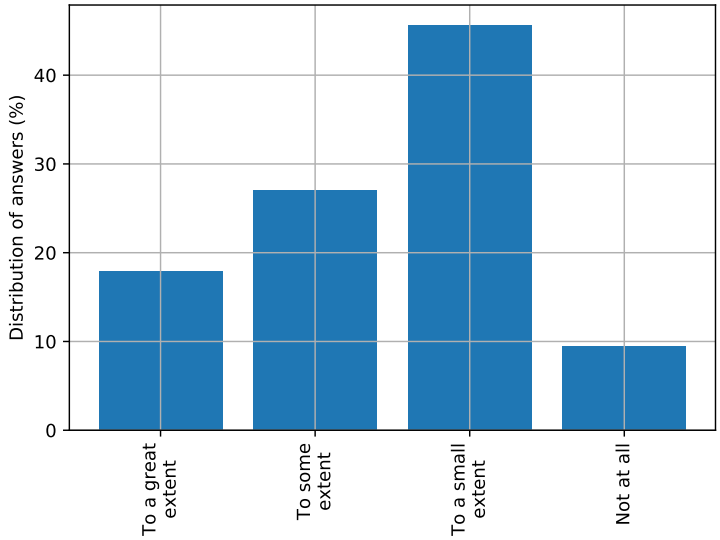

Figure 5: The extent of knowledge about archival requirements.

Table 5: The extent of knowledge about archival requirements.

Extent of knowledge	Count	Relative frequency (%)
To a great extent	49	18
To some extent	74	27
To a small extent	125	46
Not at all	26	9

2.7 Question 1g: Participation in external collaborations

Question Do you participate in external collaborations or partnerships?

The options were 'yes' and 'no'. Of the 274 respondents, 79 % indicate that they participate in external collaborations and the remaining 21% indicate that they do not.

2.8 Question 1h: Version control

Question Do you use any version control system for data?

The available options were "No, I do not use any such system", "Yes, I use manual version control with a naming scheme for files and directories" or "Yes, I use systems for automatic version control (e.g. Git, SVN). Please specify which". Multiple answers were possible.

The distribution of the 275 answers are shown in tbl. 6.

Table 6: System for version control of data.

Version control	Count	Relative frequency (%)
No system for version control	142	52
Manual version control	119	43
Automated version control	28	10

21 text answers were provided to state what system(s) are used by respondents for version control of their data. The majority (16 answers) have stated that they use a variant of Git (GitHub, GitLab or even Git-annex). A few people (3) have indicated that they use Subversion. TimeMachine, Mercurial, Fossil, LIMS and zfs snapshots have been indicated by one data producer each.

3 Question 2

Question 2 mainly deals with attitudes to and previous experiences from open data. The purpose of these questions were to determine the extent of open data publication at SLU and to examine the data producers´ attitude towards an increase in openly sharing the data they generate. We also wanted to find out what motivates and what hinders the data producers from sharing data.

3.1 Question 2a: Experience of sharing data

Question Have you shared or made research data or EMA data available in any way?

Five options were available; "No", "Yes, I have shared informally with close colleagues", "Yes, I have shared data at request from other people than close colleagues", "Yes, I have made data available via website (research project site or personal site)" and "Yes, I have made data available via data repository or data archive. Please specify which". Multiple answers were possible.

265 data producers answered this questions (see fig. 6). Of the respondents, only 17 % state that they have not shared data at all. 61% of the respondents claim to have shared data informally with close colleagues. 37% have shared with people other than close colleagues and 21% have made data available through a website. 28 % have shared data via a data repository or data archive.

The distribution of the answers is shown in fig. 6. The repositories specified by those who selected the last option are shown in tbl. 7, for repositories reported by at least two data producers. In addition to these services a number of minor services are mentioned as well as the platforms that DMGD provides.

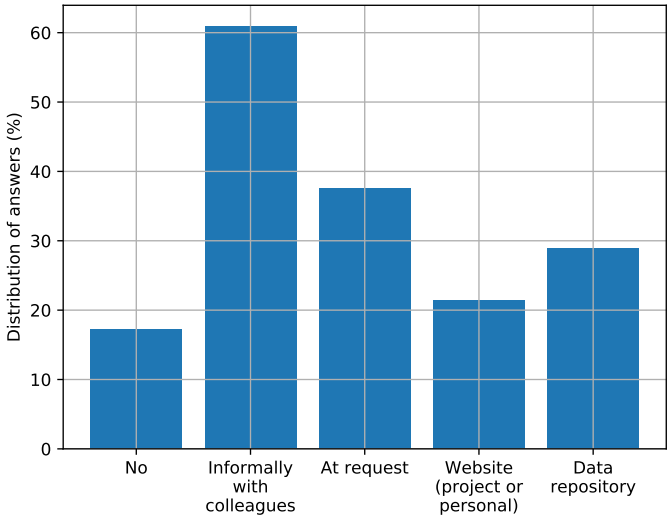

Figure 6: Distribution of the answers to question 2a, “Have you shared or made research data or EMA data available in any way?”.

