FAIRagro

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
Project number 501899475

In cooperation with
nfdi

FAIRagro Survey: Data quality in the application of agricultural system data

Jannes Uhlott[1], Anne Sennhenn[2] & Markus Möller[1]

Braunschweig, 12th June 2024

[1]Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Crop and Soil Science, Bundesallee 58, 38116 Braunschweig, Germany
[2]Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB), Central IT and Research Data Management, Max-Eyth-Allee 100, 14469 Potsdam, Germany

1. Background

In times of increasing technology in the agricultural sector, more and more digital data is being created, which leverages the importance of topics such as FAIR data management and data quality. In FAIRagro, a consortium of the National Research Data Infrastructure (NFDI), we have set the task of analysing the various aspects of data quality in the agricultural community in order to identify current demands and challenges (Ewert et al., 2023; Specka et al., 2023). This should form the basis for the development of a feedback and curation system in which the quality of research data can be documented in an user-specific format in order to increase the reuse value of research data. In preparation for the FAIRagro consortium, the 'Survey about research data management in agricultural sciences in Germany' by Senft et al., 2022 has already been conducted. In this survey, 52 questions were asked about produced and (re-)used data, data quality aspects, use of standards, publication practices and legal aspects. These were used to determine the needs and wishes of users with regard to future developments in research data management. In addition, Senft et al., 2022 identified challenges, deficits and uncertainties in the handling of research data in agricultural sciences and discussed possible solutions. Based on these results, our survey goes more into detail and focus on the area of data quality in research data management in the agricultural sciences. The results are intended to provide an overview of how data quality is currently viewed in agricultural systems research, what challenges exist and what the participants themselves are prepared to invest in terms of effort.

The survey was implemented within FAIRagro in Measure 3.3 "Measures and Application-data-matrix for Data Quality and Fitness for Use" in cooperation with Measure 2.2 "Community Participation" (Ewert et al., 2023).

2. Methods

The survey 'Data quality in the application of agricultural system data' was available online from 14 November 2023 to 21 January 2024. The target group is defined by all persons who work with data within the field of agrosystem science. In order to obtain a broad distribution of participants, the survey was sent to all FAIRagro (Co-)Applicants and Participants. All institutions were kindly asked to distribute the survey among their scientists (e.g. JKI, ZALF, ATB). Furthermore calls for participation were published in several newsletters (e.g. FAIRagro, JKI scientists). In total, the survey was completed by 321 people, 211 of whom completed the questionnaire in full. The following answers only show the results of the completed questionnaires.

The survey was created with the survey tool Limesurvey and comprises 27 questions, which are divided into five groups:

(A) General information about the participants (4)

(B) Data collection (8)

(C) Data re-use (9)

(D) Data re-use in the Practice (5)

(E) more on the topic of data quality (1).

The question blocks (B), (C) and (D) begin with a mandatory question on the participants' experience with the topic. The following questions in the question block were only asked if the participants had already gained experience themselves.

The survey was made available in both German and English, with the answers in German subsequently being translated into English during post-processing. Original answers can be found in the Appendix. Single-choice, multiple-choice, text-only and ranking questions were used in the survey, with the specific question type indicated in the image description of each question. In order to visualize possible multiple answers to multiple-choice questions, the total number of answers is given in addition to the number of people who answered the question. The percentages per answer option are rounded to absolute amounts, which may result in a value of 0 % being included in the figure or the same values being represented in bars of different lengths. In order to ensure readability, examples for the response options are only shown in abbreviated form in the illustrations. All questions and their complete answers with examples can be found in the question catalogue. The questions of the first group of participants in our survey are inspired by the FAIRagro survey conducted earlier by Senft et al., 2022. For questions with the answer option "Others" and for free text options, the text answers are shown below the figures as text.

3. Results

3.1. (A) General information about the participants

A1: In your everyday handling of data, which group do you most closely identify with?

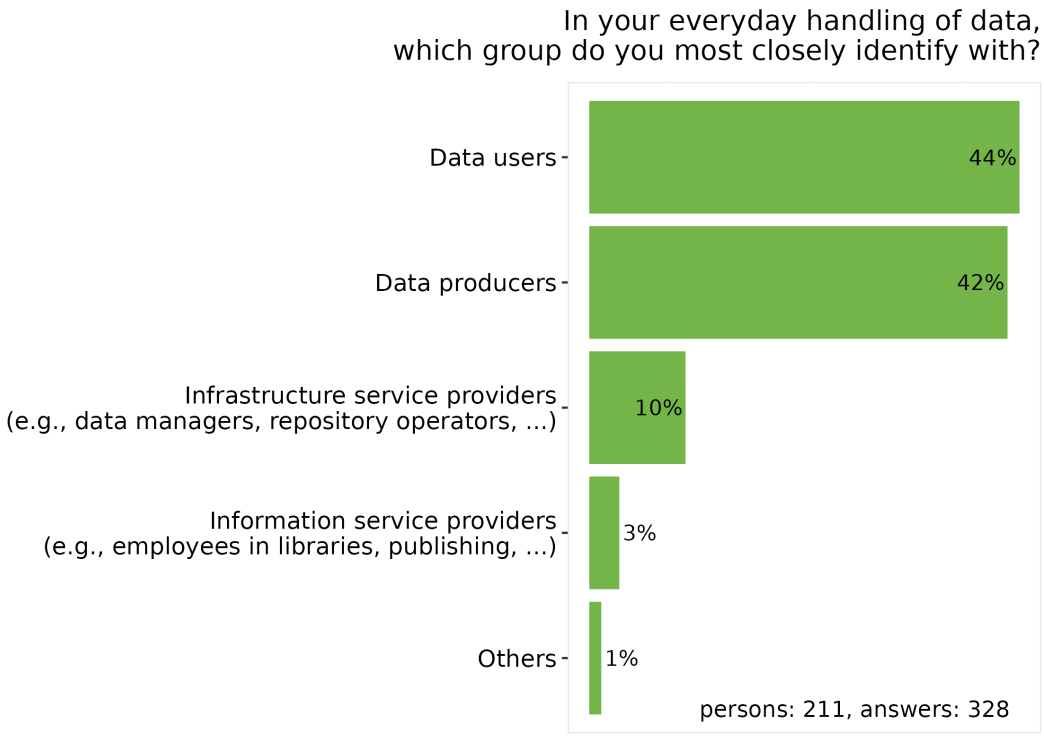

Figure 1: Results of multiple-choice question about participants' daily handling of data.

Additional answers in *Others* (n=4): Quality monitoring, Data, Wikipedia, Scientist.

A2: Which of the following groups do you primarily identify with?

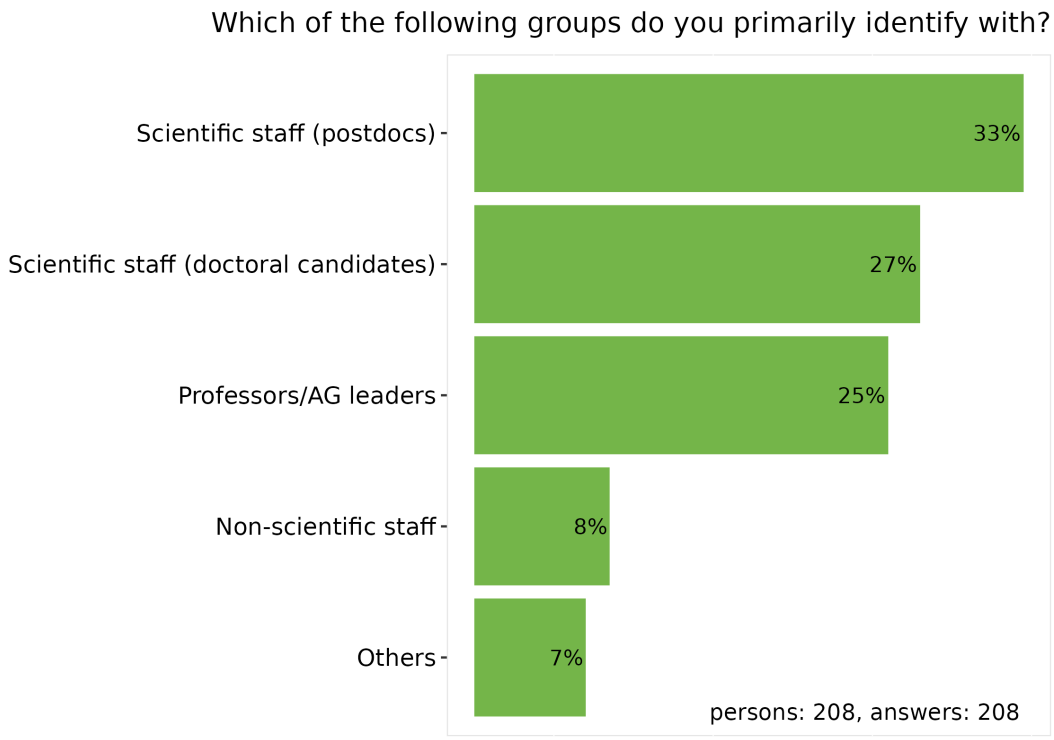

Figure 2: Results of single-choice question on the participants' working position. The response option 'Students' was not selected and are therefore not listed in the figure.

Additional answers in *Others* (n=14): scientific staff (3), scientific staff without a doctoral project (3), MSc (scientific-technical MA), research assistant, employee, farmer, research coordinators, agricultural administration, private sector, 'As a computer scientist, an intermediate between technician and scientist'

A3: Which of the following institutions do you primarily identify with?

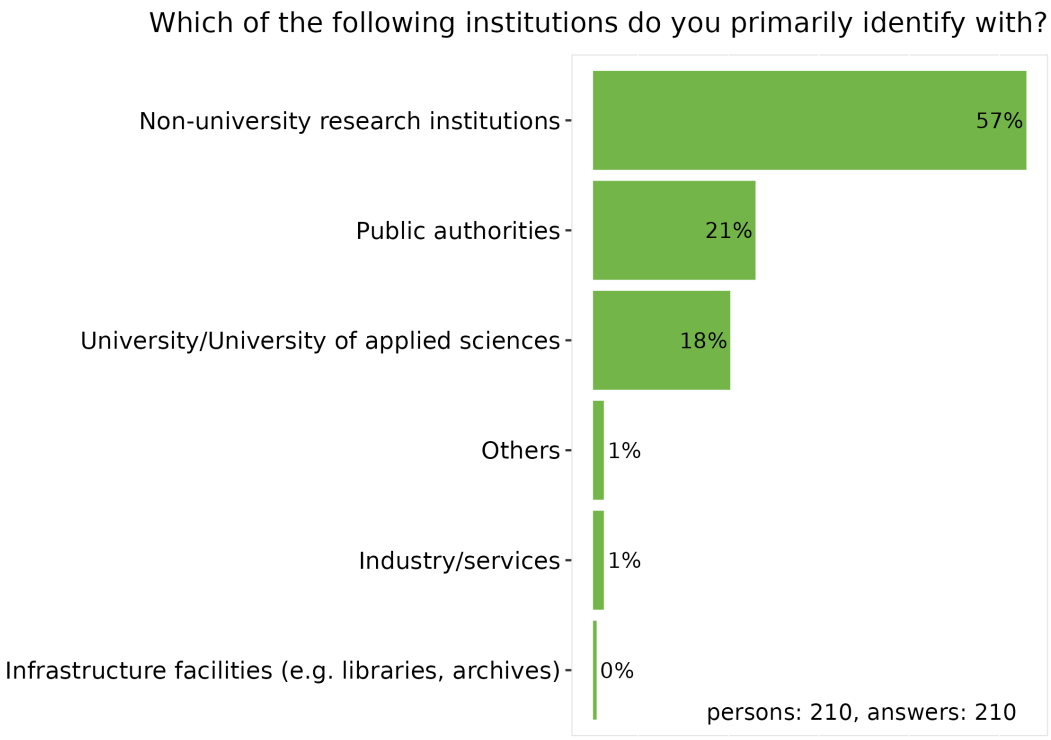

Figure 3: Results of single-choice question about the participants' institutional affiliation.

Additional answers in *Others* (n=3): Farm, Non-university research institutions, Agriculture

A4: To which of the following DFG subject groups can your current work be assigned?

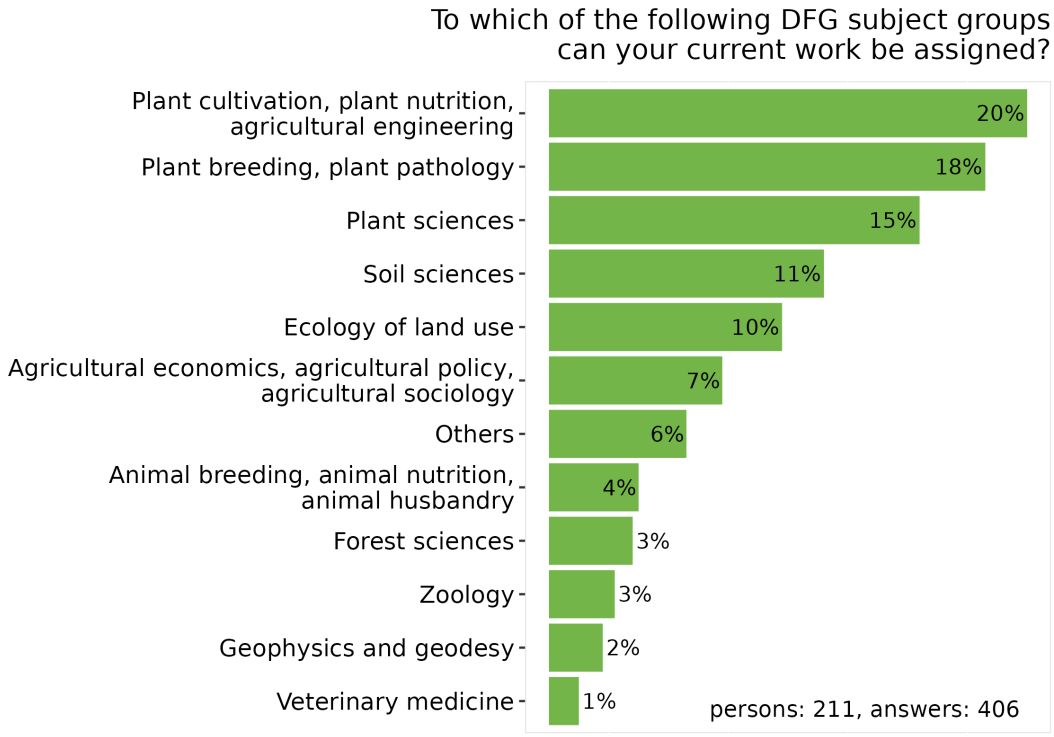

Figure 4: Resuts of multiple-choice question about the participants' affiliation to DFG Groups.

Additional answers in *Others* (n=23): Biodiversity, plant protection (4), molecular mycology, computer science, life sciences, stock protection, mechanical engineering, agroecosystem modelling, entomology (zoology is too general), water resources, management and e-mails, agricultural entomology, life science/food/beverage, entomology, phytosanitary control, hydrology, plant health, data management, bioeconomy, remote sensing, geodata analysis

3.2. (B) Data collection

B1: Do you currently collect or have you ever collected data by yourself?

