How to improve the legacy value of your dataset

Carmen S.P. TEIXEIRA[1*], David F. CHAPMAN[2] and Derrick J. MOOT[1]

[1]*Field Research Centre, Faculty of Agriculture & Life Sciences, Lincoln University*
PO Box 85084, Lincoln 7647, New Zealand
[2]*DairyNZ Ltd, PO Box 85066, Lincoln 7647, New Zealand*
*Corresponding author: carmen.teixeira@lincoln.ac.nz

Highlights

- In New Zealand, agricultural datasets are important tools for tactical and strategic decision making on farm, and to inform policy at a regional and national scale.
- A structured system for collation and storage of raw data which offers data entry in the AgYields National Database is encouraged to enhance the reliability of agricultural data and its quality.
- Consistent use of methodology for data collection and reporting is the biggest challenge ensuring data quality and utility in AgYields.
- High-quality data enhances the validity of findings, which enables meaningful and trustworthy conclusions to be drawn from it.

Keywords: Accuracy, completeness, national database, quality, yield data

Background

Agriculture datasets are an essential source of information at a global level to address the challenge to feed a growing population, from production systems that are sustainable and adapted to changing environmental and climate conditions (Runck et al. 2022). They provide evidence that allows farmers, researchers, industry organisations, commercial businesses and statutory authorities and other stakeholders to make decisions regarding, for example, farm system management, priority areas for investment in scientific research, and policies for managing environmental emissions from agricultural land uses decisions and drive progress in the agricultural sector (Weersink et al. 2018). The effectiveness of datasets for these purposes will be maximised when users have full confidence in the relevance, coverage, quality and accessibility of the data that they draw on (Carolan et al. 2015).

Datasets are the evidence base for decision making because data collection and processing aim to reduce uncertainty to allow users to make more knowledgeable decisions (Daft and Lengel 1986). To enhance the reusability of academic and commercial generated data there is a need for a concise and measurable set of principles with specific emphasis to improve the utility of data for individuals (Senft et al. 2022) and of computers to automatically find and use data. The FAIR (Findable, Accessible, Interoperable, and Reusable) principles were created and supported by a diverse set of stakeholders that represent universities, industry, funding agencies, and scientific publishers in Europe (Wilkinson et al. 2016). The FAIR principles are complemented by the CARE principles (Collective benefit, Authority to control, Responsibility, Ethics) encouraging open data movements to consider both people and purpose behind the data (Carroll et al. 2020). In New Zealand, the AgYields National Database (agyields.co.nz) is a web-based repository which operates under these principles. It serves as a comprehensive source for data on pasture and crop growth rates, yields and flowering date. It aims to pool historic data, and that from current and future studies, for all agricultural regions of New Zealand. The database also provides subclasses of important site and production parameters such as soil type, harvest method, defoliation types and irrigation. All data are organized into the New Zealand regions and then location (i.e. districts or city) and narrow down to sublocation or site (farm or even paddock level) through latitude and longitude co-ordinates.

The aim is to offer a resource that can be used to inform livestock and crop production systems across New Zealand, and guide future data collection practices by providing standardise methods to optimize the utility for stored information in recognized formats. Besides data collection, ensuring data quality and completeness is crucial in research to maintain the integrity and reliability of experimental findings (Malaverri and Medeiros 2012). This includes developing a data collection plan that uses validated methods and equipment to standardize data entry and minimize errors while gathering as much information as possible.

Data quality is a common concern in a wide range of areas and refers to the accuracy, reliability, consistency, and overall fitness for purpose. It encompasses the extent to which data are free from errors, inconsistencies, and deficiencies and can be trusted to inform decision making, conduct analysis, and support various business processes (Ferris and Rahman 2017). Reliable data are needed to determine relationships between yields and weather data, for example, to inform the potential

ISSN 2463-2872 (Print) ISSN 2463-2880 (Online) https://doi.org/10.33584/jnzg.2024.86.3681

Table 1 Summary of the format in which data are entered into the 47 different fields associated with each yield or growth rate record in the AgYields database.