Table 7: Platforms where data producers make the data available.

Name of platform	Count
Dryad	12
NCBI	6
Artdataportalen	5
European Nucleotide Archive	4
Github	3
Zenodo	3
Figshare	2
Genbank	2

3.2 Question 2b: Attitudes towards sharing data

Question To what extent are you interested in openly sharing the data you work with?

There were five available options: “To a great extent (everything or as much as possible)”, “To some extent (larger selected parts)”, “To a small extent (smaller selected parts)”, “Not at all” or “Data cannot be shared, due to e.g. secrecy”.

33% of the respondents feel that they can share data to a great extent, 43% that they can share it to some extent and 19% to a small extent. Only 3 % of the respondents feel that they cannot share data at all. The distribution of the answers are shown in fig. 7.

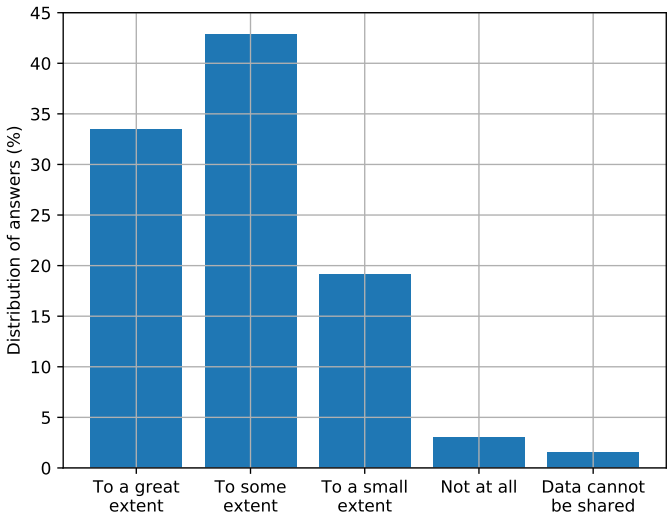

Figure 7: Distribution of the answers to question 2b,”To what extent are you interested in openly sharing the data you work with?”.

3.3 Question 2c: The importance of sharing data

Question Do you consider it important to openly share data with the research community and the general public?

The data producers were asked to choose their answer from a scale of 1 to 5, where 1 equals “absolutely not” and 5 equals “very important”. Of the 267 respondents, 35% have answered that it is very important to share data (5). 36% answered 4, 20% answered 3 and 7% answered 2. 2% of the respondents indicate that they do not think that sharing data is important (1). The distribution of the answers are shown in fig. 8.

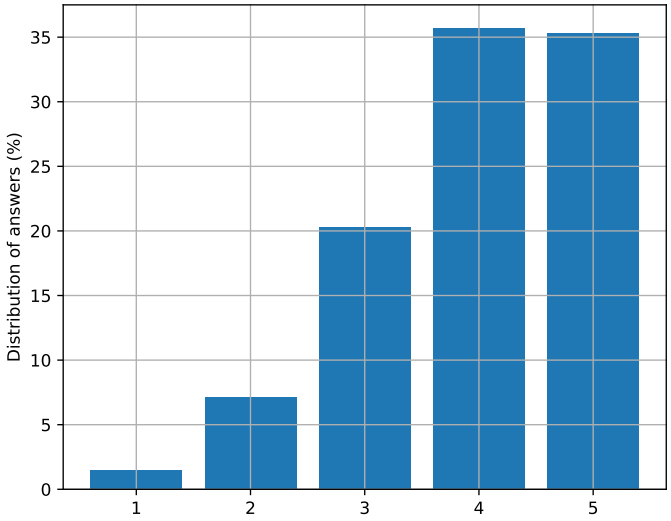

Figure 8: Distribution of the answers to question 2c,” Do you consider it important to openly share data with the research community and the general public?”.

3.4 Question 2d: The importance of sharing data – reasons given by the data producers

Question Motivate your choice in 2c

The respondents were asked to provide the reason for their answer to question 2c. The 145 text answers have been divided into two headings, those which have a positive attitude towards sharing data (tbl. 8), and those that are negative to it (tbl. 9), and then further into sub-categories under each heading. The same respondent may have given answers in more than one category and both headings.