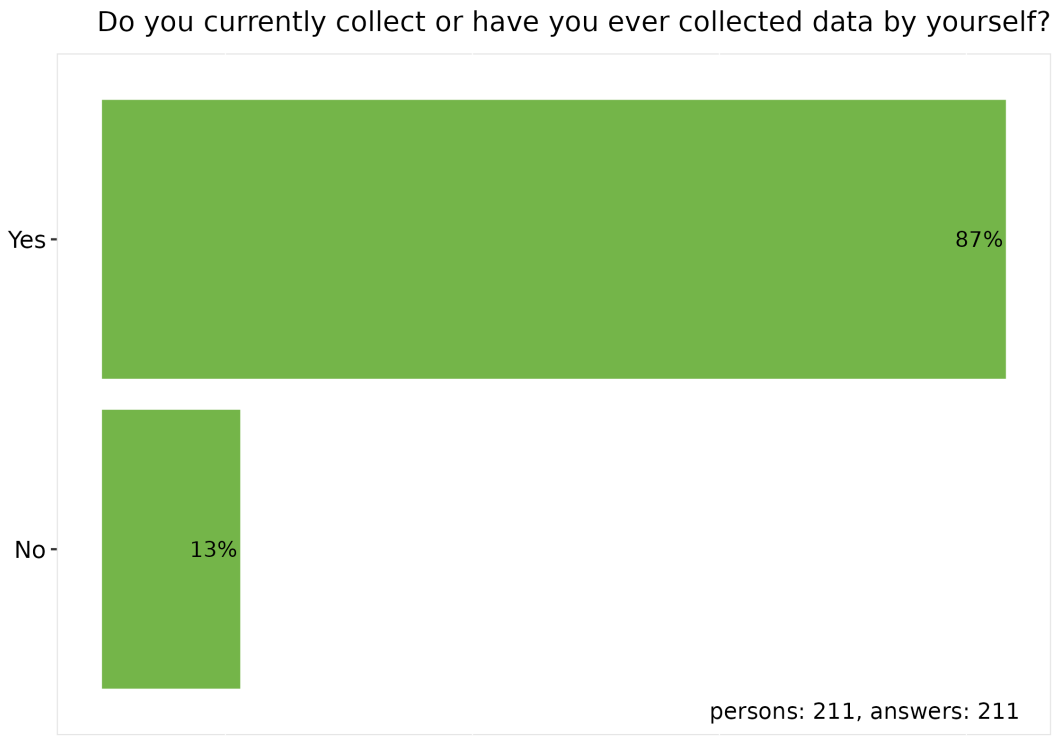

Figure 5: Results of mandatory single-choice question on participants' experiences in data collection.

A total of 183 people stated that they had already collected data themselves. The following questions in block B 'Data collection' only include the answers of those who have already collected data themselves.

B2: For what type of data did the data collection of your example primarily take place?

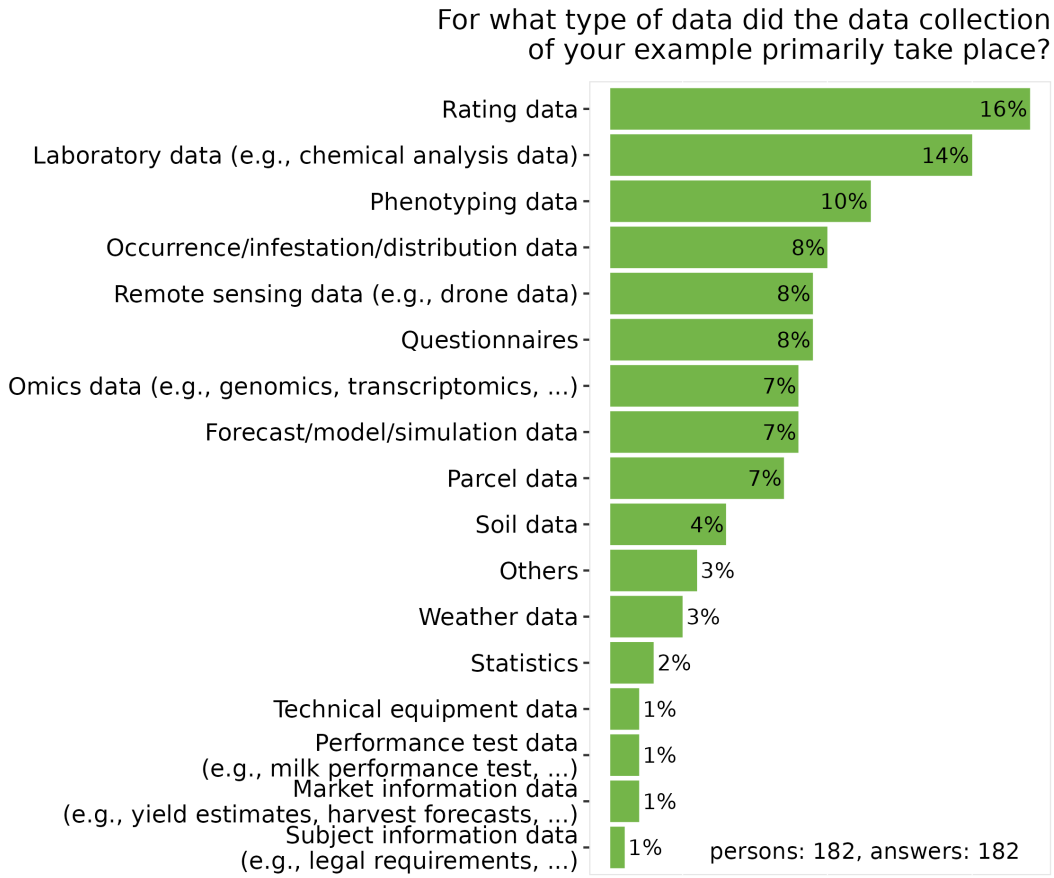

Figure 6: Results of single-choice question on the data category recorded during participants' own data collection.

Additional answers in *Others* (n=6): for various of the listed species; omics, phenotying and lab data; multiple answers would have been possible for me here; energy/carbon footprint data; animal behaviour/health; business data

B8: For what types of data did the data collection in your example take place?

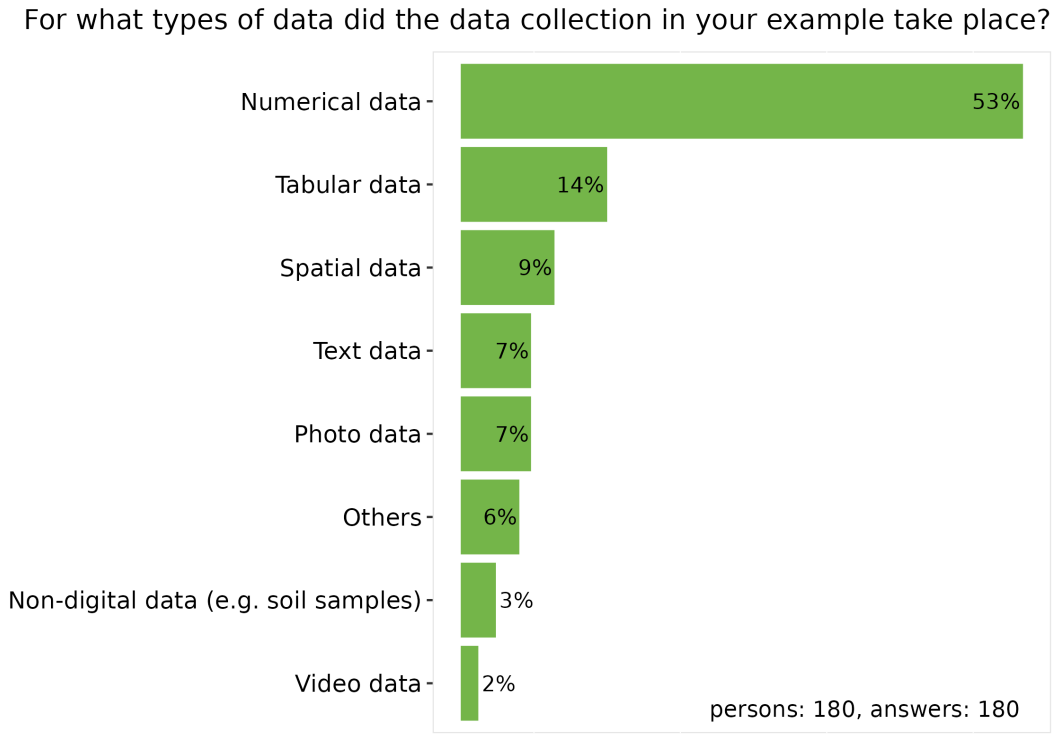

Figure 7: Results of single-choice question regarding which data types are collected by the participants. The response options 'audio data' and 'source codes' were not selected and are therefore not listed in the figure.

Additional answers in *Others* (n=10): Spectrometric data; question unclear and no combinations possible (e.g. most field data, photos, etc. are also spatial data); sequence data; multispectral data; DNA sequences; multiple: text, audio, numerical, tabular; numerical and text; text and photo; equivalent combination of video, numerical data, spatial data; unfortunately the multiple selection is missing here: non-digital data and numerical and photo data

B3: For what application area did the data collection in your example primarily take place?

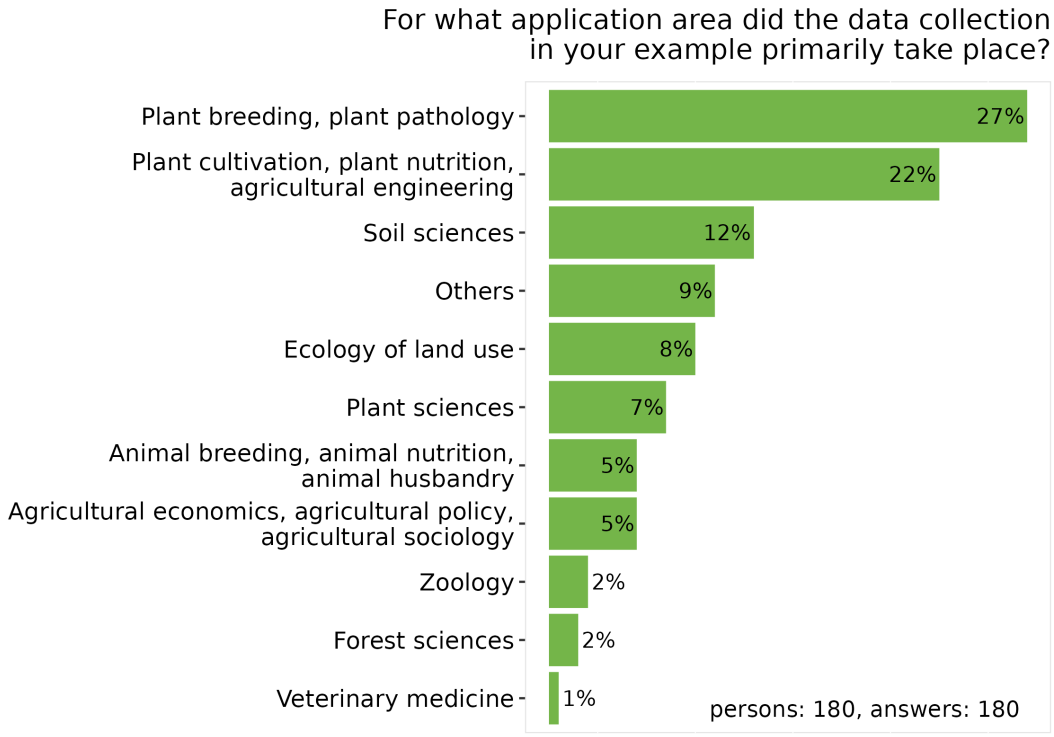

Figure 8: Results of single-choice question about application area of participants' data collection. The response option 'Geophysics and geodesy' was not selected and are therefore not listed in the figure.

Additional answers in *Others* (n=17): Chemical ecology, Plant protection (4), Molecular mycology, Stock protection, Aquatic ecology, Agricultural sciences, Sequence data, Entomology (2), Agricultural entomology, Engineering Data, Electronic equipment, Medicine, Data management purposes

B4: What would be the three most important criteria for you to describe the quality of your collected data?

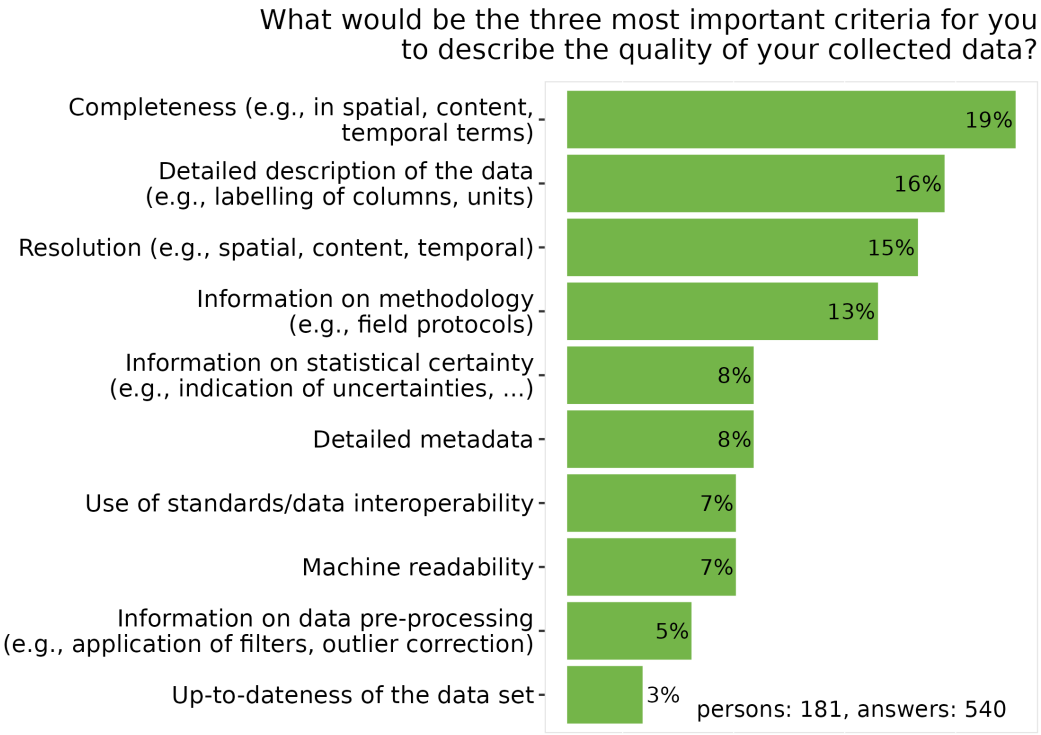

Figure 9: Results of ranking question for three out of twelve answers for the most important data quality criteria for the data collection.

B7: Was there a relevant data quality criterion missing in the previous question? If yes, which one?

