Data quality in the application of agricultural system data

Warm welcome to the survey on ´**Data Quality in the Application of Agrosystems Data**` provided by FAIRagro.

Thank you for your willingness to participate in this survey! This survey is aimed at anyone who works with data relating to the agricultural sector.

This survey focuses on the application of data quality criteria in handling research data in the agricultural field. After a few introductory questions about you, there will be a section on data quality and another on data reuse. At the end of the survey, you will have the opportunity to leave comments and/or feedback.

The aim of this survey is to learn from your own experiences and assessments regarding data quality. Please answer the questions based on your own gut feeling – there is no right or wrong. We look forward to your responses.

The survey is voluntary and anonymous. No personally identifiable information will be processed. The survey aims to enhance and optimize FAIRagro services. After completing the survey, you have the option to leave your email through a link to stay informed about our future work, surveys, and workshops related to data quality. For more information, please visit the FAIRagro website: https://www.fairagro.net/

Enjoy the survey!

On behalf of the FAIRagro consortium,
Anne and Jannes

For content-related questions, please contact: Jannes Uhlott (jannes.uhlott@julius-kuehn.de).
For technical questions or issues, please contact: Anne Sennhenn (asennhenn@atb-potsdam.de)

The duration of the survey is 10-12 minutes

There are 27 questions in this survey.

Welcome

Here we go! To start, we would like to gather some information about your scientific background and your experience with data.

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

Select all that apply

Please choose **all** that apply:

□ Data producers

□ Data users

□ Infrastructure service providers (e.g., data managers, repository operators, database developers, and IT staff)

□ Information service providers (e.g., employees in libraries, publishing, and research data management coordination)

□ Other:

Which of the following groups do you primarily identify with?

Choose one of the following answers

Please choose **only one** of the following:

○ Students

○ Scientific staff (doctoral candidates)

○ scientific staff (postdocs)

○ Non-scientific staff

○ Professors/AG leaders

○ Other

Which of the following institutions do you primarily identify with?

❗ Choose one of the following answers

Please choose **only one** of the following:

○ University/University of applied sciences
○ Non-university research institutions
○ Public authorities
○ Infrastructure facilities (e.g. libraries, archives)
○ Industry/services
○ Other ________________

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

❗ Select all that apply

Please choose **all** that apply:

□ Soil sciences
□ Plant breeding, plant pathology
□ Plant cultivation, plant nutrition, agricultural engineering
□ Ecology of land use
□ Agricultural economics, agricultural policy, agricultural sociology
□ Forest sciences
□ Animal breeding, animal nutrition, animal husbandry
□ Veterinary medicine
□ Zoology
□ Plant sciences
□ Geophysics and geodesy
□ Other: ________________

Data collection

In this section, there will be questions related to conducting your own data collection.

If you are not currently collecting data, you may be able to recall a previous data collection effort. If you have conducted multiple data collections or are currently collecting data from multiple sources, please focus on **one** specific example. Please answer the following questions with reference to your chosen example.

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

❗ Choose one of the following answers

Please choose **only one** of the following:

○ Yes

○ No

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

Only answer this question if the following conditions are met:
B1 == "A1" or B1 == "A2"

❗ Choose one of the following answers
Please choose **only one** of the following:

○ Parcel data

○ Omics data (e.g., genomics, transcriptomics, proteomics, metabolomics)

○ Subject information data (e.g., legal requirements, authorisation data for plant protection products and seeds, livestock movement data)

○ Market information data (e.g., yield estimates, harvest forecasts, planning guidelines, breeding value estimates)

○ Forecast/model/simulation data

○ Weather data

○ Occurrence/infestation/distribution data

○ Soil data

○ Rating data

○ Laboratory data (e.g., chemical analysis data)

○ Remote sensing data (e.g., drone data)

○ Technical equipment data

○ Phenotyping data

○ Performance test data (e.g., milk performance test for cattle or meat performance test for pigs)

○ Questionnaires

○ Statistics

○ Other ▭

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

Only answer this question if the following conditions are met:
Answer was 'Yes' at question ' [B1]' (Do you currently collect or have you ever collected data by yourself?)

❗ Choose one of the following answers
Please choose **only one** of the following:

- ○ Numerical data
- ○ Text data
- ○ Audio data
- ○ Photo data
- ○ Video data
- ○ Tabular data
- ○ Spatial data
- ○ Source Codes
- ○ Non-digital data (e.g. soil samples)
- ○ Other

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

Only answer this question if the following conditions are met:
B1 == "A1" or B1 == "A2"

Choose one of the following answers
Please choose **only one** of the following:

○ Soil sciences
○ Plant breeding, plant pathology
○ Plant cultivation, plant nutrition, agricultural engineering
○ Ecology of land use
○ Agricultural economics, agricultural policy, agricultural sociology
○ Forest sciences
○ Animal breeding, animal nutrition, animal husbandry
○ Veterinary medicine
○ Zoology
○ Plant sciences
○ Geophysics and geodesy
○ Other

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

Only answer this question if the following conditions are met:
B1 == "A1" or B1 == "A2"

❗ All your answers must be different and you must rank in order.
❗ Please select at most 3 answers
Please number each box in order of preference from 1 to 10
Please choose no more than 3 items.