Among the answers that indicate a positive attitude to sharing data positive to sharing data, the main motive is that it furthers research in some way (46 answers). The comments touch upon reasons such as the use of data by more than one research group, and that research data has great potential that is not currently used to the full. A large proportion of these respondents believe that data can be used for other purposes and for several kinds of analysis and therefore should be made openly available. The common denominator is that a greater understanding can be achieved if data is made openly available, and that this is the main purpose of research.

The second largest category of answers express a positive attitude to making data openly available as this

contributes to the public good (34 answers). Most of these answers do not specify how, but express an over-all positive attitude to data sharing. Some answers however specify that spreading research results to a larger extent to the general public is what constitutes the previously mentioned "public good". Another major reason for sharing is that it is a matter of principle as the results are generated with funds from the taxpayers (21 answers).

The opportunity to re-examine already published results is indicated by 25 respondents as a reason to share data. They argue that review of research results is difficult or even impossible if the underlying data sets are not publicly available. These answers express an opinion that availability of data increases the research quality and contributes to accessibility. Additional respondents mean that availability of data increases the transparency within research (2 answers) and that long-term preservation increases the durability of research (2 answers).

A smaller group of data producers (7 answers) have provided answers that focus on an increased potential for collaboration between researchers by sharing data.

The respondents who voiced negative opinions towards sharing data express worries that data will be mis-interpreted (14 answers) and others feel that it is too time-consuming both to share data (7 answers) and to properly describe data before sharing (6 answers) and the resources required for data sharing are not available. Several respondents are concerned about sharing raw data (13 answers), but are happy to share the data after it's been processed. Some are reluctant to share data due to its sensitive nature (12 answers). These and other negative opinions on sharing data are further described in tbl. 9.

Table 8: Categories of answers expressing a positive attitude to sharing data.

Category	Description	Count
furthering research	Making data available is seen as a way to further research. Can relate to speed, the possibility to answer questions et.c.	46
for the common good	Making data available is beneficial for the common good. The answers in this category represent a positive attitude to sharing data without any further details into how.	34
review of research results	Making data available is seen as warrant that the research is conducted in an appropriate way or as a means to prevent fraud.	25
publicly financed	A positive attitude to data sharing based on the belief that data funded by public funds should be publicly available.	21
collaboration	Sharing data is beneficial to research collaborations.	7
reuse/big data	The possibility to reuse data and to generate large datasets to answer research questions.	5
exterior demands	Making data available because major players (e.g. funders or journals) demand it.	3
long-term use	Future use of data.	2
transparency	Increased transparency into the research process.	2

Table 9: Categories of answers expressing a negative attitude to sharing data.

Category	Description	Count
timing	Comments concerning the timing of data publishing. The general view is that data should not be made openly available until the results of the study have been published in a journal or elsewhere.	18
risk of mis-interpretation	Not keen to share data in case it is mis-interpreted.	14
data type	A disinclination to share raw data. Processed data however is OK to share.	13
secrecy/sensitive data	Reluctance to share data because of the nature of data that can be sensitive or imposed with secrecy.	12
available resources	Making data openly available is resource and time consuming and will be done at the expense of other tasks.	7
metadata	Respondents are of the opinion that data must be thoroughly described, and that this is too time consuming.	6
scooped	A concern that other researchers will use the data for their own purposes and get credit for the results before the data producer has had a chance to fully benefit from the data themselves.	6
competitive edge	Reluctance to share data as it is viewed as the competitive edge and the reason that the data producer is funded.	6
misinterpretation by general public	The general public is not able to correctly interpret data.	4
citation and merits	It must be meritable to publish data. Citation of data must become the norm, or at least more meritable.	4
defintion of data	The respondents comment on the inadequate definition of data in the question.	3
target audience	Reluctance to publish data as the data producer feels the need to determine who gets access to the data.	3
Not standard procedure	Sharing data is unusual in the discipline they are working in.	1
size of datasets	The datasets are too large to easily share.	1

3.5 Question 2e: Making data openly available

Question What would induce you to start sharing data openly, or to make data openly available to a greater extent than you do today?

The data producers were given 9 options, and the possibility to select more than one of them. The options were as follows: “Higher demand from other researchers working in my subject area”, “Access to tools or platforms for sharing data”, “Making data available would be seen as meritorious”, “Better knowledge about sharing data”, “More support”, “Impact for the open data is measurable”, “SLU policy or decision by the vice-chancellor for open data”, “Other” and “Do not know”. In addition to this it was possible for respondents to leave a comment to their answer.

266 data producers responded to this question and the distribution of the answers are illustrated in fig. 9 and tbl. 10.