Text answers (unsorted, n=28):

- I use RMSE
- Correctness
- Annotation of the data, as there is often not only recorded data but also data that has already been calculated
- I also consider the other suggestions to be very important to ensure the quality and usability of the data!
- Information on the methodology, comprehensive description of the data
- Size of data (number of samples)
- As far as possible, consideration of reference measurements and independent replicates
- Reproducibility
- I also find metadata very important
- Sufficient validation of the data, for example missing in GenBank
- Bias
- Are the data representative (basis for generalisable AI models)
- Locality of data
- Internal consistency of the data sets, labelling of the data with quality flags
- Quality of measurements
- I would like to be able to name and describe several examples as I collect a variety of data types
- Accuracy of the collected (scoring) data
- Usability (i.e. that if data is available, it may be used and not not used for data protection reasons, e.g. host plant availability in all federal states)
- Number of repetitions of a measurement series
- Completeness
- Machine readability
- Comparability (always create the same conditions when taking photos: Background, lighting, camera settings and work with tripod, . .)
- Question not easy to answer because it is not 100% clear what it is aimed at
- Reproducibility
- Comparison/control analyses/measurements
- Yes, because you never collect only one type of data, as you had to choose, but usually need several data for a meaningful interpretation.
- Plausibility check

B5: How do you ensure the quality of your collected data?

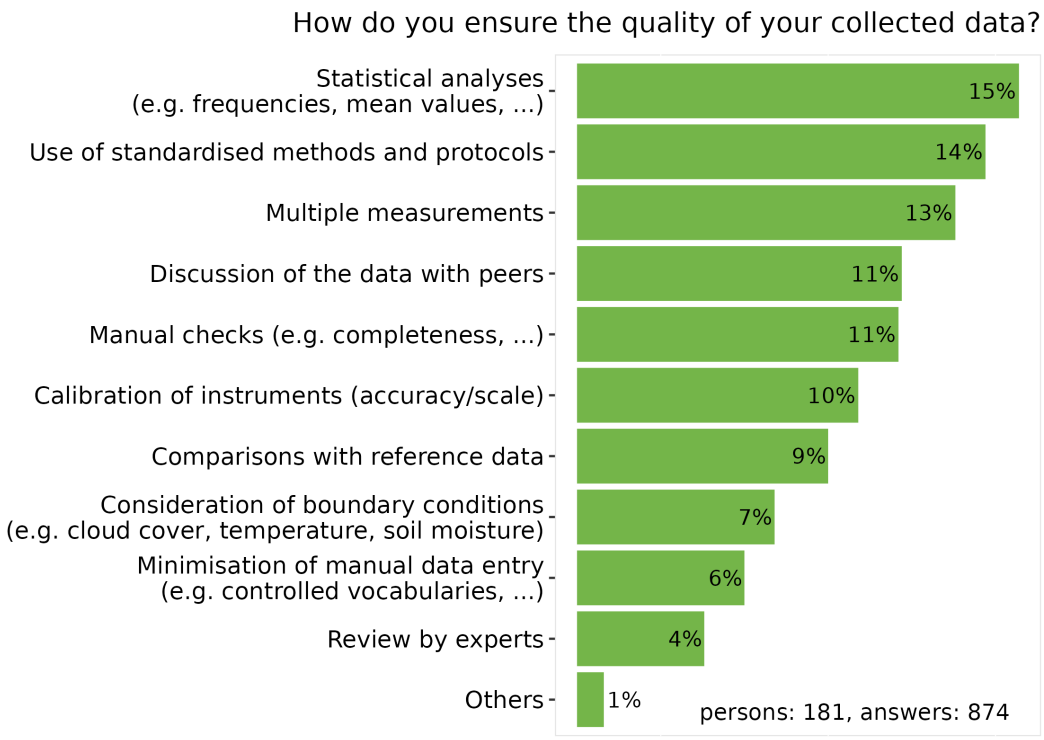

Figure 10: Results of multiple-choice question on the methods used by the participants to currently ensure data quality.

Additional answers in *Others* (n=8):

- Use of negative controls
- Review process
- Repetitions
- Reflexive, transdisciplinary practices e.g. participant validity checks and co-creation of knowledge
- Automated data analysis of the photos: defined colour proportions
- Blinding, randomisation, observer matching, plausibility check
- Modelling approaches
- Extensive 'digitalisation' of the entire data production process - no paper entries.

B6: What are the biggest challenges for you in data collection regarding data quality?

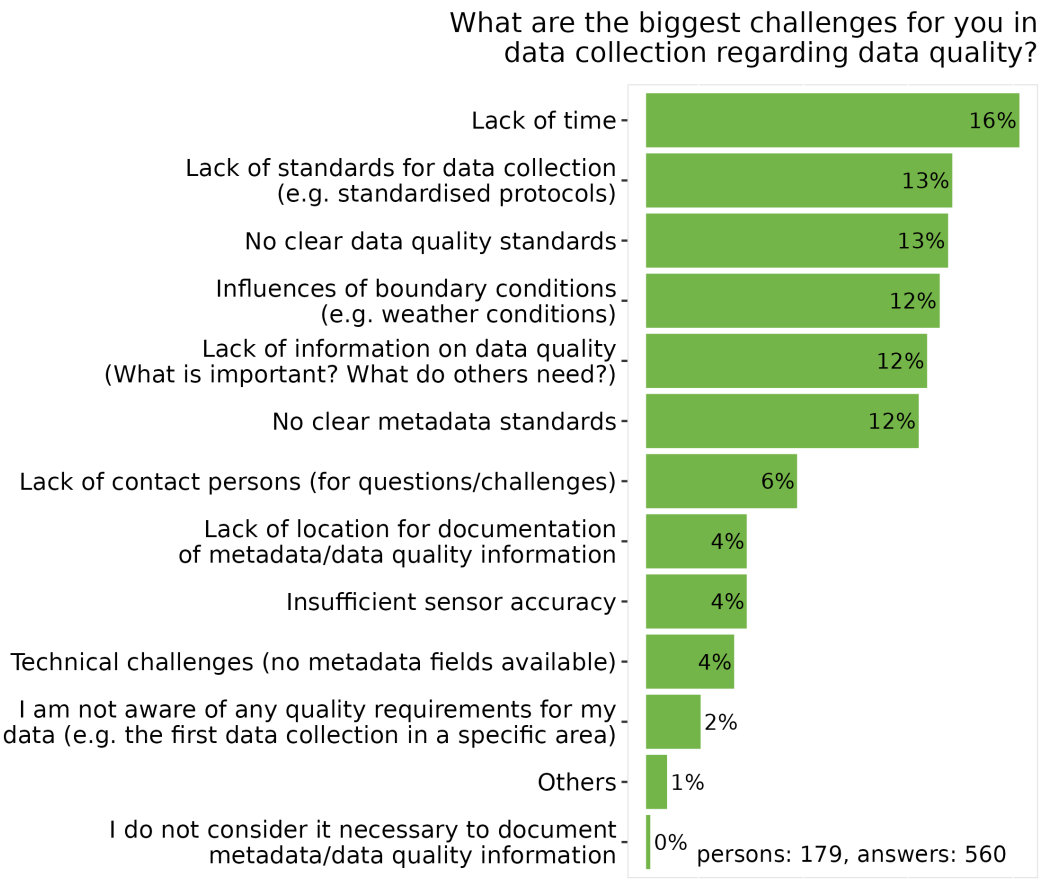

Figure 11: Results of multiple-choice question about the participants' challenges on data quality during data collection.

Additional answers in *Others* (n=5):

- Lack of infrastructure at the JKI (computer power, etc.)
- operational blindness
- Lack of objectivity in manual surveys
- Protocols are not adhered to
- For the data we collect, throughput is the major challenge

3.3. (C) Data reuse

C1: Do you use or have you ever used data that you did not collect yourself?

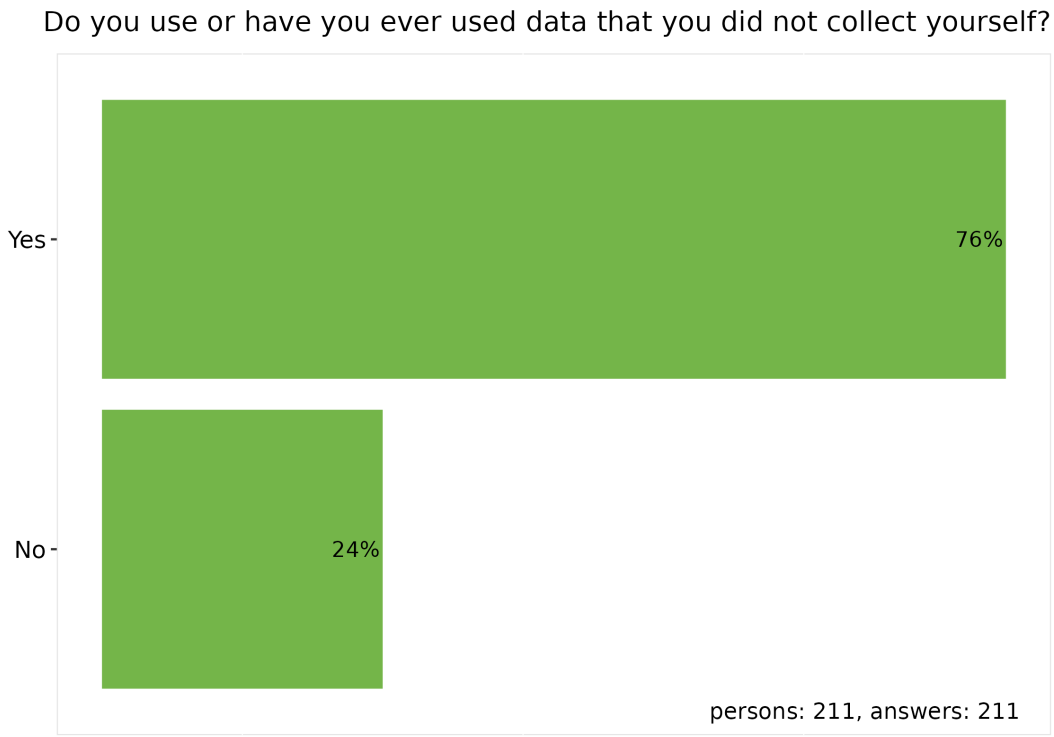

Figure 12: Results of mandatory single-choice question on the experiences of the participants in data reuse.

The following results only include the responses of the 161 people who already work with other people's data.

C2: What types of data have you primarily used in your example?

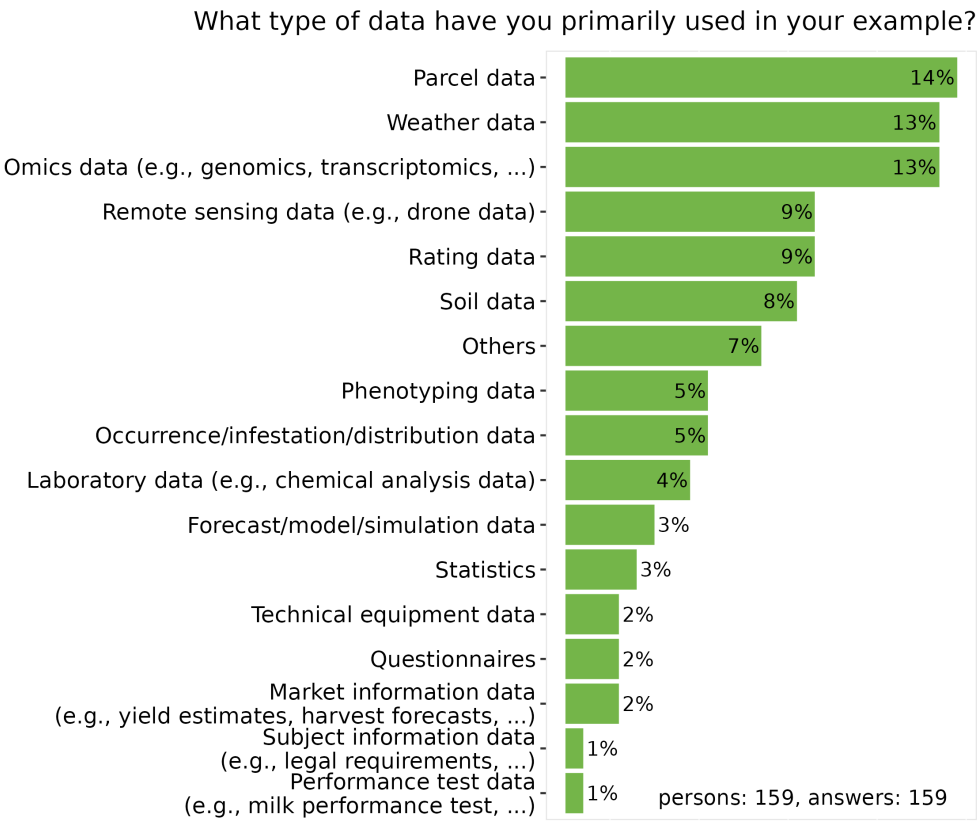

Figure 13: Results of single-choice question regarding which data categories are reused by the participants.

Additional answers in *Others* (n=11):

- Data from publications
- Almost all species shown
- Geodata
- Population data
- Weather data and occurrence data
- Metadata
- Laboratory data & phenotyping data
- Multiple selection would be helpful
- many different ones: Weather, impact, phenotypic data, omics data; technical equipment data. Here multiple choice would be important
- Administrative data
- Multiple selection is missing here: Field data, technical information data, soil data, statistics

C8: What type of data have you primarily used in your example?

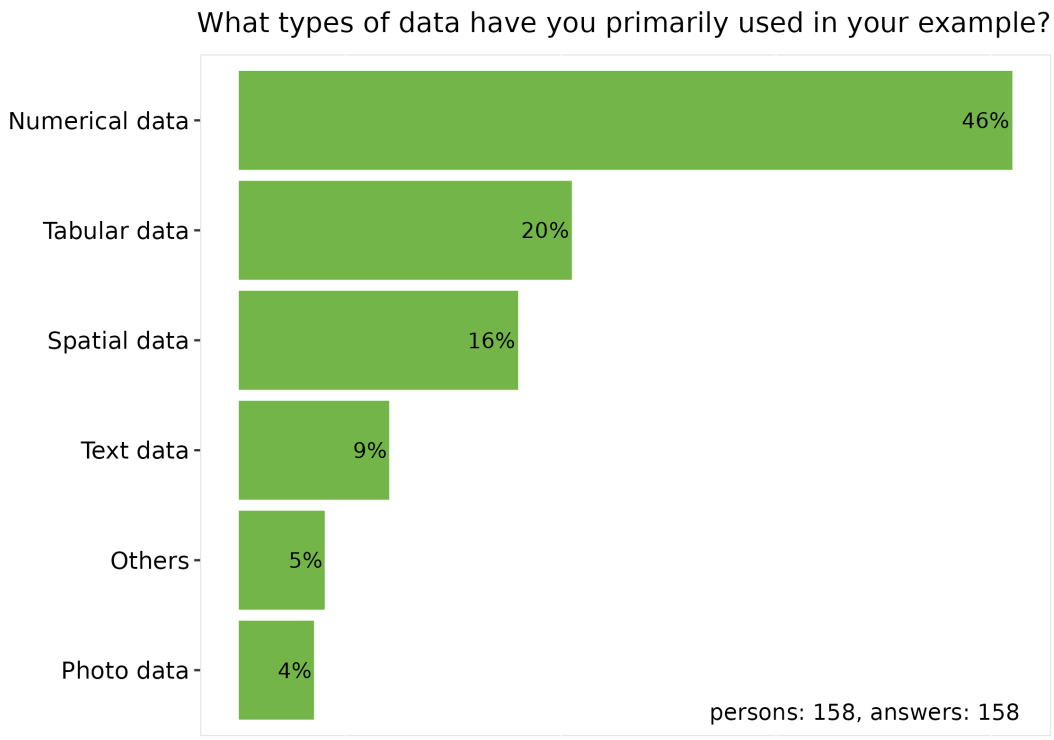

Figure 14: Results of single-choice question regarding which data types are reused by the participants. The response options 'audio data', 'video data', 'source codes' and 'non-digital data' were not selected and are therefore not listed in the figure.