Data Format	Number of fields	%	Details
Date	5	10	DD/MM/YYYY
Label	13	28	Text option from list or drop-down sub menu
Numeric	16	34	Numerical value entered by the user
Text	13	28	Free text entered by the user

and realistic yields in different locations nationwide. Incomplete data, resulting from missing values and accompanied meta-data, impairs data quality and utility (Gandar and Kerr 1980). Missing data can occur due to various factors, such as careless data entry, or physical loss of records. Data completeness is one key measure of how well a dataset captures relevant information and its suitability for querying, analysis, and mining (Liu et al. 2016).

Efforts to enhance the reliability of agricultural data include, for example, methodologies for data collection and analysis, development of novel database systems and software applications (Malaverri and Medeiros 2012) such as the AgYields National Database (Moot et al. 2021). Since prevention is more effective than correction, data collection and compilation are the first quality issues that need to be considered in the generation of data that are fit for use (Chapman 2005). For instance, non-reporting data, incomplete coverage of data, imprecise concepts and standard definitions are common problems faced during the collection and compilation of data on land use (FAO 2021).

New Zealand agronomic research is not immune from these problems. In an analysis of 100 papers reporting agronomic trials in the Proceedings of the Agronomy Society of New Zealand, and the New Zealand Journals of Experimental Agriculture and of Agricultural Research, Gandar and Kerr (1980) found that 37% of papers contained no information on climatic conditions experienced during trials, while a further 24% included just a comment on climate in the text. Thus, only ~ 40% of publications provided any quantitative information on climate. No analyses of data completeness in New Zealand agronomic research have been undertaken in the 44 years since this analysis was conducted.

The aim of the study reported here was to assess the completeness and accuracy of pasture and crop data available from unpublished and published datasets held in the AgYields National Database. The primary focus was on the reporting methods and standards required to allow relevant and accurate data to be included into the database and add value to legacy datasets. The high-level objective was to provide individuals and organisations with guidance on the additional parameters that should be included when publishing a field-based dataset to maximise its utility now and into the future.

Material and Methods

Two hundred and ninety-six (296) national submitted datasets were obtained from the AgYields National Database and compiled into a two-dimensional data frame (Supplement Material 1). The data frame contained a total of 47 cells, which were populated with entries according to the formats shown in Table 1.

The total number of rows extracted was 29195: 47% from published datasets and 53% from unpublished sources (i.e., farmers notes, commercial company trials, theses). Data frames were then classified into two groups: Pasture (defined as forage and conserved feed species) and Crop (harvested for grain or seed) after the compulsory entry field: “Is it pasture/crop?”, on the Site attributes entry mode page. The journals and conference proceedings from which published datasets were extracted are shown in Figure 1. Completeness was calculated as the percentage of the required data entry cells for which a numeric value or text information was entered (Dong and Peng 2013). For example, if information was entered for all 33 of the cells, then completeness was 100%; if only 28 cells were complete, then completeness was 85%.

Calculation and Analysis

The ISO/IEC 25012 standard defines completeness as “the degree to which subject data associated with an entity has values for all projected attributes and related entity instances in a specific context of use”. The completeness analysis considered the dynamic features using periodic data profiling to identify completeness (Guerra-García et al. 2023). The quantification of data categorical completeness for Pasture and Crop datasets used the aggregate Sum function (colSums()) to compute missing data in rows and columns which were blank (Table 2).

The qualitative evaluation of the reported data was performed by considering the non-compulsory entry field: “Description Notes” (Moot et al. 2021, Supplement 1) and the frequency of appearance of