□ Resolution (e.g., spatial, content, temporal)

□ Completeness (e.g., in spatial, content, temporal terms)

□ Information on statistical certainty (e.g., indication of uncertainties, verified consistency of content)

□ Machine readability

□ Use of standards/data interoperability

□ Information on data pre-processing (e.g., application of filters, outlier correction)

□ Detailed metadata

□ Information on methodology (e.g., field protocols)

□ Detailed description of the data (e.g., labelling of columns, units)

□ Up-to-dateness of the data set

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

Only answer this question if the following conditions are met:
Answer was 'Yes' at question ' [B1]' (Do you currently collect or have you ever collected data by yourself?)

Please write your answer here:

How do you ensure the quality of your collected data?

Only answer this question if the following conditions are met:
B1 == "A1" or B1 == "A2"

❗ Select all that apply
Please choose **all** that apply:

- □ Calibration of instruments (accuracy/scale)
- □ Multiple measurements
- □ Review by experts
- □ Use of standardised methods and protocols
- □ Minimisation of manual data entry (e.g. controlled vocabularies, code lists, selection lists)
- □ Manual checks (e.g. completeness, duplicate entries)
- □ Statistical analyses (e.g. frequencies, mean values, dispersion, outlier values)
- □ Comparisons with reference data
- □ Discussion of the data with peers
- □ Consideration of boundary conditions (e.g. cloud cover, temperature, soil moisture...)
- □ Other:

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

Only answer this question if the following conditions are met:
B1 == "A1" or B1 == "A2"

❗ Select all that apply
Please choose **all** that apply:

□ No clear metadata standards
□ No clear data quality standards
□ Lack of information on data quality (What is important? What do others need?)
□ Lack of contact persons (for questions/challenges)
□ Lack of time
□ Lack of standards for data collection (e.g. standardised protocols)
□ Insufficient sensor accuracy

□ Influences of boundary conditions (e.g. weather conditions)
□ Lack of location for documentation of metadata/data quality information
□ Technical challenges (no metadata fields available)
□ I do not consider it necessary to document metadata/data quality information.
□ I am not aware of any quality requirements for my data (e.g. the first data collection in a specific area)
□ Other:

Use of data

In this section, there will be questions related to the (re-)use of already existing data. If you are currently not using data that you did not collect yourself, you may recall **one** previous occasion of utilizing existing data to answer the questions in this section. If you use / have used several data sets, please choose **one example** in the following questions. Please answer the following questions in relation to your chosen example.

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

❗ Choose one of the following answers

Please choose **only one** of the following:

○ Yes

○ No

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

Only answer this question if the following conditions are met:
C1 == "A1" or C1 == "A2"

❗ Choose one of the following answers
Please choose **only one** of the following:

- ○ Parcel data
- ○ Omics data (e.g., genomics, transcriptomics, proteomics, metabolomics)
- ○ Subject information data (e.g., legal requirements, authorisation data for plant protection products and seeds, livestock movement data)
- ○ Market information data (e.g., yield estimates, harvest forecasts, planning guidelines, breeding value estimates)
- ○ Forecast/model/simulation data
- ○ Weather data
- ○ Occurrence/infestation/distribution data
- ○ Soil data
- ○ Rating data
- ○ Laboratory data (e.g., chemical analysis data)
- ○ Remote sensing data (e.g., drone data)
- ○ Technical equipment data
- ○ Phenotyping data
- ○ Performance test data (e.g., milk performance test for cattle or meat performance test for pigs)
- ○ Questionnaires
- ○ Statistics
- ○ Other ________________

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

Only answer this question if the following conditions are met:
Answer was 'Yes' at question ' [C1]' (Do you use or have you ever used data that you did not collect yourself?)

❗ Choose one of the following answers
Please choose **only one** of the following:

○ Numerical data
○ Text data
○ Audio data
○ Photo data
○ Video data
○ Tabular data
○ Spatial data
○ Source Codes
○ Non-digital data (e.g. soil samples)
○ Other ________________

What is the purpose of your data reuse?