Additional answers in *Others* (n=8):

- almost all species
- Sequence data (3)
- numeric and text
- photo, text, non-digital
- many criteria here too: Numerical data, text, photo, spatial; tabular, source codes
- Information on variant calling, normalised gene expression, genome sequence

C5: What is the purpose of your data reuse?

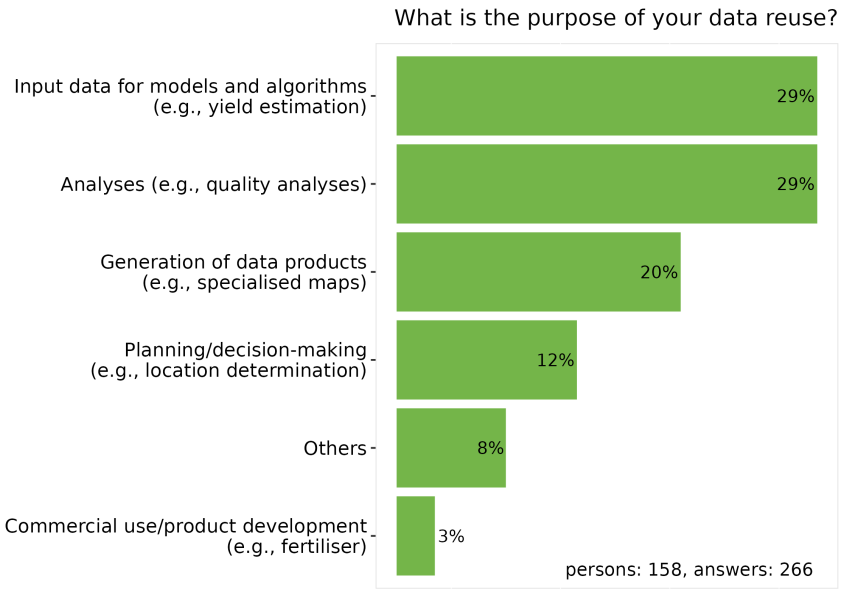

Figure 15: Results of multiple-choice question for the purpose of data reuse.

Additional answers in *Others* (n=20):

- Creation of a database
- point cloud for 3d semantic segmentation
- Enrichment of own data
- Calculation of indicators
- for comparison with own data (2)
- Population evolution
- Teaching
- QTL calculation
- Use of reference data; comparison of own data with published data
- Presentation in a lecture, by agreement
- Explanation for yield data
- Statutory reporting to the EU
- Communication to (future) importing countries
- Variety resistance evaluation, effectiveness
- Statistical analyses and quantitative estimates
- Evaluation of other data
- Surveys on pests
- Identification of orthologous and paralogous proteins.
- Enlargement of the database

C3: Where does this data come from? Please specify the data sources you have used.

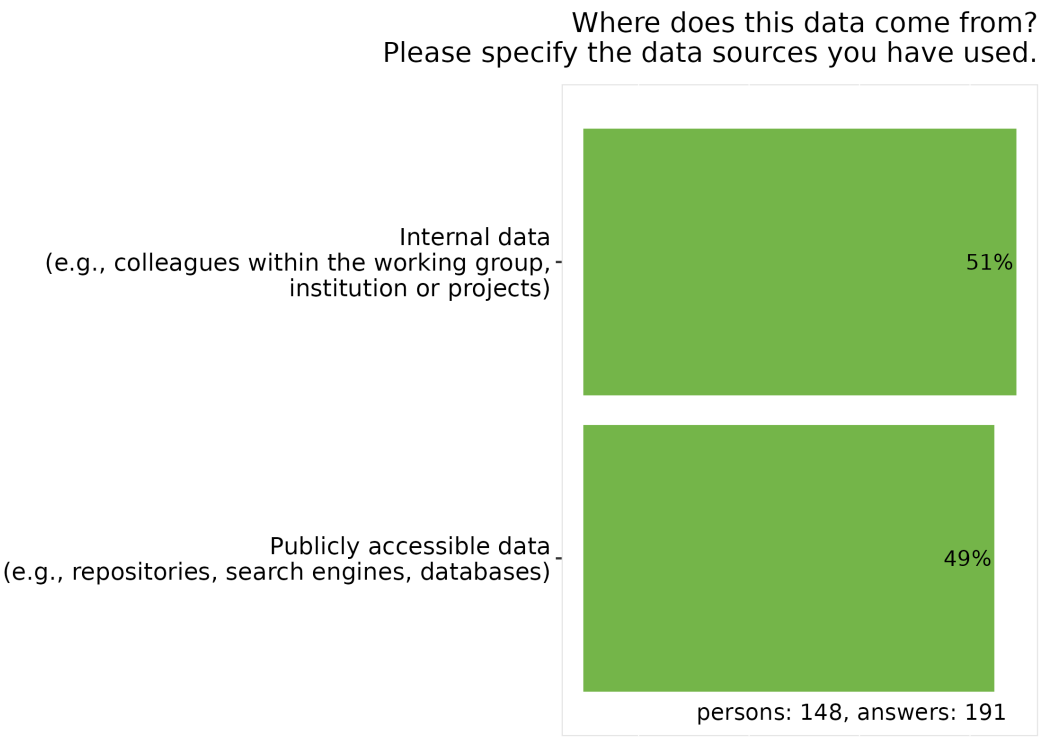

Figure 16: Results of multiple-choice question on usage of internal and external sources.

Additional answers for *Internal data* (n=72, unsorted):

- Various projects that are brought together
- CN, NIRS, ICP
- Soil data
- JKI internal weather data
- Official agricultural statistics surveys and measurements
- Experimental farm
- Soil data from FLI
- From related projects
- Measurement data from the site
- Project partners (2)
- Data products generated by a service provider
- Laboratory data (measurement data)
- State offices/state authorities (2)
- Chemical+microbiological data, metabolomics, microscopy-fluorescence

images yeasts+bacteria+leaf, photos of green growth

- Population surveys
- Colleagues within AG/institute/university (15)
- Internally collected statistics, e.g. prices for seeds, fertiliser, etc.
- Soil condition survey Germany
- Standardised data collection
- Predecessor projects
- Data collection of other
- JKI internal database (2)
- Excel tables of older colleagues
- Data collected by colleagues or project partners (4), e.g. field trial data incl. weather data, soil properties
- all mentioned
- Exchange within the framework of bi/multilateral co-operations
- Scoring data from cooperation projects,
- Plant protection services, JKI
- All of the examples given
- Research institution
- Evaluation of an internal database of JKI and plant protection services of the federal states
- Project partners and subcontractors (3)
- LLH
- Collection of field data by external parties within the framework of BMEL projects
- Student generated data
- Own assessments/data (2)
- Ministries of Agriculture
- Surveys from the agricultural administration, joint application
- GOV
- processed agricultural data
- Administrative data from the federal states
- Data from various projects/project periods
- Plant protection services, research institutes
- Technical staff
- Data from practice
- LSV, BSA
- Private research institutes; colleagues; peers
- cooperating breeding companies
- shared folder
- shared databases

Additional answers for *Public data* (n=71, unsorted):

- NCBI (6) [National Center for Biotechnology Information]
- Map server
- Weather data
- Public weather data
- ModelNet 40, an old school one
- Corine Landcover of the EU, WorldClim climate data, elevation models etc.
- NASA
- Federal and state offices
- Published data
- Federal Statistical Office
- Weather data (DWD) (5)
- Destatis
- https://esdac.jrc.ec.europa.eu/projects/lucas
- Information from state offices and specialised databases
- BKG
- Weather and soil data
- GBIF
- Transcriptomics databases, gene expression databases
- Weather data
- Pubmed, NCBI, Uniprot
- e.g. DSW2, FRED (IGB), Alkis
- GenBank, Q-Bank
- European Nucleotide Database
- Genesis online, FAOSTAT, EUROSTAT, Forest Condition Survey; DWD
- Soil estimation
- Agricultural statistics
- PS Info, hortigate
- Metadata of published studies
- USDA, US State College Ag Programmes
- Time series database
- DWD, Google Earth Engine, Copernicus Services, BGK
- BGR
- Field experimental data from publications, individual scientists, research institutions, open-data sources
- DWD, BGR (BÜK200)
- Recorded yields viticultural practice
- NCBI SRA
- all mentioned and publications
- e.g. sequence databases
- Weather data from DWD measurement networks and plant protection services
- Databases

- Sequencing Read Archive (SRA), ensembl, European Nucleotide Archive (ENA)
- NCBI GenBank
- mainly DWD
- https://cds.climate.copernicus.eu/
- BLE reports (https://www.ble.de/DE/BZL/Daten-Berichte/daten-berichte_node.html), https://trade.ec.europa.eu/access-to-markets/en/home
- IACS, LUCAS
- Geoportals
- FAOstat, ISRIC soilgrids, climate models from ISIMIP
- Search engine
- Agricultural statistics, promotion, specialised geodata
- Stat. Data (agricultural statistics)
- NHI- TCGA
- Libraries
- On request from the respective federal states and partly freely available.
- Soil data, soil maps
- Soil database
- BARLEX database IPK, James Hutton Institute, RAP database rice, MSU database rice, Ensembl, NCBI
- BonaRes repository
- Federal Statistical Office and BVL
- DWD (ftp server, cliamte data centre - download) ESA
- FLUXNET, ICOS

Additional answers in *Others* (n=8, unsorted):

- Service provider
- e.g. sequencing data
- Non-publicly accessible data from others
- Data collected by other authorities
- IACS
- Available in Germany on request and with luck or connections
- Invekos data
- Invekos data provided by the federal states

C4: What are the three criteria for data quality that should be met at a minimum for you to be able to reuse data effectively?

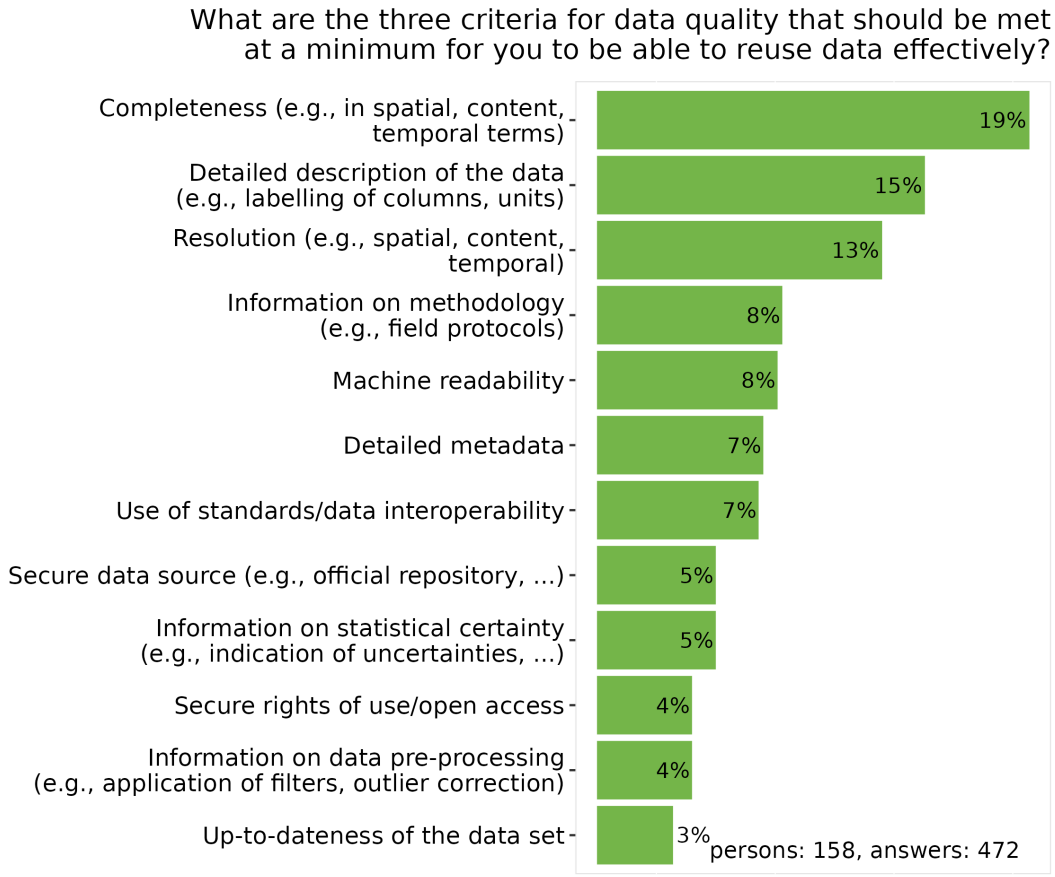

Figure 17: Results of ranking question for three out of twelve answers for the most important data quality criteria for the reuse of data.

C6: What would be reasons why you could not use the data?

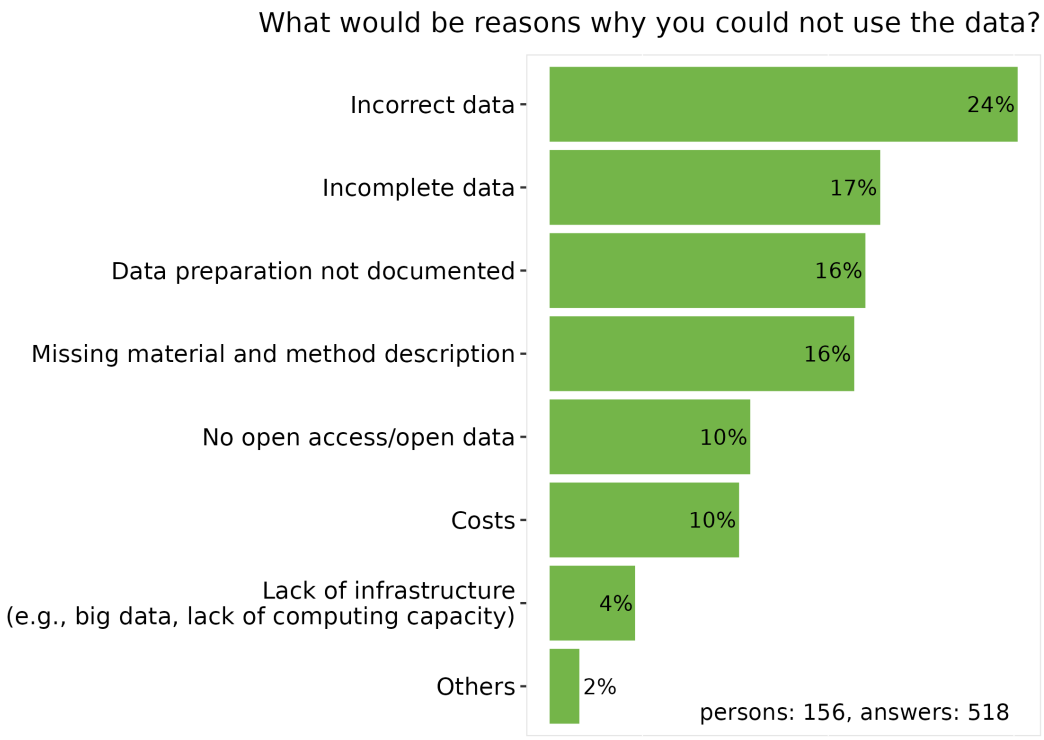

Figure 18: Results of multiple-choice question on possible reasons for not reusing data.