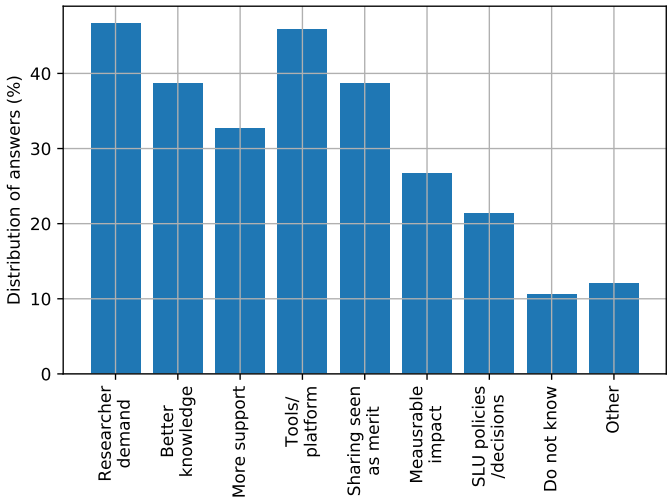

Figure 9: The distribution of the answers to the question “What would induce you to start sharing data openly, or to make data openly available to a greater extent than you do today?”

Table 10: Categories of opinions on making data openly available.

Category	Count	Relative frequency (%)
Higher demand from researchers	124	47
Better knowledge about data sharing	103	39
More support	87	33
Access to tools/platform	122	46
Sharing seen as merit	103	39
Impact for open data meausrable	71	27
SLU policies/decisions	57	21
Do not know	28	11
Other	32	12

31 respondents also provided comments to their answer. The answers were divided into categories that are listed in tbl. 11. One respondent may have given answers in more than one category.

Table 11: Categories of opinions on making data openly available.

Category	Description	Count
resources	Need more time and/or financial resources.	14
meritable	Would like it to be meritable to share data.	5
data curation support	Would like support from the Data Curation Unit.	3
infrastructure	Would like better infrastructure for sharing data (storage space, bandwidth etc.).	3
bottom-up initiative	Opposes a top-down management.	1
only share if others share	There is a risk for loss of competitive edge if SLU researchers are the only ones to share data.	1
continuous use of data	Wants to use data for future, personal projects.	1
secrecy	Requires that sources let go of secrecy demands.	1
control, risk for misinterpretation	Wants control over use and interpretation of data to prevent misinterpretation.	1
existing demands	Publication of data is an existing demand and the data producer is currently looking to create a platform for publication.	1
freely available	Let anyone access data that has been generated.	1
open standards	Data must be accessible to people using open source software. There must be an open standard for data storage.	1

3.6 Question 2f: Data management plan

Question Do you have a data management (DMP) plan for your research/EMA data?

264 data producers responded to this question. 28% of them have a DMP, 43% of them don't and 29% don't know what a DMP is, as illustrated by fig. 10.

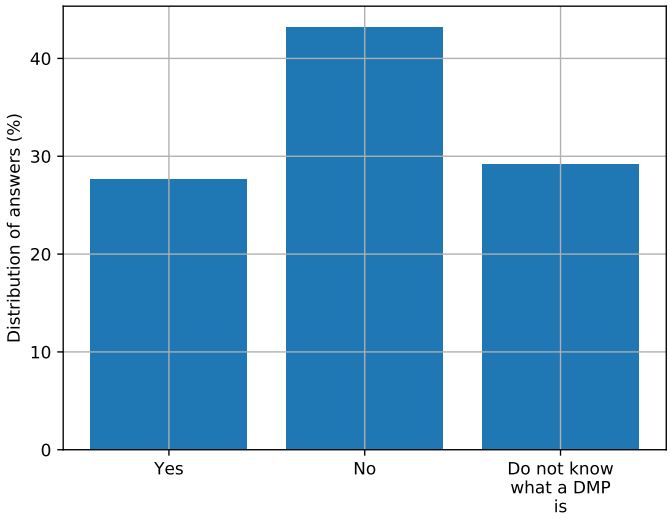

Figure 10: The distribution of the answers to the question "Do you have a data management (DMP) plan for your research/EMA data?".

4 Question 3

4.1 Question 3a: Affiliation

Question What institution or organization are you primarily affiliated with?

In order to determine the distribution of the respondents at the university, we asked them to indicate their primary affiliation. 262 of the recipients answered this question and the distribution of the respondents are shown in tbl. 12.

Table 12: Primary affiliation.