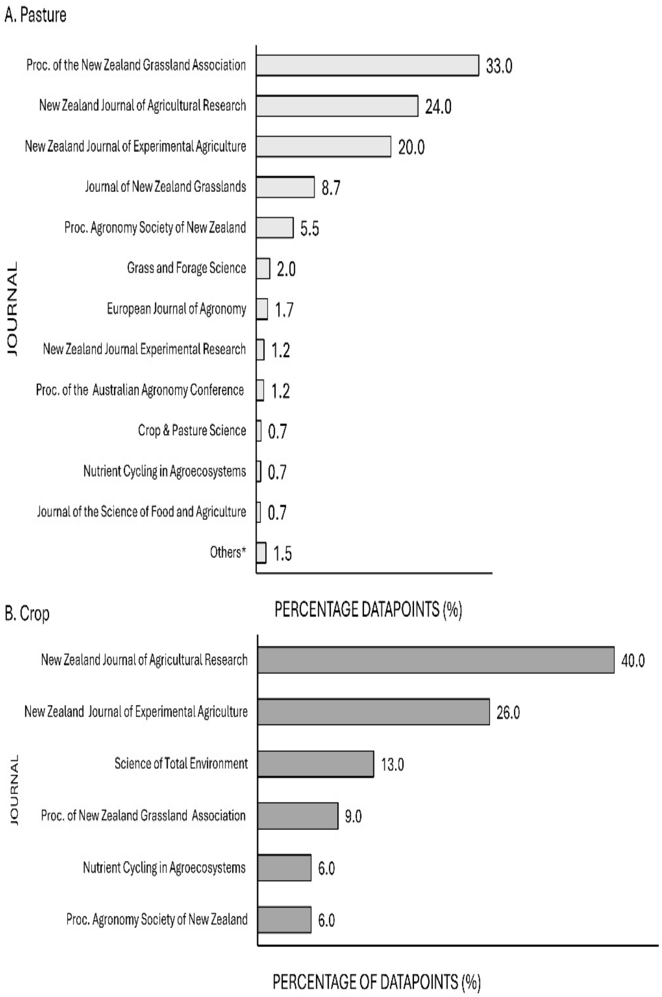

Figure 1 Journals from which published datasets were extracted for Pasture (A) and Crop (B) in the AgYields National Database. *Others represented 1.5% and refer to Journals: Agriculture, Ecosystems & Environment, Ministry of Agriculture and Fisheries, Agroforestry Systems, Proc. International Nitrogen Initiative Conference, Agronomy N.Z., Animal Production Science and Proceedings of the International Grassland Congress.

Table 2 Methodological steps used to compare crop and pasture data frames and quantify data categorical completeness.

Step	Procedure	Output
1	Crop and Pastures data frames were compared with a complete data frame (no missing categorical values = 0% of missing data	The overall completeness is reported as percentage (%) of general completeness (rows and columns).
2	Within both data frames (Crop and Pasture) the categories with the higher gaps were identified (blank cells in rows within each column)	Results were reported as percentage of completeness (%).

key words “assume(d), assumption, estimate(d), consider(ed)”. This was examined within attributes or categories which had the highest (or lowest) degree of accuracy, which is expressed as percentage (accuracy indicator).

Results and Discussion

Datasets overview

Of the total data available, 96% (27394 data rows) were from pasture trials with 4% (1801 data rows) from crop trials (Table 3). This represented a total of 247 different dataset titles for pastures and 51 dataset titles for crops. The dominance of pasture datasets is because the database was originally established to assemble pastoral data. Furthermore, since 2018 research effort has been directed at pasture species through funding from the livestock sector (e.g. National Forage Database /Dairy NZ, the Hill Country Futures Programme and the current Advancing AgYields to support forage/crop decision making (by Beef and Lamb NZ).

While this is the first comprehensive database for forage and crop yield, most data compiled to date came from three main regions (Table 3). The majority of datapoints (29%) were from measurements taken in Bay of Plenty followed by pasture measurements taken in Northland (~21% of the datapoints) and forage and crop records from Canterbury (20%). The percentage of records from these three regions reflect the previous research focus on collating data from those regions (Teixeira et al. 2023, Teixeira et al. 2023a) and the presence of experimental areas on those regions combined with the volume of experiments conducted over time (e.g. Lincoln University).

Overall data completeness

Across both the pasture and crop categories, overall completeness of the data accompanying each DM yield or growth rate value was 60%. Thus, a considerable amount of information was missing in each case, especially considering several data fields are mandatory (Supplement 1). This process revealed that some of the articles and unpublished datasets lack critical information and therefore, simply cannot be inserted into a database or can only be inserted using proxies. Most importantly, these results show that articles and experimental descriptions can be greatly improved. For instance, within the trial site attributes a mean of 81.5% of the records lack the altitude parameter.