Additional answers in *Others* (n=8):

- Resolution too low
- Availability with easy-to-use user interface
- Changes in the method of provision
- No standardised data collection standards and subjectivity of scoring
- Not machine-readable (creatively formatted Excel or Word files)
- Comparability
- Lack of accessibility, data not digitised (resulting in poor readability)
- High acquisition and processing costs

C9: What steps would you take to make datasets usable for the desired purposes despite insufficient quality?

Figure 19: Results of multiple-choice question for subsequent improvement of data quality.

Additional answers in *Others* (n=6):

- Request missing information from the service provider
- own experience
- request authors
- Modelling approaches
- I mainly use genome sequence information. It cannot be used if the data is incomplete.
- If this type of data is incomplete, there is not much you can do.

C7: What are the biggest challenges for you in data reuse concerning data quality?

Freetext question on challenges in data reuse regarding data quality (n=66)

- Understand possible preprocessing of the data, there is usually missing metadata. Data (excel tables) usually do not match with reports, e.g. different number of samples...
- i only use good quality dataset that is published and opensource. so no challenge, coz they all good sorted. if it is bad quality, i will not use it. about how to measure if it is good or not, it should be good documented and structured, and data quality looks good (manural overview them).
- Data source must be reliable.
- Lack of standards and metadata
- Data quality and comparability with own data
- no metada available, small sample size
- Traceability of decision-making processes
- Lack of information on data quality, including the comparability of data sets
- Methodology used for obtaining the data, reproducibility
- Too many systems, each with their own requirements, no 'easy to use' output system, data constantly requires different specialised knowledge
- Hardly any validation by experts, no clear quality standards
- No standardised data collection criteria, incompleteness
- I have no idea about it.
- Trust in the institution providing the data
- Access to data collected by the individual federal states. Incompleteness of the data. Data time series too short
- Data quality is often difficult to assess at first glance. It takes a lot of time to assess/check whether the data quality is satisfactory.
- The data I need for my work is not available in the appropriate resolution or the costs are too high.
- Data correctness
- Background information/metadata
- Missing/inaccurate data labelling, careless errors during data entry
- Documentation and accessibility of data
- Out of date data, lack of locality detail
- Data errors not checked
- Missing data on key system components

- Introduction and use of metadata standards
- Traceability of data collection,
- Insufficient description of data (QTL calculation) and insufficiently accurate documentation and collection of yield data (viticultural practice)
- Assessment of the reliability and completeness of the data in surveys of agricultural management and balance sheet data from practical farms
- Non-transparent data collection methods, poor collection protocols including incomplete data.
- Reliability of the data and the conclusions based on it.
- The time to deal with external data and their structures, which does not always lead to a result. In third-party funded projects, it is generally not intended to base research solely on the data of others.
- Description of the methodology and labelling
- In relation to drone data, for example, this would be: lack of documentation of the survey methodology lack of information on resolution, flight altitude, date of recording, crop on the field lack of information on the technical parameters of the drone (type of sensor, drone type, etc.)
- Data quality strongly dependent on the operator.
- Partly different units that are difficult to convert. Lack of standards for set of parameters/results that should be published in all cases (even if the focus was not on this defined 'standard data').
- Authenticity of the data
- statistics
- Missing labelling of columns or labels that are not self-explanatory, but there is also no data dictionary for them.
- Comparability of data and summarisation of data according to a given standard
- Resolution - temporal and spatial
- Missing information about the mapped values. No metadata.
- Comparability; at first glance, identical information is usually assigned different figures. Unclear what the differences are and what figure could be used for my use.
- Obtaining complete data sets.
- The creativity for incorrect entries is infinite
- Data not freely available (open access), e.g. IACS
- Use of Excel as data entry form

- Incomplete description of the survey methodology, incomplete metadata, lack of harmonisation, incomplete validation
- Interoperability between data-generating, proprietary systems
- Machine readability interoperability
- Standardised responses
- Lack of standardised process of data management and manipulation
- Insufficient labelling of columns, units etc. and varying formatting of Excel tables in different years. This makes the actual analysis very time-consuming because everything has to be formatted, transferred etc. first.
- Missing metadata, missing standards for data content, types, structuring, data protection, machine readability or missing data interfaces
- The poorer the data quality, the longer it takes to process and the more assumptions have to be made.
- Data protection
- Increasing data entry time with standardised forms - it is not possible to document/write down everything 'for others' that the expert on site takes into account in the interpretation from experience; there is no guarantee that data was really collected as discussed - major source of error for incorrect derivations; lack of IT equipment and freedom in the public sector for cloud/software/computer performance, etc.
- Standardisation, completeness and correctness of the data
- For my chosen example, the different countries have different standards for how they provide the data sets. In addition, there is hardly any documentation on the data quality of the data and hardly any metadata standards.
- Digitisation of written or oral data
- For the datatype I use, one of the biggest challenge is the inconsistent nomenclature for updated information on genotyping data.
- Related to the data mentioned above: It was difficult to find unique identifiers for the respective genes/proteins as it was shotgun sequencing data and hypothetical proteins based on it.
- The processing pipeline should be customized for each dataset due to a lack of standardisation
- Standardisation of datasets (e.g. across multiple environments), data completeness
- Data preparation
- I also use secondary data if primary data is missing. There is enough general data. Unfortunately, there is not always the specific secondary

data that I need (specific location/work step/machines/countries).

- The lack of knowledge about the quality of the data and missing or insufficient documentation.

3.4. (D) Data reuse in practice

D1: Imagine you wanted to use a new dataset. During your research, you find not only the metadata, but also information on how well (or poorly) the dataset was already usable in a similar use case. I find this information:

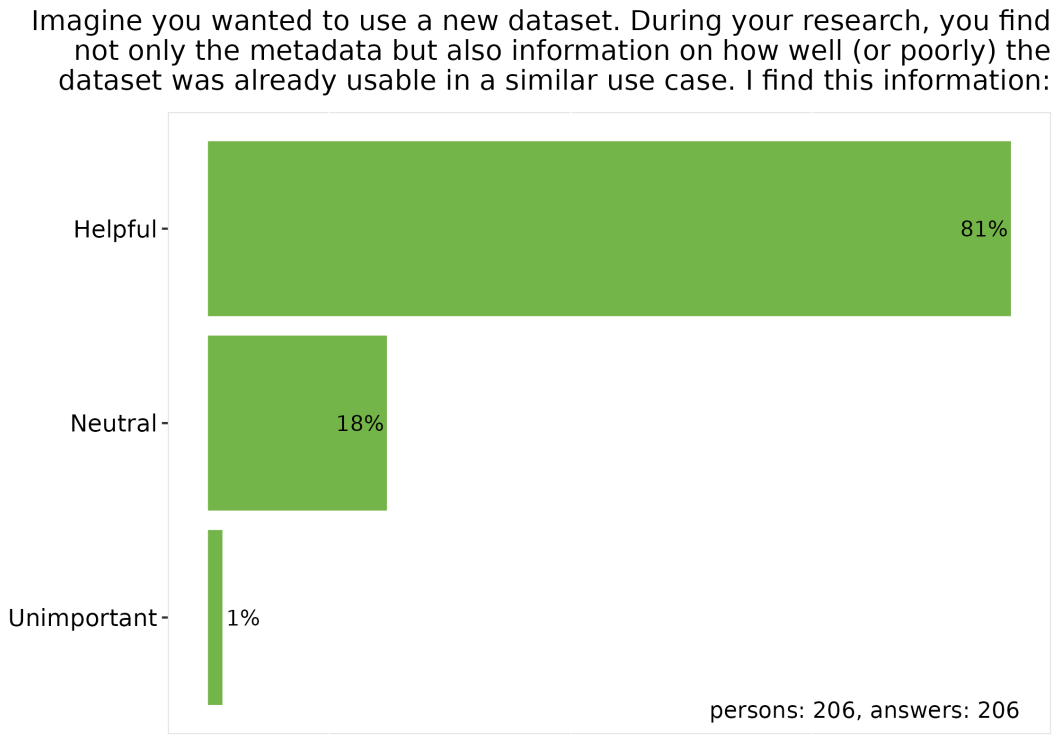

Figure 20: Results of single-choice question on the helpfulness of information on the usability of data sets.

D4: Would you be willing to invest time in documenting completed applications of a dataset in its metadata?

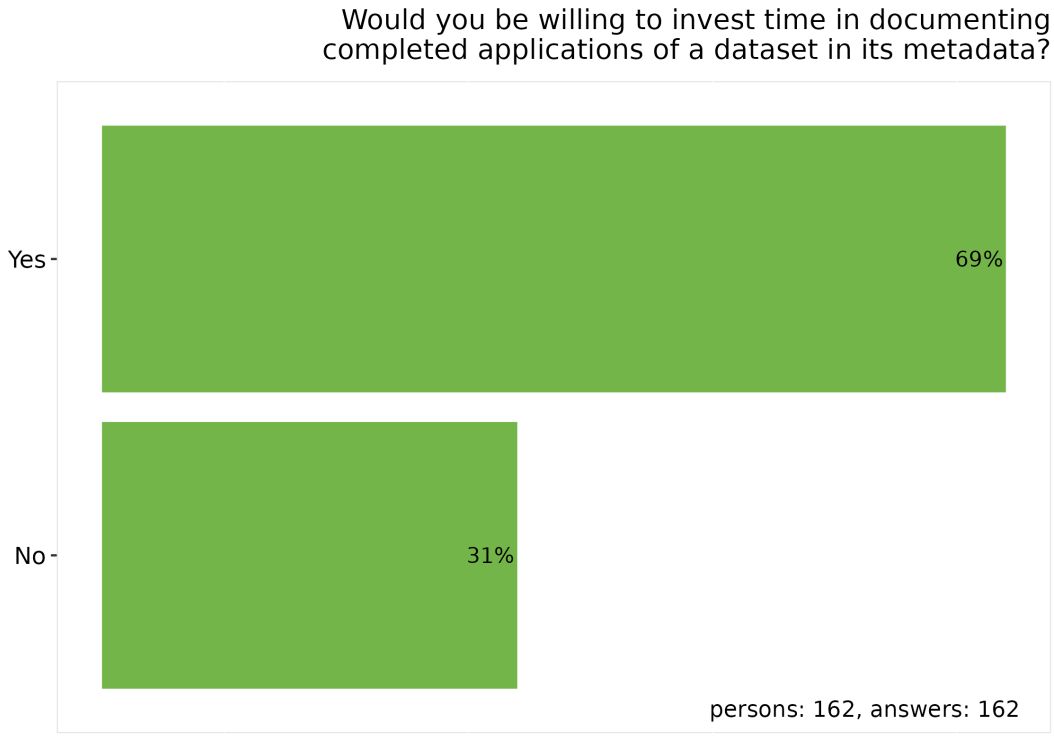

Figure 21: Results of single-choice question on the willingness of participants to invest time in the documentation of a completed application of a data set.

D5: How much time (in minutes) would you be willing to invest to document the successful application of a data set in its metadata?

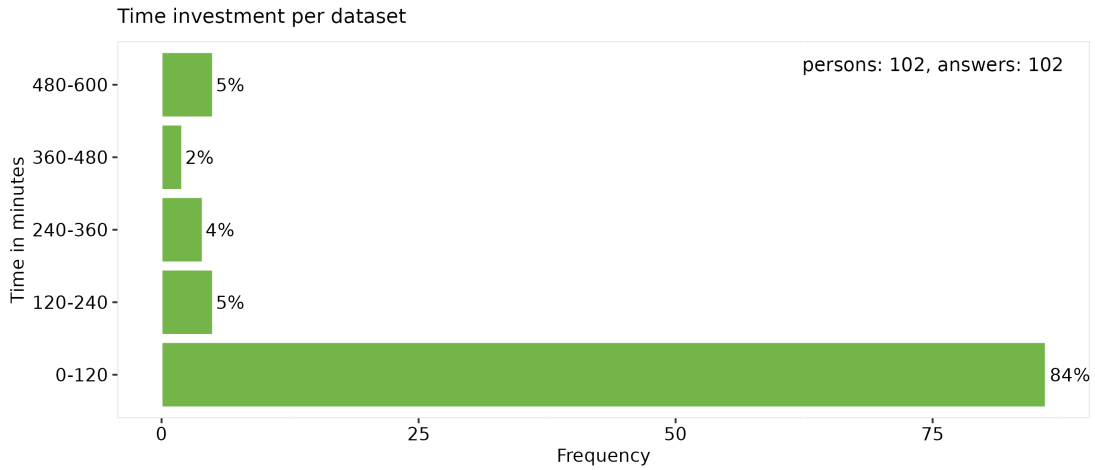

Figure 22: Time that participants would be willing to invest to document the successful application of a data set. For ease of reading, the individual numbers given in minutes were summarised in blocks of two hours.

D2: Have you ever compared different input data for a use case before computation or for a model (or similar) to choose the 'best' dataset to use?

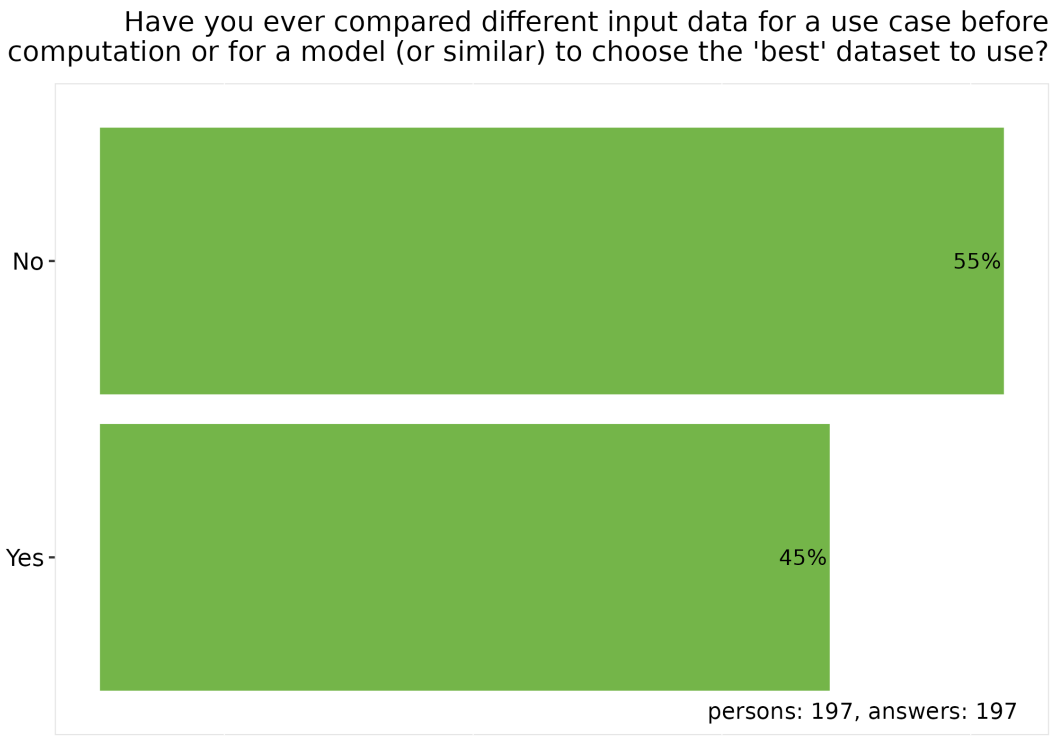

Figure 23: Results of single-choice question, about the participants' experience in selecting a 'best' data set.