Institution/department	Count	Relative frequency (%)
Department of Aquatic Resources	116	44
Department of Work Science, Business Economics and Environmental Psychology	34	13
Department of Biomedical Sciences and Veterinary Public Health	31	12
Department of Biosystems and Technology	31	12

Institution/department	Count	Relative frequency (%)
Department of Ecology	35	13
Department of Economics	20	8
Department of Animal Nutrition and Management	3	1
Department of Animal Breeding and Genetics	9	3
Department of Clinical Sciences	17	6
Department of Landscape Architecture, Planning and Management LAPM	2	1
Department of Soil and Environment	16	6
Department of Molecular Sciences	8	3
Department of Agricultural Research for Northern Sweden	4	2
Department of Forest Biomaterials and Technology	3	1
Department of Forest Ecology and Management	9	3
Department of Forest Genetics and Plant Physiology	5	2
Department of Forest Mycology and Plant Pathology	6	2
Department of Forest Resource Management	29	11
Department of Urban and Rural Development	11	4
Southern Swedish Forest Research Centre	11	4
Department of Aquatic Sciences and Assessment	6	2
Department of Wildlife, Fish, and Environmental Studies	10	4
Department of Plant Biology	6	2
Department of Plant Breeding	7	3
Department of Crop Production Ecology	1	0
Department of Plant Protection Biology	19	7
School for Forest Management	2	1
Other	9	3

4.2 Question 3b: Research centre/activity

Question Are you affiliated with a research centre/an EMA activity/other activity?

160 recipients responded to this question. 39% indicate that they belong to a research centre, 26% to an EMA activity and 35% to another activity, as shown by tbl. 13.

Table 13: Activity of the respondents.

Institution	Count	Relative frequency (%)
Research centre	62	39
EMA activity	42	26
Other activity	56	35

4.3 Question 3c: Specification of activity of the respondents

Question Please specify which research centre or EMA activity

32 text answers were provided to this question and they indicate that the respondents come from a large variety of research centres and activities at SLU as the same answer was rarely given by more than one respondent. Those few answers given by more than one respondent included National Inventory of Landscapes in Sweden (NILS), the Institute of Freshwater Research and Swedish National Forest Inventory.

4.4 Question 3d: Job title

Question Which title describes your occupation most accurately?

Four options were available; “Professor”, “PhD student”, “Associate professor, postdoc or other research occupation” or “other”. 263 recipients have responded to this question and the distribution of the answers are as follows: Professor, 19%, PhD student, 8%, Associate professor, postdoc etc., 60% and other, 13%. The answers are illustrated in fig. 11.

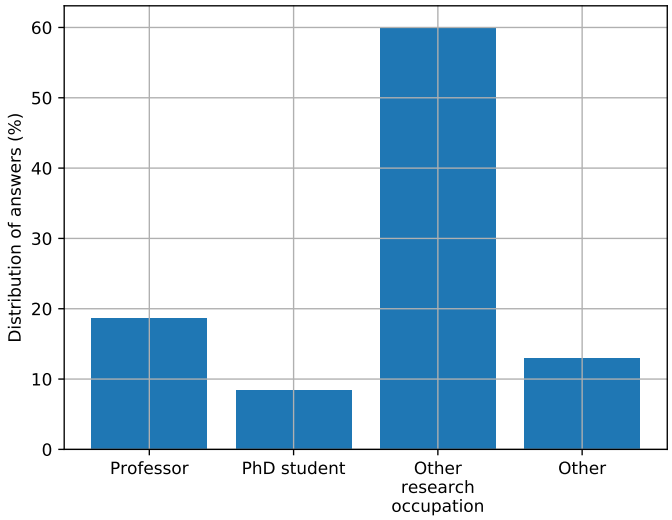

Figure 11: The distribution of answers to the question “Which title describes your occupation most accurately?”.

4.5 Question 3e: Funding

Question Is your current activity wholly or partially externally funded?

The available options were “yes” and “no”. 259 recipients responded to this question, of which 17 % answered “no” and 83% answered “yes”.

4.6 Question 3f: Other comments

Question Other comments

This was an opportunity for the respondents to freely comment upon the content of the survey or issues related to the questions in the survey. 22 respondents have taken the opportunity to leave a comment. Several have

expressed a concern that archiving and sharing data will be yet another burden for data producers, others express a wish for further education concerning these activities.

5 Conclusions and discussion about the results

The initial survey was sent to 1441 data producers at SLU. Most of the questions have been answered by just below 20% of the recipients. Due to the relatively low response rate the interpretation of the results must be done with some caution. We can't assume that the respondents are representative for data producers at SLU. It is entirely possible that the recipients that chose to respond to this survey have a greater interest in open data than the average data producer and are therefore more involved in issues regarding data management. Data producers that already share data and are in favour of open data may still be reluctant to deliver data to Tilda due to the perceived extra work load demanded of them, if they already have published data in an external repository/data portal. Hopefully, the added value of Tilda (archiving as well as publishing) can convince them. At the same time those that lack an adequate repository/data portal may find Tilda extra useful.