Only answer this question if the following conditions are met:
C1 == "A1" or C1 == "A2"

❗ Select all that apply
Please choose **all** that apply:

□ Generation of data products (e.g., specialised maps)
□ Commercial use/product development (e.g., fertiliser)
□ Analyses (e.g., quality analyses)
□ Input data for models and algorithms (e.g., yield estimation)
□ Planning/decision-making (e.g., location determination)
□ Other:

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

Only answer this question if the following conditions are met:
C1 == "A1" or C1 == "A2"

❗ Comment only when you choose an answer.
Please choose all that apply and provide a comment:

□ Publicly accessible data (e.g., repositories, search engines, databases)

□ Internal data (e.g., colleagues within the working group, institution or projects)

Other:

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

Only answer this question if the following conditions are met:
C1 == "A1" or C1 == "A2"

All your answers must be different and you must rank in order.

Please select at most 3 answers

Please number each box in order of preference from 1 to 12

Please choose no more than 3 items.

□ Resolution (e.g., spatial, content, temporal)

□ Completeness (e.g., in spatial, content, temporal terms)

□ Information on statistical certainty (e.g., indication of uncertainties, verified consistency of content)

□ Machine readability

□ Use of standards/data interoperability

□ Information on data pre-processing (e.g., application of filters, outlier correction)

□ Detailed metadata

□ Information on methodology (e.g., field protocols)

□ Detailed description of the data (e.g., labelling of columns, units)

□ Up-to-dateness of the data set

□ Secure data source (e.g., official repository, published data)

□ Secure rights of use/open access

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

Only answer this question if the following conditions are met:
C1 == "A1" or C1 == "A2"

Select all that apply

Please choose **all** that apply:

□ Incomplete data

□ Incorrect data

□ Data preparation not documented

□ Missing material and method description

□ No open access/open data

□ Costs

□ Lack of infrastructure (e.g., big data, lack of computing capacity)

□ Other:

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

Only answer this question if the following conditions are met:
Answer was 'Yes' at question ' [C1]' (Do you use or have you ever used data that you did not collect yourself?)

❗ Select all that apply
Please choose **all** that apply:

□ Interpolation (e.g., spatial, temporal, content-related)
□ Extrapolation (e.g., spatial, temporal, content-related)
□ Data reduction
□ Data cleansing
□ Data addition
□ Normalisation/standardisation
□ Utilising expert knowledge
□ Other: ▭

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

Only answer this question if the following conditions are met:
C1 == "A1" or C1 == "A2"

Please write your answer here:

▭

Data reuse in practice

When using external data, it can happen that multiple datasets are available for the intended use case. In this section, there will be questions designed to capture the challenges in such situations.

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:

Choose one of the following answers

Please choose **only one** of the following:

○ Helpful

○ Neutral

○ Unimportant

○ Other □

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

Choose one of the following answers

Please choose **only one** of the following:

○ Yes

○ No

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

Only answer this question if the following conditions are met:
Answer was 'Yes' at question ' [D4]' (Would you be willing to invest time in documenting completed applications of a dataset in its metadata?)

❗ Only numbers may be entered in this field.
Please write your answer here:

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?

❗ Choose one of the following answers
Please choose **only one** of the following:

○ Yes
○ No

How did you select the 'appropriate' dataset?

Only answer this question if the following conditions are met:
D2 == "A1"

❗ Select all that apply
Please choose **all** that apply:

□ Highest resolution (e.g., spatial, content, temporal)
□ Most detailed metadata
□ Most detailed description of the data preparation
□ Best references (e.g., recommendations by others)
□ Most dense time series
□ Fewest outliers
□ Highest quality (e.g., lowest error rate)
□ Highest repetition rate (e.g., every year instead of every two years)

□ Largest coverage (e.g., all of Germany instead of a single federal state)
□ Best results in application
□ Most appropriate semantics (e.g., classes)
□ Lowest costs
□ Open access
□ Smallest file size
□ Other: ________

More on the topic of data quality

This final section provides you with the opportunity to express your personal thoughts on the topic of data quality.

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

Please write your answer here:

Thank you very much for your participation in the survey on data quality and for your valuable contribution to a FAIR data future in agricultural systems research!

Are you specifically interested in the topic of data quality and further collaboration opportunities within the framework of FAIRagro? If so, you can leave your contact information in a separate contact form.

The results of the survey will be presented at the FAIRagro workshop on 13 + 14 March 2024 in Braunschweig. More information will follow in the FAIRagro newsletter. Feel free to subscribe to the FAIRagro newsletter mailing list to receive further information about FAIRagro services and future events.

You can find more information about the NFDI initiative FAIRagro and FAIR data management on the FAIRagro websites.

For current updates on FAIRagro activities, follow FAIRagro on X (formerly known as Twitter; http://twitter.com/FAIR_agro) and/or Mastodon (http://nfdi.social/@FAIRagro)!

01-20-2024 – 23:59
Submit your survey.
Thank you for completing this survey.