D3: How did you select the 'appropriate' dataset?

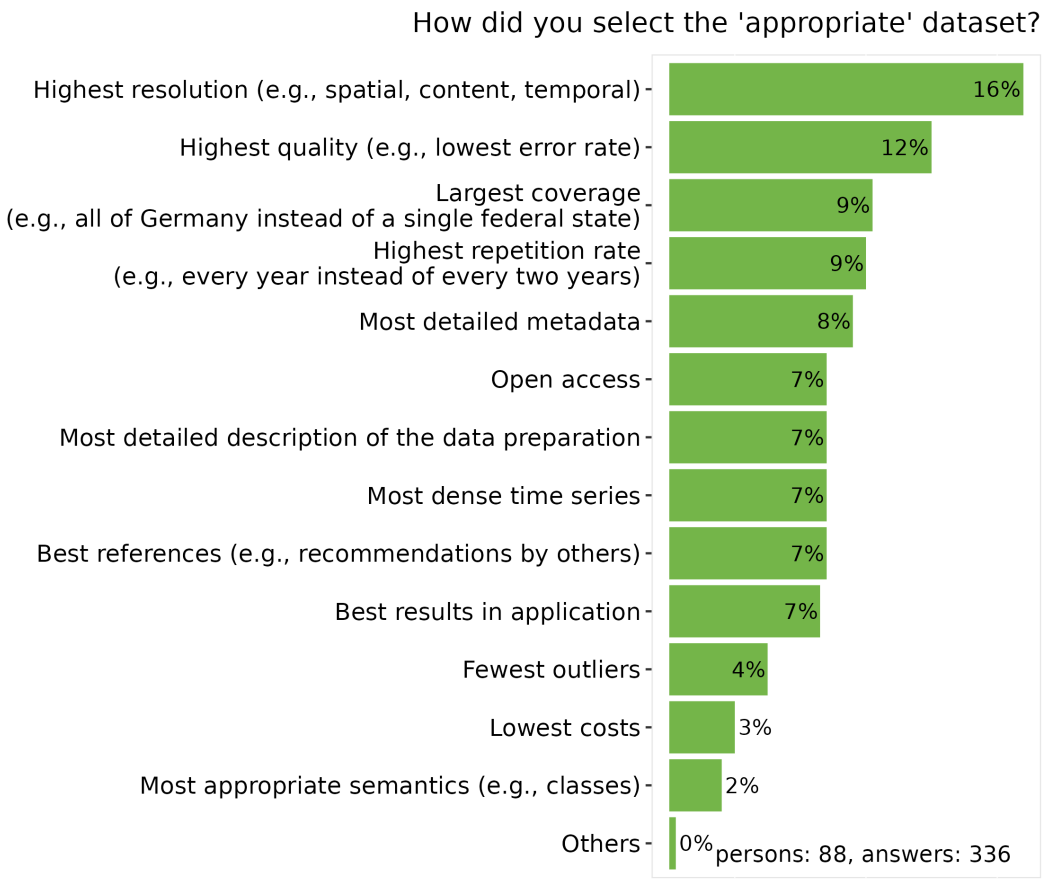

Figure 24: Result of multiple-choice question regarding how participant choose the 'appropriate' dataset for themselves. The response option 'Smallest file size' was not selected and is therefore not listed in the figure.

Additional answers in *Others* (n=1): Plausibility, stability of the results

3.5. (E) More on the topic of data quality

E2: Are there any additional information or thoughts on the topic of data quality that you would like to share with us?

Answers of free text question about additional thoughts on data quality (n=37):

- I find it difficult to judge data quality, because I think for publications usually already cleaned data is used, so maybe the raw data was not optimal but could be cleaned to a good dataset. I would give the dataset a good data quality, but maybe it wasn't originally like that.
- Different data type's quality need to be measured by different methods.
- Statistical measures would be important. Origin of the data is crucial.
- In my view, the most important point is that it is still almost impossible to reproduce research data from the scarce M&M part of a publication. Such a seal of approval (e.g. confirmation of experimental data by truly independent replicates, ideally in other laboratories) would make it much easier to reuse data. A good description of the data and the purpose of the data collection helps to assess the data quality for your own application. This data is often missing. Unfortunately, a large proportion of data sets are usually excluded from consideration because they are incomplete for your own use case. This can very easily lead to a distortion of the data situation.
- I overlooked the following aspects: i) Ensuring the reproducibility of results, emphasizing the need for not just replicating final outputs (e.g., graphs for an article) but also documenting the detailed steps involved in data preprocessing. This underscores the significance of utilizing programming languages like R or Python to script all steps comprehensively, from raw data to processed datasets. ii) Highlighting the crucial role of proper data version control, including details such as date, authorship, and data origin, to maintain a systematic record of dataset cycle.
- My example came from the behavioural ecology (especially visual ecology) of insect pests. This questionnaire was probably aimed more at the agricultural context. For us it is particularly important:
  - The experimental design (must be well documented)
  - Controlled conditions or (documentation)
  - Measure everything that cannot be controlled as well as possible (with sensors) (documentation)
  - Good selection of co-variables, if necessary with stat. Methods

(statistics, biological relevance)

  - Model building and model evaluation using well selected, possibly standardised and biologically informed data sets (physiology etc.)

- 1. most data in the 'agricultural sciences' are not detailed enough, e.g. QTL are written as a location, but there is usually not even an overview of at least presumably interesting genes there. 2. data consist only of gene IDs, but almost never of or with functional annotations, all of which can be obtained. . . . but the GO terms or protein function prediction would make such transcript data (as an example) sortable - ah, all calcium binding proteins, I have often been able to quickly generate good new hypotheses in the modelling system with them

- Even if research institutions have great data management concepts on paper, they fail when it comes to practical implementation (no clear guidelines for collecting, validating and storing research data, lack of infrastructure, insufficient computing capacity).

- I am shocked at how little value is placed on actual data quality and its validity in our institution, although there is talk of 'good scientific practice' etc. everywhere. However, these quality standards are not adhered to in practice and if you point this out, you are told that everyone does it that way.

- For my project, I am working with high-resolution geodata, which must have a resolution of at least 50 metres, preferably even lower. Unfortunately, the availability of a wide range of geodata (soil data, topographical data, erosion data, etc.) is quite poor. Only low-resolution data is usually freely available. High-resolution data is very expensive or not available at all.

- Application: Form of data in e.g. databases that change and are sometimes no longer applicable.

- A first helpful impression on data quality can be achieved via a simple visualization of data sets by looking for plausibility, completeness, outliers, trends, variable comparisons and logic dependencies of variables.

- In my work, it is very important to achieve the most accurate and reliable assessments possible, as the consequences of my assessment can be very far-reaching. The better the quality and availability of the data, the more reliable the assessment can be.

- Data quality depends on the persons who generate data!! This can be improved if useful templates for data entry would exist

- The idea of FAIR is good, but it means that some researchers have to invest a lot of time and effort in data documentation so that statistics-oriented

researchers and modellers can use it - without any benefit.

- This questionnaire is true, but probably invalid. I filled it out anyway at her insistence.
- I find it most important to have clarity about the content and quality of the data. If a dataset is not complete or is likely to be of poor quality in parts, this is not an exclusion criterion for me if it is clearly labelled. Because then I may still be able to use the data set in a reduced form.
- I would like to recieve more information about this
- There would need to be a data codex that is taught at German universities or graduate schools. Many current university teachers are not competent to teach this, but need staff who know and can apply the standards.
- Like in Cochrane reviews, publication of the financial support of the data research by which donators.
- Compare similar or different datas to the topics.
- A standardised method of data collection and evaluation by the persons collecting the data is necessary. This requires appropriate standards and digital recording and analysis options.
- I sometimes lack the basic knowledge to be able to provide metadata adequately, as I generally do not generate data, but only analyse data or use analysed data. Perhaps information material on this could be made available.
- It is extremely important and helpful for the analysis to train the data collectors and to provide clear specifications, preferably in the form of selection lists, for the main components such as harmful organisms and pesticides
- The survey is generic - there is no other way. However, it would be helpful to base the answers on specific examples, as different participants may mean different things with the same statements.
- I'm not quite sure what exactly is meant by 'data quality' in this survey. On the subject of data utilisation, I would like to point out that it would be very nice for many questions if data, e.g. on animal health on farms, were accessible for science (and possibly also for politics). However, data protection regularly slows us down here. Within the university, research data management is still in its infancy and there is still a lack of experience with preregistration or data repositories
- Mixed questions about quality - quality of the measurement process (data collection) vs quality of the equipment (sensor/equipment resolution). These are two different aspects of data quality.
- We would like to see more and more

standards and centralised databases and data interfaces created in Germany in order to finally reduce the effort involved in acquiring data and making it usable, thus enabling efficient and targeted scientific work.

- Inconsistent classes of animal data are very annoying. In the (frequently occurring) extreme case, it is only possible to compare 'pigs' or 'cattle' when comparing two data bases, because finer subgroups are selected incomparably.
- Please do not create a construct that has many noble goals but makes scientific data collection and analysis terribly cumbersome!
- A plausibility check of the data should precede any data analysis.
- The areas of livestock sciences and WiSoLa of agricultural sciences were not sufficiently considered in the response options (plant focus) in this survey.
- All extra work that will be required in the future to document and evaluate work/results must be as little additional time as possible. Therefore, tools that are (semi-)automated and made available to researchers would be useful. For example, the automated creation of metadata files or the identification of quality standards would be helpful.
- Rather the problem of external data. Multi-year field histories obtained by the farmer. The problem here is less one of quality and more one of digitisation. Very important: what quality is needed at all and what is technically and economically necessary? What data is needed at all. Rather the problem of drilling the right 'thick' boards.
- Genotype data was not mentioned.
- If there was some kind of certification label, at least for state actors (authorities, universities, etc.), so that you can see at a glance whether the data has been checked or that a certain standard has been followed in its creation.

Acknowledgements

We would like to thank all participants for their willingness to take part in the survey as well as all the supporters and pretesters who enriched the survey with their answers and ideas, in particular: F. Beyer, P. Brandt, D. Gabriel, S. Gedicke, C. Hoffmann, O. Kirchgeßner, D. Martini, S. Risvi, M. Senft, U. Stahl

This work was created as part of the NFDI consortium FAIRagro (www.fairagro.net). We gratefully acknowledge the financial support of the German Research Foundation (DFG) – project number 501899475.

A. Appendix

Original answers to 'Others' and free text questions.

Appendix A: General information about the participants

A1: In your everyday handling of data, which group do you most closely identify with?

Original answers in *Others* (n=4): Qualitätsüberwachung, Daten?, Wikipedia, Wissenschaftlerin

A2: Which of the following groups do you primarily identify with?

Original answers in *Others* (n=14): „Als Informatiker ein Zwischending zwischen techniker und Wissenschaftler", MSc (wiss.-technische MA), wissenschaftliche Mitarbeitende, wissenschaftliche Mitarbeiterin, Scientific staff, Farmer, scientific staff, other (pre-/non-PhD, Forschungskoordinatoren, Angestellte, wissenschaftliche Mitarbeitende ohne Promotionsvorhaben, Agrarverwaltung, wissenschaftliche Mitarbeitende ohne Promition im höhren Dienst, Privatwirtschaft, Wissenschaftlicher Mitarbeiter

A3: Which of the following institutions do you primarily identify with?

Original answers in *Others* (n=3): Farm, Außeruniversitäre Forschungseinrichtungen, Landwirtschaft

A4: To which of the following DFG subject groups can your current work be assigned?

Original answers in *Others* (n=23): Biodiversität, Pflanzenschutz (4), Molecular mycology, Informatik, Lebenswissenschafte, Vorratsschutz, Maschinenbau, Agrarökosystem-Modellierung, Entomologie (Zoologie ist zu allgemein), Wasserressourcen, Verwaltung und e-Mails, Agrarentomologie, Life science/food/beverage, Entomologie, pflanzengesundheitskontrolle, Hydrologie, Pflanzengesundheit, data management, Bioökonomie, Fernerkundung, Geodatenanalyse

Appendix B: Data collection

B2: For what type of data did the data collection of your example primarily take place?

Original answers in *Others* (n=6): für verschiedene der aufgeführten Arten; omics, phenotying and lab data; hier wären Mehrfachantworten für mich möglich gewesen; Energy/Carbon Footprint Data; Verhalten/Gesundheit von Tieren; Betriebswirtschaftliche Daten

B8: For what types of data did the data collection in your example take place?

Original answers in *Others* (n=10): Spektrometrische daten; Frage unklar und keine Kombinationen möglich (z. B. sind die meisten Felddaten, Fotos, etc. auch räumliche Daten); Sequenzdaten; Mulitispektraldaten; DNA-Sequenzen; Multiple: Text, audio, numerical, tabular; numerisch und Text; text and photo; gleichwertige Kombination von Video, numerischen Daten, räumlichen Daten; hier fehlt leider die Mehrfachauswahl: nicht digitale Daten und numerische sowie Fotodaten

B3: For what application area did the data collection in your example primarily take place?

Original answers in *Others* (n=17): Chemische Ökologie, Pflanzenschutz (4), Molecular mycology, Vorratsschutz, Gewässerökologie, Agrarwissenschaften, Sequenzdaten, Entomologie (2), Agrarentomologie, Engieering Data, Electronic equipment, Medizin, Data management purposes

B7: Was there a relevant data quality criterion missing in the previous question? If yes, which one?

Original text answers (unsorted, n=28):

- I use RMSE
- Korrektheit
- Annotation der Daten. da häufig nicht nur erfasste Daten sondern auch bereits berechnete Daten vorliegen.
- Auch die anderen Vorschläge halte ich für sehr wichtig, um die Qualität und Nutzbarkeit der Daten sicherzustellen!
- Informationen zur Methodik, Umfangreiche Beschreibung der Daten
- size of data (number of samples)
- soweit möglich, Berücksichtigung von Referenzmessungen und unabhängige Wiederholungen
- reproducibility
- Metadaten finde ich auch sehr wichtig

- ausreichende Validierung der Daten, zum Beispiel fehlend bei GenBank
- Bias
- Sind die Daten repräsentativ (Basis für generallisierbare KI-Modelle)
- Locality of data
- Innere Konsistenz der Datensätze, Markierung der Daten mit Qualitätsflags
- quality of measurements
- Ich hätte gerne die Möglichkeit mehrere Beispiele zu nennen und zu Beschreiben da ich eine Vielzahl der Datenarten erhebe
- Treffsicherheit der erhobenen (Bonitur)daten
- Nutzbarkeit (also dass wenn Daten vorhanden sind, diese auch genutzt werden dürfen und nicht aus Datenschutzgründen nicht verwendet werden dürfen, z.B. Wirtspflanzenverfügbarkeit in allen Bundesländern)
- Anzahl der Wiederholungen einer Meßreihe
- Vollständigkeit
- Nein.
- Maschinenlesbarkeit
- Vergleichbarkeit (bei der Erstellung von Fotos immer dieselben Bedingungen schaffen: Hintergrund, Beleuchtung, Kameraeinstellungen und Arbeit mit Stativ,. . .)
- Frage nicht leicht zu beantworten, weil t 100% klar, worauf sie ab zielt
- Reproduzierbarkeit
- Vergleichs/Kontrollanalysen/messungen
- Ja, denn man erhebt nie nur eine Art von Daten, wie man auswählen musste, sondern braucht idR mehrere Daten für die sinnvolle Interpretation.
- Plausibilitätsprüfung

B5: How do you ensure the quality of your collected data?