5.1 Conclusions from the answers to Question 1: Type of data, sensitive data and archiving of data

From the answers to questions 1 a, b and c we can conclude that data producers at SLU work with multiple data types, software and data formats. Among the more common data types we see numerical and statistical data, genome sequences, spreadsheets, databases and documents and reports. To manage these data types the data producers use a large variety of software, both open source and proprietary programs. The most common being Microsoft Excel, but a number of other software; R, ArcGIS, SAS, Microsoft Word and QGIS to name a few, are also relatively common. In general, the types of software used at SLU provide a possibility to export data in open formats, which from a reusability and archiving point of view is a very positive finding. The DCU must however remember to take this into account when a submission agreement, i.e. the contract between data producers and SLU that describes the conditions for delivery and management of data for preservation, is established to make sure that the information is quality assured with regards to e.g. data format.

We also needed to understand what amount of data that the data producers at SLU already stores to enable us to ensure enough storage space in Tilda. As seen by the answers to question 1d, the amount of data that the data producers need to archive varies greatly. A fair number (19 %) of respondents claim that they have more than 1 TB of data that needs archiving and the text answers say up to PB of data. However, it is possible that several of the respondents are referring to the same data set when they indicate the volume they need to archive. This is by no means an exhaustive analysis of the amount of data that SLU data producers possess, but it indicates that the volumes that we need to be able to store are substantial and are growing continuously, and more likely in the PB than the TB range.

Now that the General Data Protection Regulation (EU 2016) has taken effect, it is very important for the development of Tilda to determine to which extent the SLU data producers collect sensitive or personal data. The majority of the respondents (59%) claim to have no sensitive data of any kind. 35% say that they have some kind of sensitive/personal data and 18% of the respondents are not sure if they have this kind of data. Some respondents are not sure about the difference between raw and processed data, and the extent of awareness of the Swedish archival requirements vary from those that are well aware of them (9 %) to those who know nothing at all (18%). The majority of respondents have some knowledge about the requirements (some extent 46 % and small extent 27%). The insight into laws and regulations, or perhaps the lack thereof, clearly indicates that education of data producers concerning the legal aspects of data management is required.

An overwhelming majority (79%) of the respondents participate in external collaborations. Any education

of the SLU data producers must therefore clarify to them that the data management in collaborative projects must take into account the demands from Swedish law, e.g. the principle of public access to information (SLU 2009) and archival requirements already at project start.

Relatively few of the researchers that have answered the survey have specified a system for automated version control, but many of the more common systems, Git in particular, are represented among the answers.

One possible conclusion is that those who responded to the survey do not realize that, for example, Git is included as automated version management. It is also unclear (already in the questionnaire) if it involves version management of data and/or software source code and the like.

5.2 Conclusions from the answers to question 2: Open data

Most of the respondents have at least some experience from sharing data and/or making data available. Only 17% say that they haven't shared data at all, which is fewer than expected. Of the respondents that have shared data, a majority (61%) have done so informally, with close colleagues. Surprisingly 50% of the data producers that responded to our survey have made data available via web sites, data repositories or data archives. The large proportion of respondents that have already made data available in some way is reflected in the answers to question 2b, "To what extent are you interested in openly sharing the data you work with?" and question 2c, "To what extent are you interested in openly sharing the data you work with?". A majority of the respondents (75%) are ready to share their data to a great (everything or as much as possible) or some (larger selected parts) extent. Only 3% are not willing to share data at all. And on a scale from 1 to 5, where 1 equals "absolutely not" and 5 "very important", 71% answered 4 when asked whether they consider it important to openly share data with the research community and the general public. The average to this question was 3.7 which shows that the SLU data producers on average are open to the concept of open data and making data available to the research community and a general public.

The respondents that have a positive attitude towards sharing data believe that sharing will benefit the research community and the greater good. They believe that transparency in research is beneficial and a way to prevent fraud. Among those who are more reluctant to share data the prevailing opinion is that they want to control the timing of data publication to occur after the results have been published (e.g. in a journal). Another reason not to share data is the risk for misinterpretation by whoever wants to use it. Another common concern is that sharing data will take time and money from the research activities. A majority of the research community does not seem to include the process of creating and publishing open data as a basic research activity. This attitude needs to change if open data, and Tilda, is to become successful, but will most likely be a slow process. Relatively few of the comments concerned worries about being scooped by other researchers or about losing the competitive edge by sharing data. All these opinions considered, we again perceive a need for information to and education of the SLU data producers. Apart from information about the possibilities with Tilda, information about the archival requirements in Sweden may be necessary. At the same time we are surprised with the number of comments that are in favour of sharing research data and hope that this also indicates a general interest in the platform for archiving and publication being developed for their use.