Data completeness - Pastures

Data for the exact latitude and longitude of sites were missing for 27 and 31% of the pasture entries respectively. Information respectively on soil type and altitude was missing for 66% and 77% (Figure 2). This contrasts with Gandar and Kerr (1980) who found only 13% incompleteness with 87% of papers identifying the trial soil type. This discrepancy could be due to the unpublished datasets which do not have an indication of soil type.

Table 3 Total number of individual data points for DM yield or growth rate evaluated for completeness by region and category.

Region	Category	
	Pasture	Crop
Northland	6272	0
Waikato	2040	85
Bay of Plenty	8436	56
Hawkes Bay	1213	276
Whanganui-Manawatu	1400	206
Canterbury	4833	963
Otago	1212	102
Southland	1605	108
Other	383	5
Total (% total)	27394 (96)	1801 (4)

For pastures, entries can be either from resident or sown pastures. When sowing date was selected, 5% were assumed because authors did not provide the exact date. For instance, when authors mention “sown in September 1977” the assumed date is the 15th of September of 1977. When year is not mentioned, or the information provided is too general (for example sown in spring) a date cannot be entered, or ultimately the user must infer a hypothetical date. In this case it is recommended to insert an observation in the Description notes field.

Of more concern regarding data quality, the method used for measuring DM yield or growth rate was not specified for 19% of the entries, and no information was provided for the defoliation management of trial plots (cut, grazed etc.) for 48% of entries. Where the measurement method was given, exclusion cages and quadrat cuts were used to derive 38% and 21% of the data respectively.

Unlike measurement technique, there is no drop-down menu for the trial defoliation management in AgYields. Users can enter information as free text (Table 1) in the ‘Defoliation Method’ cell. For example, the user can indicate if the trial was cut mechanically or grazed: if it was grazed, the species of grazing animal (sheep, cattle), grazing method (continuous grazing, mob stocking rotational grazing etc.), and grazing intensity and frequency (rotation length, pre- and post-graze mass/height etc.) can be specified. The observation that

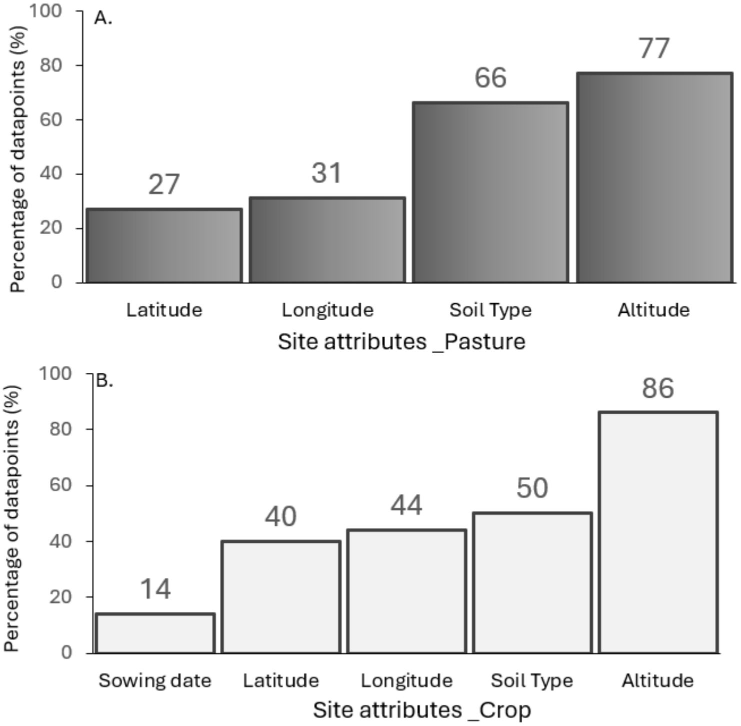

Figure 2 Percentage of datapoints missing site attributes information from pasture species (A) and crop (B) datasets. Pastures are defined as forage crops and conserved feed. Crops are defined as plants grown for grain or seed harvest.

this information is missing for half of all points to a major omission in the recording and reporting of trial details that should be addressed in future.