Original answers in *Others* (n=8):

- Verwendung von Negativkontrollen
- Reviewprozess
- Wiederholungen
- Reflexively, transdisciplinary practices e.g. participant validity checks and co-creation of knowledge
- maschinelle Datenauswertung der Fotos: definierte Farbanteile

- Verblindung, Randomisierung, Beobachterabgleich, Plausibilitätsprüfung
- Modellansätze
- weitgehende “Digitalisierung” des gesammten Datenproduktionsprozesses -keine Papiereingaben.

B6: What are the biggest challenges for you in data collection regarding data quality?

Original answers in *Others* (n=5):

- mangelnde Infrastruktur am JKI (Computerpower etc)
- Betriebsblindheit
- fehlende Objektivität bei manueller Erhebung
- Protokolle werden nicht eingehalten
- For the data we collect, throughput is the major challenge.

Appendix C: Data reuse

C2: What types of data have you primarily used in your example?

Original answers in *Others* (n=11):

- Daten aus Publikationen
- fast alle dargestellten Arten
- Geobasisdaten
- Populationsdaten
- Wetterdaten und Vorkommensdaten
- Metadaten
- laboratory data & phenotyping data
- Mehrfachauswahl wäre hilfreich
- sehr viele verschiedene: Wetter, Schlag, Phenotypic data, Omics data; technische Geräte Daten. Hier wäre mehrfachnennung wichtig
- Verwaltungsdaten
- hier fehlt Mehrfachauswahl: Schlagdaten, Fachinformationsdaten, Bodendaten, Statistiken

C8: What type of data have you primarily used in your example?

Original answers in *Others* (n=8)

- fast alle Arten
- Sequenzdaten
- Sequenzdaten
- numerisch und Text
- photo , text. mon digital
- Sequenzdaten
- auch hier viele Kriterien: Numerische Daten, Text, Foto, Räumliche; tabellarische, Souce codes
- Information on variant calling, normalized gene expression, genome sequence

C5: What is the purpose of your data reuse?

Original answers in *Others* (n=20):

- Erstellung einer Datenbank
- point cloud for 3d semantic segmentation
- Anreicherung eigener Daten
- Berechnung von Indikatoren
- zum Vergleich mit eigenen Daten
- Vergleich zu eigenen Daten

- Populationsevolution
- Lehre
- QTL Berechnung
- Nutzung von Referenzdaten; Vergleich eigener mit publizierten Daten
- Darstellung in einem Vortrag, nach Absprache
- Erklärung für Ertragszahlen
- Gesetzliche Berichterstattung an die EU
- Kommunikation an (zukünftige) Importländer
- Sortenresistenzbewertung, Wirksamkeit
- Statistische Analysen und quantitative Abschätzungen
- Bewertung anderer Daten
- ERhebungen zu Schädlingen
- Identifizierung von orthologen und paralogen Proteinen.
- Vergrößerung der Datenbasis

C3: Where does this data come from? Please specify the data sources you have used.

Original answers for *Internal data* (unsorted, n=72):

- verschiedene Projekte, die zusammengeführt werden
- CN, NIRS, ICP
- soil data
- JKI interne Wetterdaten
- official agricultural statistics surveys and measurements
- Versuchsbetrieb
- Bodendaten vom FLI
- aus verwandten Projekten
- Erfassung von Kollegen zur Weiterverarbeitung
- Messdaten vom Standort
- Projekt intern
- von einem Dienstleister generierte Datenprodukte
- Labordaten (Messdaten)
- Landesämter
- Chemische+ mikrobiol. Daten, Metabolomics, Mikroskopie- Fluoreszenzaufnahmen Hefen+Bakterien+Blatt, Fotos Begrünungsaufwuchs
- populationserhebungen
- Kolleg:innen innerhalb AG
- intern erhobene Statistiken, z.B. Preise für Saatgut, Dünger etc.
- Bodenzustandserhebung Deutschland
- Projektpartner

- Aufnahmen von Drohnenflügen durch einen Kollegen
- JKI interne Daten; Daten die von Landesbehörden zur Verfügung gestellt wurden
- colleagues within the working group
- standardisierte Datenerhebungen
- andere AG der Uni
- Kolleg:Innen innerhalb der AG
- Vorgänger-Projekte
- Datenerhebung anderer
- andere hochschulinterne AG
- JKI interne Datenbank
- Wetterdatenabfrage aus FG Agrarmeteorologie der gleichen Einrichtung
- Excel Tabellen älterer Kollegen
- AG interne Erhebungen von Kollegen
- Daten von Kollegen oder Projetpartnern erhoben, z.B. Feldversuchsdaten incl. Wetterdaten, Bodeneigenschaften
- Project/ colleagues
- alle genannten
- Austausch im Rahmen von bi/multilateralen Kooperationen
- Boniturdaten aus Kooperationsprojekten,
- Drohnenaufnahmen anderer Kollegen, anderer Abteilungen, oder Aufnahmen externer Dienstleister
- Kolleg:innen von anderen Instituten
- Ehemalige Kollegen
- Kollegen
- Kolleg*innen des Instituts, Projektpartner
- Pflanzenschutzdienste, JKI
- alle der angegebenen Beispiele
- Forschungseinrichtung
- Auswertung einer internen Datenbank von JKI und Pflanzenschutzdiensten der Bundesländer
- Projektpartner und Unterauftragnehmer
- LLH
- Erhebung von Schlagdaten durch Externe in Rahen von BMEL-Projekten
- Student generated data
- eigene Bonituren
- Landwirtschaftsministerien
- Erhebungen aus der Landwirtschaftverwaltung, gemeinsamer Antrag
- GOV
- von Kolleg:innen des eigenen Instituts
- aufbereitete Agrardaten

- Verwaltungsdaten der Länder
- Institutskolleg:innen, Projektpartner:innen
- Daten aus verschiedenen Projekten/Projektzeiträumen
- Projektpartner
- Pflanzenschutzdienste, Forschungsinstitute
- technisches Personal
- Daten aus der Praxis
- LSV, BSA
- eigene Daten, die über die letzten Jahre generiert wurden
- Private research institutes; colleagues; peers
- kooperierende Zuchtfirmen
- Teil eines Großprojektes eines Kollegen
- shared folder
- geteilte Datenbanken
- Projektpartner

Original answers for *Public data* (unsorted, n=71):

- NCBI (6) [National Center for Biotechnology Information]
- Kartenserver
- weather data
- öffentliche Wetterdaten
- ModelNet 40, an old school one
- Corine Landcover der EU, WorldClim-Klimadaten, Höhenmodelle etc.
- NASA
- Bundes- und Landesämter
- Publizierte Daten
- Statistisches Bundesamt
- Wetterdaten (DWD) (5)
- destatis
- https://esdac.jrc.ec.europa.eu/projects/lucas
- Information von Landesämtern und fachspezifischen Datenbanken
- BKG
- Wetter und Bodendaten
- GBIF
- transcriptomics databases, gene expression databases
- Wetterdaten
- Pubmed, NCBI, Uniprot
- z.B. DSW2, FRED (IGB), Alkis
- GenBank, Q-Bank
- European Nucleotide Database
- Genesis online, FAOSTAT, EUROSTAT, Waldzustandserhebung; DWD

- Bodenschätzung
- Agrarstatistiken
- DWD DataHub
- PS Info, hortigate
- Metadaten publizierter Studien
- USDA, US State College Ag Programs
- Zeitreihendatenbank
- DWD, Google Earth Engine, Copernicus Services, BGK
- BGR
- Field experimental data from publications, individual scientists, research institutions, open-data sources
- DWD, BGR (BÜK200)
- Erfasste Erträge Weinbaupraxis
- NCBI SRA
- alle genannten sowie Veröffentlichungen
- z.B. Sequenz-Datenbanken
- Wetterdaten aus Messnetzen des DWD und der Pflanzenschutzdienste
- Datenbanken
- Sequencing Read Archive (SRA), ensembl, European Nucleotide Archive (ENA)
- NCBI GenBank
- https://cds.climate.copernicus.eu/
- BLE-Berichte (https://www.ble.de/DE/BZL/Daten-Berichte/daten-berichte_node.html), https://trade.ec.europa.eu/access-to-markets/en/home
- IACS, LUCAS
- Geoportale
- FAOstat, ISRIC soilgrids, Klimamodelle aus ISIMIP
- search engine
- Agrarstatistik, -förderung, Geofachdaten
- stat. Daten (Agrarstatistik)
- NHI- TCGA
- Bibliotheken
- Auf Anfrage bei den jeweiligen Bundeländern und teilweise frei verfügbar.
- Bodendatren, Bodenkarten
- Datenbank
- BARLEX database IPK, James Hutton Institute, RAP database rice, MSU database rice, Ensembl, NCBI
- BonaRes repository
- Statistisches Bundesamt und BVL
- Deutscher Wetterdienst
- DWD (ftp server, cliamte data center - download) ESA
- FLUXNET, ICOS

Additional answers in *Others* (n=8, unsorted):

- Dienstleister
- z.b. Sequenzierungsdaten
- Nicht öffentlich zugängliche Daten anderer
- Erhebungsdaten durch andere Behörden
- InVeKoS
- Erhält man in D auf Anfrage und mit Glück bzw. Beziehungen
- Invekos Daten
- von den Ländern bereitgestellte Invekos Daten

C6: What would be reasons why you could not use the data?

Original answers in *Others* (n=8):

- Zu geringe Auflösung
- Verfügbarkeit mit leicht zu nutzendem user interface
- Änderungen in der Art der Bereitstellung
- Keine einheitliche Datenerhebungsstandards und Subjektivität der Bonitur
- nicht maschinenlesbar (kreativ formatierte Excel oder Word-Dateien)
- Vergleichbarkeit
- mangelnde Zugänglichkeit, Daten nicht digitalisiert damit einhergehend mangelnde Lesbarkeit)
- hoher Akquise und Aufbereitungsaufwand

C9: What steps would you take to make datasets usable for the desired purposes despite insufficient quality?

Original answers in *Others* (n=8):

- keine
- fehlende Informationen beim Dienstleister anfragen
- eigene Erfahrung
- Autoren anfragen

- keine
- Modellansätze
- I mainly use genome sequence information. It cannot be used if the data is incomplete.
- Wenn diese Art von Daten unvollständig sind, kann man nicht viel machen.

C7: What are the biggest challenges for you in data reuse concerning data quality?

Original answers on freetext question on challenges in data reuse regarding data quality (n=66, unsorted):

- mögliches preprocessing der Daten verstehen, dazu gibt es meist fehlende metadaten. Daten (excel-tabellen) stimmen meist nicht mit Berichten überein, z.B. unterschiedliche Sample-Anzahl..
- i only use good quality dataset that is published and opensource. so no challenge, coz they all good sorted. if it is bad quality, i will not use it. about how to measure if it is good or not, it should be good documented and structured, and data quality looks good (manural overview them).
- Datenquelle muss zuverlässig sein.
- Lack of standards and metadata
- Datenqualität und Vergleichbarkeit mit eigenen Daten
- no metada available, small sample size
- Nachvollziehbarkeit von Entscheidungsprozessen
- Fehlende Informationen zur Datenqualität, u.a. auch zur Vergleichbarkeit von Datensätzen
- methodology used for obtaining the data, reproducubility
- Zuviele Systeme mit jeweils eigenen Anforderungen, kein “easy to use” Outputsystem, Daten bedürfen permanent anderer Spezialkenntnisse
- kaum Validierung durch Fachexperten, keine klaren Qualitätsstandards
- Keine einheitlichen Erhebungskriterien, Unvollständigkeit
- Ich habe keine Ahnung davon.
- Vertrauen in die Institution die Daten bereitstellt
- Der Zugang zu Daten, die von den einzelnen Bundesländern erhoben werden. Unvollständigkeit der Daten. Zu kurze Zeitreihen der Daten
- Die Datenqualität ist auf den ersten Blick oftmals schwer einzuschätzen. Ein hoher Zeitaufwand wird benötigt um einzuschätzen/überprüfen, ob die Datenqualität zufriedenstellend ist.

- Die Daten, die ich für meine Arbeit benötige sind in der entsprechenden Auflösung nicht verfügbar oder die Kosten sind zu hoch.
- Data correctness
- Hintergrundinformationen/Metadata
- Fehlende/ungenaue Datenbeschriftungen, Flüchtigkeitsfehler bei der Datenerfassung
- Dokumentation und Zugänglichkeit von Daten
- Out of date data, lack of locality detail
- Nicht überprüfte Datenfehlern
- missing data on key system components
- Einführung und Nutzung von Metadatenstandards
- Nachvollziehbarkeit der Datenerhebung,
- Unzureichende Beschreibung der Daten (QTL- Berechnung) und nicht ausreichende genaue Dokumentation und Erhebung der Ertragsdaten (Weinbaupraxis)
- Einschätzung der Zuverlässigkeit und Vollständigkeit der Daten bei Erhebungen von landwirtschaftlichen Bewirtschaftungs- und Bilanzdaten von Praxisbetrieben
- Intransparent data collection methods, poor collection protocols including incomplete data.
- Zuverlässigkeit der Daten und der darauf basierenden Schlussfolgerungen.
- Die Zeit, sich mit fremden Daten und deren Strukturen zu befassen, die nicht immer zu einem Ergebnis führt. In Drittmittelprojekten ist es in der Regel nicht vorgesehen, Forschung nur auf Daten anderer aufzubauen.
- Beschreibung der Methodik und Beschriftung
- Bezogen z. B. auf Drohnendaten wären das: fehlende Dokumentation der Erhebungsmethodik fehlende Angaben zu Auflösung, Flughöhe, Aufnahmedatum, Kultur auf dem Schlag fehlende Angaben zu technischen Parametern der Drohne (Art des Sensors, Drohnentyp etc.)
- Datenqualität stark abhängig vom Bonituer.
- Tlw. unterschiedliche Einheiten, die sich nur schwer umrechnen lassen. Fehlende Standards für Set an Parameter/Ergebnisse, die in allen Fällen publiziert werden sollten (auch wenn der Fokus nicht auf diesen definierten “Standarddaten” lag).
- Authenzität der Daten
- statistics
- Fehlende Beschriftungen von Spalten bwz. Beschriftungen, die nicht selbsterklärend sind, aber es auch kein data dictionary dazu gibt.