When asked what would induce survey recipients to start sharing data openly, or to make data openly available to a greater extent than today, "Higher demand from other researchers working in my subject area" is the top priority of the respondents and "Making data available would be seen as meritorious" is also high-ranking. Neither of these are areas that Tilda or the DCU can influence directly, but all datasets in Tilda will be provided with a digital object identifier (DOI; a persistent identifier of a digital object) to enable easier citation of the dataset, which in turn will make the published dataset meritorious for the data producer. The three areas where the DCU can make a difference to SLU data producers; "Access to tools or platforms for sharing data", "Better knowledge about sharing data" and "More support" all get a fairly large proportion of the answers. Tilda will be a platform for sharing data generated at SLU and we need to make sure that it is easily accessible and

comprehensible to really become a support for data producers, rather than a burden. Along with education about the Tilda system, the DCU will provide support and training that will contain information about many of the aspects of data management and data sharing that have been identified through this survey, hoping that this will persuade more data producers to make their data publicly available, ideally through Tilda.

The awareness of archival requirements is fairly low at SLU, as seen by answers to question 1f and to some extent question 1e. In addition to the training, the concept of a data management plan and a submission agreement may need to be introduced to a larger proportion of the SLU data producers to increase the knowledge of data management in a long-term preservation perspective. A better understanding of good data management practice early on in the research process will substantially decrease the need for resources when data archiving and publication are concerned.

The DCU plan to publish the data underlying this report in the Tilda system as soon as it is available. Until then data will be supplied on request to dcu@slu.se.

References

EU. 2016. “EU Regulation 2016/679.” April 27, 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679/&rid=1.

SLU. 2009. “Offentlighets- Och Sekretesslag (2009:400).” May 20, 2009. http://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/offentlighets--och-sekretesslag-2009400_sfs-2009-400.

———. 2016. “SLU:S Strategi 2017–2020.” June 16, 2016. https://www.slu.se/globalassets/mw/org-styr/styr-dok/vision-strategi/slus-strategi-2017-2020-faststalld-160616.pdf.

Appendix A The survey form

Note that this is a verbatim copy of the English survey form, and that links may not be up to date.

Survey about research and environmental monitoring and assessment data at SLU

This survey is part of the development of Tilda, the new SLU system for publishing and archiving research and EMA (Environmental Monitoring and Assessment) data. Tilda will become the central repository for long-time preservation of data created at SLU. All data deposed in Tilda will be assigned a persistent identifier (DOI) and licenses, which will facilitate dissemination and citation of your data. All data is archived in accordance with Swedish archival practices, which will ensure long-term data quality and security.

Tilda is not intended for short-time data storage of working material.

In order to make sure that Tilda and the work of the Data Curation Unit (DCU), the new support function for data management at SLU, will be based on the best information possible, we would like you to answer a few short questions about what types of data you work with and how you manage them.

Your answers will be processed anonymously. If you want to know more about Tilda, or have questions about data management, feel free to contact DCU (dcu@slu.se). You can also look at the Tilda webpage.

1. Questions about research and EMA data

1 a) Which data types do you work with? Please select one or more of the following options.

- ☐ I do not work with digital data
- ☐ Automatically generated data from computer applications
- ☐ Databases
- ☐ Digital photos and other raster graphics
- ☐ Digital audio files
- ☐ Digital video files
- ☐ Documents or reports
- ☐ Self-developed software within project
- ☐ Geographical data
- ☐ Spreadsheets
- ☐ Numerical data
- ☐ Data collected with sensors or instruments
- ☐ Statistical data
- ☐ Textual data
- ☐ Vector graphics and drawings
- ☐ Websites
- ☐ XML, JSON, and similar formats
- ☐ Other data types, please specify which _____

1 b) Which software products do you use when working with data?

_____
_____
_____
_____
_____

1 c) Do you save data in other formats than those that are standard in the software products specified in 1b? If so, please specify which.

_____
_____
_____
_____
_____

1 d) How large volumes of data do you need to publish and archive? Please select one or more of the following options.

- ☐ 1–500 MB
- ☐ 500 MB–1 GB
- ☐ 1–500 GB
- ☐ 500 GB–1 TB
- ☐ >1 TB
- ☐ Please specify in free-text if none of the alternatives above applies.
  _____

1 e) Do you work with sensitive data? More than one option may be chosen.

- ☐ I am not sure whether or not I work with sensitive data.
- ☐ No, I do not work with sensitive data.
- ☐ Yes, data which is regulated according to Swedish openness and secrecy law.(*)
- ☐ Yes, information about persons.
- ☐ Yes, other sensitive data. Please specify which kinds.
  _____

1 f) To what extent are you familiar with the archival requirements a Swedish governemnt agency is required to fulfill, and how those requirements affect your data management?

- ☐ To a great extent.
- ☐ To some extent.
- ☐ To a small extent.
- ☐ Not at all.

1 g) Do you participate in external collaborations or partnerships?

- ☐ Yes
- ☐ No

1 h) Do you use any version control system for data? More than one option may be chosen.

- ☐ No, I do not use any such system.
- ☐ Yes, I use manual version control with a naming scheme for files and directories.
- ☐ Yes, I use systems for automatic version control (e.g. Git, SVN). Please specify which. _____

*** The law is available in Swedish at http://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/offentlighets--och-sekretesslag-2009400_sfs-2009-400**

Language

- ☐ English
- ☐ Svenska

2. Open data

2 a) Have you shared or made research data or EMA data available in any way? Please select one or more of the following options.

- ☐ No
- ☐ Yes, I have shared informally with close colleagues.
- ☐ Yes, I have shared data at request from other people than close colleagues.
- ☐ Yes, I have made data available via website (research project site or personal site).
- ☐ Yes, I have made data available via data repository or data archive. Please specify which. _____

2 a) Have you shared or made research data or EMA data available in any way? Please select one or more of the following options.

- ☐ No
- ☐ Yes, I have shared infomally with close colleagues.
- ☐ Yes, I have shared data at request.
- ☐ Yes, I have made data available via website (research project site or personal site).
- ☐ Yes, I have made data available via data repository or data archive. Please specify which. _____

2 b) To what extent are you interested in openly sharing the data you work with?

- ☐ To a great extent (everything or as much as possible).
- ☐ To some extent (larger selected parts).
- ☐ To a small extent (smaller selected parts).
- ☐ Not at all.
- ☐ Data cannot be shared, due to e.g. secrecy.

2 c) Do you consider it important to openly share data with the research community and the general public?

- ☐ Absolutely not
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ Very important

2 d) Motivate your choice in 2c, if you would like to.

_____
_____
_____
_____
_____

2 e) What would induce you to start sharing data opnely, or to make data openly available to a greater extent than you do today? Please select one or more of the following options.

- ☐ Higher demand from other researchers working in my subject area.
- ☐ Better knowledge about sharing data.
- ☐ More support.
- ☐ Access to tools or platforms for sharing data.
- ☐ Making data available would be seen as meritorious.
- ☐ Impact for the open data is measurable.
- ☐ SLU policy or decision by the vice-chancellor for open data.
- ☐ Do not know.
- ☐ Other _____

2 f) Do you have a data management plan for your research/EMA data?

- ☐ Yes
- ☐ No
- ☐ I do not know what a data managment plan is.

3. Contextual questions

3 a) What institution or organization are you primarily affiliated with?

- ☐ Department of Aquatic Resources
- ☐ Department of Anatomy, Physiology and Biochemistry
- ☐ Department of Work Science, Business Economics and Environmental Psychology
- ☐ Department of Biomedical Sciences and Veterinary Public Health
- ☐ Department of Biosystems and Technology
- ☐ Department of Ecology
- ☐ Department of Economics
- ☐ Department of Energy and Technology
- ☐ Department of Animal Environment and Health
- ☐ Department of Animal Nutrition and Management
- ☐ Department of Animal Breeding and Genetics
- ☐ Department of Clinical Sciences
- ☐ Department of Landscape Architecture, Planning and Management LAPM
- ☐ Department of Soil and Environment
- ☐ Department of Molecular Sciences
- ☐ Department of Agricultural Research for Northern Sweden
- ☐ Department of Forest Biomaterials and Technology
- ☐ Department of Forest Ecology and Management
- ☐ Department of Forest Products
- ☐ Department of Forest Genetics and Plant Physiology
- ☐ Department of Forest Mycology and Plant Pathology
- ☐ Department of Forest Resource Management
- ☐ Department of Forest Economics
- ☐ Department of Urban and Rural Development
- ☐ Southern Swedish Forest Research Centre
- ☐ Department of Aquatic Sciences and Assessment
- ☐ Department of Wildlife, Fish, and Environmental Studies
- ☐ Department of Plant Biology
- ☐ Department of Plant Breeding
- ☐ Department of Crop Production Ecology
- ☐ Department of Plant Protection Biology
- ☐ School for Forest Management
- ☐ Other _____

3 b) Are you affiliated with

- ☐ a research centre
- ☐ an EMA activity
- ☐ other activity, please specify which _____

3 c) please specifiy which research centre or EMA activity

_____
_____
_____
_____
_____

3 d) Which title describes your occupation most accurately?

- ☐ Professor
- ☐ PhD student
- ☐ Associate professor, postdoc or other research occupation
- ☐ Other _____

3 e) Is your current acitivity wholly or partially externally funded?

- ☐ No
- ☐ Yes, please specify funder(s) _____

3 f) Other comments

_____
_____
_____
_____
_____

Thank you for your answers!

If you have any questions about publishing or archiving data, please contact dcu@slu.se.