- Vergleichbarkeit der Daten und Zusammenfassung der Daten nach einem gegebenen Standard
- Resolution - temporal and spatial
- Fehlende Informationen über die abgebildeten Werte. Keine Metadaten.
- Vergleichbarkeit; meist sind auf den ersten Blick identische Auskünfte mit unterschiedlichen Zahlen belegt. Unklar, worin die Unterschiede liegen und welche Zahl für meine Nutzung verwendet werden könnte.
- Vollständige Datensätze zu erhalten.
- Die Kreativität für fehlerhafte Eingaben ist unendlich
- Daten nicht frei verfügbar (Open Access), zB InVeKoS
- Nutzung von Excel als Erfassungsformular
- unvollständige Beschreibung der Erhebungsmethodik, unvollständige Metadaten, fehlende Harmonisierung, unvollständige Validierung
- Interoperabilität zwischen datengenerierenden, proprietären Systemen
- Maschinenlesbarkeit, Interoperabilität
- standartisierte Antworten
- Lack of standardized process of data management and manipulation
- Unzureichende Beschriftung von Spalten, Einheiten etc. und variierende Formatierung von Exceltabellen in verschiedenen Jahren. Das macht die eigentliche Auswertung sehr aufwendig, weil erst alles formatiert, übertragen etc. werden muss.
- fehlende Metadaten, fehlende Standards für Dateninhalte, typen, strukturierung, Datenschutz, Maschinenlesbarkeit bzw. fehlende Datenschnittstellen
- Je schlechter die Datenqualität, desto länger dauert die Aufbereitung, und desto mehr Annahmen müssen getroffen werden.
- Datenschutz
- Zunehmender Erfassungszeitaufwand bei standardisierten Formularen - man kann gar nicht alles “für Fremde” dokumentieren/aufschreiben, was die Expertin vor Ort mit in der Interpretation berücksichtigt aus Erfahrung; man hat keine Garantie, dass Daten wirklich so erhoben wurden wie besprochen - große Fehlerquelle für falsche Ableitungen; fehlende IT-Ausstattung und -Freiheiten beim öffentlichen Dienst für Cloud/Software/Rechnerleistung etc.
- Vereinheitlichung, Vollständigkeit und Korrektheit der Daten
- Für mein gewähltes Beispiel, haben die unetrschiedlichen Länder uterschiedliche Standards, wie sie die

Datensätze anbieten. Außerdem gibt es kaum Dokumentationen zur Datenqualität der Daten und auch kaum Metadatenstandards.

- Digitalisierung schriftlich oder mündlich gewonnener daten
- For the datatype I use, one of the biggest challenge is the inconsistent nomenclature for updated information on genotyping data.
- Bezogen auf oben genannte Daten: Es war schwer eindeutige Identifier für die jeweiligen Gene/Proteine zu finden, da es sich um Shotgun Sequencing Daten und darauf basierende hypothetische Proteinen handelte.
- The processing pipeline should be customized for each dataset due to a lack of standardization
- Vereinheitlichung der Datensätze (zB über mehrere Umwelten), Datenvollständigkeit
- Datenaufbereitung
- Ich verwende auch Sekundärdaten, wenn Primärdaten fehlen. Es gibt genügend allgemeine Daten. Leider gibt es nicht immer die spezifischen Sekundärdaten, die ich genau brauche (bestimmter Ort/Arbeitsschritt/Maschinen/Länder).
- Das Unwissen über die Qualität der Daten und fehlende bzw. unzureichende Dokumentation.

Appendix D: Data reuse in practice

D3: How did you select the 'appropriate' dataset?

Original answers in *Others* (n=1): Plausibilität, Stabilität der Ergebnisse

Appendix E: More on the topic of data quality

E2: Are there any additional information or thoughts on the topic of data quality that you would like to share with us?

Answers of free text question about additional thoughts on data quality (n=37):

- ich finde s schwierig Datenqualität zu beurteilen, denn für Publikationen werden denke ich meist schon bereinigte Daten verwendet, also vielleicht war die raw data nicht optimal konnte aber zu einem gute Dataset afgereinigt werden. Bei dem dataset würde ich denn eine gute datenqualität geben aber vielleicht war es ursprünglich nicht so.
- different data type 's quality need to be measured by different methods.
- Statistische Maßzahlen wären wichtig. Herkunft der Daten entscheidend.
- Wichtigster Punkt ist aus meiner Sicht ist, dass es nach wie vor nahezu unmöglich ist, aus dem knappen M&M-Teil einer Publikation Forschungsdaten zu reproduzieren. Ein derartiges Gütesiegel (z.B. Bestätigung von experimentellen Daten durch wirklich unabhängige Replikate, idealerweise in anderen Laboren) würde es sehr erleichtern, Daten nachzunutzen. Eine gute Beschreibung der Daten und des Erhebungszwecks hilft dabei, die Datenqualität für die eigene Anwendung einzuschätzen. Diese Daten fehlen oft. Leider fallen auch meist ein Großteil von Datensätzen aus der Betrachtung heraus, weil sie für den eigenen Anwendungsfall lückenhaft sind. Dadurch kann es sehr leicht zu einer Verzerrung der Datenlage kommen.
- I overlooked the following aspects: i) Ensuring the reproducibility of results, emphasizing the need for not just replicating final outputs (e.g., graphs for an article) but also documenting the detailed steps involved in data preprocessing. This underscores the significance of utilizing programming languages like R or Python to script all steps comprehensively, from raw data to processed datasets. ii) Highlighting the crucial role of proper data version control, including details such as date, authorship, and data origin, to maintain a systematic record of dataset cycle.
- Mein Beispiel kam aus der Verhal-

tensökologie (speziell visuelle Ökologie) von Schadinsekteninsekten. Dieser Fragebogen zielte wahrscheinlich eher auf den Agrarkontext ab. Für uns ist besonders wichtig:

  - Das Versuchsdesign (muss gut dokumentiert sein)
  - Kontrollierte Bedingungen bzw. (Dokumentation)
  - Alles was nicht kontrollierbar ist möglichst gut (mit Sensoren) messen (Dokumentation)
  - Gute Auswahl der Ko-Variablen, ggf. mit stat. Methoden (Statistik, biologische Relevanz)
  - Modellbildung und Modellevaluation anhand von gut ausgewählten, ggf. normierten und biologisch informierten Datensätzen (Physiologie etc.)

- 1.Die meisten Daten in den "Agrarwissenschaften" sind nicht detailliert genug, z.B. QTL werden als Ort bschrieben, aber es erfolgt meist nicht mal eine Übersicht über zumindest vermutlich interessante Gene dort. 2. Daten bestehen nur aus Gene-IDs, aber so gut wie nie aus oder mit funktionalen Annotationen, die kann man alle beschaffen. . . , aber die GO-Terms oder Protein Funktionsvorhersage würde solche Transkriptdaten (als Bsp) sortierbar machen - ah, alles Calcium-Bindeproteine, damit habe ich schon oft im Modelsystem schnell gute neue Hypothesen generieren können
- Auch wenn auf dem Papier Forschungseinrichtungen tolle Datenmanagementkonzepte habe, scheitert es bei der praktischen Umsetzung (keine klaren Vorgaben für Erhebung, Validierung und Speicherung von Forschungsdaten, mangelnde Infrastruktur, keine ausreichenden Rechnerkapazitäten).
- Ich bin entsetzt, wie wenig Wert in unserer Einrichtung auf tatsächliche Datenqualität und deren Validität gelegt wird, obwohl überall von "Guter Wissenschaftlicher Praxis" etc. geredet wird. Diese Qualitätsstandards werden aber in der praktischen Umsetzung nicht eingehalten und wenn man darauf hinweist wird gesagt, dass macht doch jeder so.
- Für mein Projekt arbeite ich mich hochauflösenden Geodaten, die eine Auflösung von mindestens 50m, am besten noch geringer haben müssen. Die Datenverfügbarkeit einer breiten Masse an Geodaten (Bodendaten, Topografische Daten, Erosionsdaten etc.) ist leider recht schlecht. Frei verfügbar sind meist nur Daten mit einer geringen Auflösung. Hochauflösende Daten sind sehr teuer oder gar nicht erst verfügbar.
- Anwendung: Form der Daten in zB Datenbanken, die sich verändern und zT nicht mehr anwendbar sind.
- no
- A first helpful impression on data qual-

ity can be achieved via a simple visualization of data sets by looking for plausibility, completeness, outliers, trends, variable comparisons and logic dependencies of variables.

- Bei meiner Tätigkeit ist es sehr wichtig, zu möglichst genauen und sicheren Einschätzungen zu kommen, da die Konsequenzen meiner Einschätzung sehr weitreichend sein können. Je besser die Datenqualität und -verfügbarkeit, um so sicherer kann auch die Einschätzung sein.
- data quality depends on the persons who generate data!! This can be improved if useful templates for data entry would exist
- Die Idee von FAIR ist gut, führt aber dazu, dass einige Forschende viel Zeit und Arbeit in Datendokumentation investieren sollen, damit die Statistikorientierten Forschende und Modeller sie nutzen können - ohne jeglichen Benefit.
- Dieser Fragebogen ist wahr, aber vermutlich ungültig. Ich habe ihn auf ihr Drängen hin trotzdem ausgefüllt.
- Ich finde es am wichtigsten, Klarheit über den Inhalt und die Qualität der Daten zu haben. Wenn ein Datensatz nicht komplett ist, oder teilweise mit schlechter Qualität zu rechnen ist, ist das für mich kein Ausschlusskriterium, wenn das so klar gekennzeichnet ist. Denn dann kann ich den Datensatz in reduzierter Form eventuell trotzdem nutzen.
- I would like to recieve more information about this
- Es bräuchte einen Datencodex, der an den Deutschen Hochschulen oder Graduiertenschulen verbindlich unterichtet wird. Viele amtierenden Hoschschullehrer sind nicht kompetent, das zu unterrichten, benötigen aber Mitarbeiter, die die Standards kennen und anwenden können.
- like in Cochrane reviews, publication of the financial support of the data research by which donators.
- Compare similar or different datas to the topics.
- Einheitliche Art der Datenerfassung und -bewertung durch die erhebenden Personen ist notwendig. Hierfür sind entsprechende Standards notwendig und digitale Erfassungs- und Analysemöglichkeiten.
- Mir fehlen mitunter Grundkenntnisse, um Metadaten adäquat zur Verfügung stellen zu können, da ich in der Regel keine Daten generiere, sondern nur Daten auswerte oder ausgewertete Daten nutze. Vielleicht könnte Informationsmaterial hierzu zur Verfügung gestellt werden.
- Extrem wichtig und hilfreich für die Auswertung ist eine Schulung der Datenerheber und klare Vorgaben, am

besten in Form von Auswahllisten, für die Hauptkomponenten wie z.B. Schadorganismen, Pflanzenschutzmittel

- Die Abfrage ist generisch gehalten - anders geht es auch nicht. Allerdings wäre es hilfreich, die Antworten an konkreten Beispielen festzumachen, da verschiiedene Teilnehmende womöglich Verschiedenes mit gleichen Aussagen meinen.
- Mir ist nich so ganz klar, was in dieser Umfrage unter “Datenqualität” genau gemeint ist. Zum Thema Datennutzung möchte ich anmerken, dass es für viele Fragen sehr schön wäre, wenm Daten zB zu Tiergesundheit in landwirtschaftlichen Betrieben, für die Wissenschaft (ggfs auch für die Politik) zugängich wären. Der Datenschutz bremst uns hier aber regelmäßig aus. Uni-intern steckt das Forschungsdatenmanagement noch in den Kinderschuhen, und es fehlt noch an Erfahrungen zu Präregistrierung oder Datenrepositorien
- nein
- Mixed questions about quality - quality of the measurement process (data collection) vs quality of the equipment (sensor/equipment resolution). These are two different aspects of data quality.
- Die bitte, dass in Deutschlands zunehmend Standards und zentrale Datenbanken bzw. Datenschnittstellen geschaffen werden, um den Aufwand Daten zu akquirieren und nutzbar zu machen endlich zu reduzieren und so ein effizientes und zielführendes wissenschatliches Arbeiten zu ermöglichen.
- Uneinheitliche Klassen von TIerdaten sind sehr ärgerlich. Im (häufig vorkommenden) Extremfall kann man im Vergleich von zwei Datengrundlagen nur noch “Schweine” oder “RInder” miteinander vergleichen, weil feinere Untergruppen unvergleichar gewählt sind.
- Bitte kein Konstrukt schaffen, das viele hehre Ziele hat, aber das wissenschaftliche Datenerheben und -auswerten furchtbar umständlich macht!
- Eine Plausibilitätsprüfung der Daten sollte vor jeder Datenanalyse vorangestellt werden.
- Die Bereiche Nutztierwissenschaften und WiSoLa der Agrarwissenschaften ist in dieser Umfrage nicht ausreichend in den Antwortmöglichkeiten (Pflanzenschwerpunkt) berücksichtigt worden.
- Alle Arbeiten die zukünftig extra anfallen, um Arbeiten/Ergebnisse zu dokumentieren und zu bewerten, müssen möglichst kaum zusätzlicher Zeitaufwand sein. Daher wären Tools, die (semi-)automatisiert,

Forscher*innen zur Verfügung gestellt werden, eine sinnvolle Sache. Bspw. die automatisierte Erstellung von Metadaten-Dateien oder Ausweisung von Qualitätsstandards wären hilfreich.

- Eher das Problem externer Daten. Mehrjährige Feldhistorien, die vom Landwirt gewonnen werden. Dabei weniger das Problem der Qualität sondern der Digitalisierung. Ganz wichtig, welche Qualität ist überhaupt nötig und technisch und wirtschaftlich nötig. Welche Daten sind überhaupt nötig. Eher das Problem die richtigen "dicken" Bretter bohren.
- Genotypdaten wurden nicht erwähnt.
- Wenn es eine Art Prüfsiegel gäbe zumindest bei Staatlichen Akteuren (Behörden, Uni etc.), Damit man auf einem Blick sieht ob die Daten geprüft wurden oder das nach einer bestimmten NORM bei der Ertellung vorgegangen wurde.

References

Ewert, F., Arend, D., Senthold, A., Boehm, F., Feike, Til, Fluck, Juliane, Gackstetter, David, Gonzalez-Mellado, Aida, Hartmann, Thomas, Haunert, Jan-Henrik, Hoedt, Florian, Hoffmann, Carsten, König, Patrick, Lange, Matthias, Lesch, Stephan, Lindstädt, Birte, Lischeid, Gunnar, Martini, Daniel, Möller, Markus, … Weiland, Claus. (2023). FAIRagro FAIR data infrastructure for agrosystems - proposal 2021. https://doi.org/10.5281/ZENODO.7528172

Senft, M., Stahl, U., & Svoboda, N. (2022). Research data management in agricultural sciences in Germany: We are not yet where we want to be. *PloS one*, *17*(9), e0274677. https://doi.org/10.1371/journal.pone.0274677

Specka, X., Martini, D., Weiland, C., Arend, D., Asseng, S., Boehm, F., Feike, T., Fluck, J., Gackstetter, D., Gonzales-Mellado, A., Hartmann, T., Haunert, J.-H., Hoedt, F., Hoffmann, C., König, P., Lange, M., Lesch, S., Lindstädt, B., Lischeid, G., … Ewert, F. (2023). FAIRagro: Ein Konsortium in der Nationalen Forschungsdateninfrastruktur (NFDI) für Forschungsdaten in der Agrosystemforschung: Herausforderungen und Lösungsansätze für den Aufbau einer FAIRen Forschungsdateninfrastruktur. *Informatik Spektrum*, *46*(1), 24–35. https://doi.org/10.1007/s00287-022-01520-w