Data completeness - Crops

Crop datasets were more concise and less complex because they mostly come from monoculture trials established on one sowing date whereas pasture datasets are more complex in terms of species mixtures and different sowing dates (autumn, spring sown). Nonetheless, sowing date was stated in 14% of the datasets included in the analysis (Figure 2 B). The exact latitude and longitude of sites was poorly described with 40 and 44% of entries lacking the latitude and longitude respectively. Soil type information was not entered for 50% of all datapoints. For the sown crop entries only 14% of the records had no indication of sowing date.

Data accuracy

From all sowing dates which were informed 97.5±0.75% had exact date (Table 4). However, from all datasets combined, 29% of the sown crops and pastures records had indicated that the species were sown but no exact date has been provided nor "estimated" dates have been entered. For instance, the author(s) only reported "sown in November", "sown in early autumn". In this case, proxies can be used to estimate the sowing dates.

This process of filling in missing data with an estimate or predicted value is commonly referred as "data imputation" (Bennet 2001). The goal is to provide a more complete dataset for further analysis or modelling by replacing missing values with realistic estimates or proxies based on the available information whenever possible (Yi et al. 2021).

Regarding measurement dates, 10±0.4% of the harvest dates (End date) were assumed. Different from sowing dates, the measurement dates must be all entered (except when annual yield is inserted) so the degree of accuracy is slightly lower (mean of 90±2.8%) than the degree of accuracy of the sowing dates.

From all datasets in which irrigation was mentioned (for example "trial was irrigated") 63±14.5% of the records had no exact amount of water applied in irrigation so an estimated value had been inserted. Regarding species, <0.5% of pasture species were assumed while for the crop datasets, as expected, the species were accurately reported. In AgYields users have the possibility to select an option "Resident pasture of unknown composition" when only resident pasture is mentioned. In terms of species the most

Table 4 The degree of accuracy for selected variables within Pasture and Crop datasets from AgYields datasets.

Variable	Pasture Degree of accuracy *(%)	Crop Degree of accuracy *(%)
Sowing date	99	96
Measurement dates	98	82
Irrigation	77	48

common inaccuracy identified was when authors write "yield of ryegrass" or "clover yield" without specifying the species (e.g. perennial vs. annual ryegrass; white or red clover). The AgYields species list to date, has five options of ryegrasses (perennial, perennial hybrid, annual, Italian, hybrid) and 28 clover species. Over the decades the name of forage species might have changed or will change or still become a synonym. One example is kikuyu grass, recently named as *Chenchrus clandestinum* (Sierra et al. 2023) known previously as *Pennisetum clandestinum* (Percival 1980). The changes in nomenclature can cause some degree of uncertainty. It is recommended to users to consider as much as possible the information reported on the original dataset to ensure that the datum is accurate and matches the source.

Data collection in New Zealand

According to Hendy et al. (2018) New Zealand is "data poor" compared with other countries such as USA, Japan, France, Germany, India, China, and Australia. This is a limiting factor in the quality of agriculture and crop modelling results. The authors identified the (i) need to regularly update digital maps of land use that are available to all and (ii) to improve the acquisition of farm-level data which can be used to analyse, for instance, species suitability and economic performance of farms. Options for developing more useful data may include randomised controlled experiments to estimate key parameters (e.g. yields, growth rates), particularly on regions working with Landcorp to conduct experiments. Landcorp, also known by its Māori brand name Pāmu, is a New Zealand government state-owned enterprise. Its core business is pastoral farming, including dairy, sheep, beef, and deer farms. The data collected by a Landcorp (Pāmu) is useful to enhance farm performance, animal welfare, environmental responsibility, and overall industry progress. However, Gandar and Kerr (1980) reported that the effectiveness of agronomic research in New Zealand is relatively low due to poor targeting of the research to end user needs; poor design of trials relative to the (often unspecified) aims or targets (e.g. inadequate treatment levels, short duration, failure to replicate in different environments); and incompleteness of associated information provided in papers especially with respect to environmental variables.

Since 1991, New Zealand's seed companies and plant breeders implemented comprehensive nationwide trials known as the National Forage Variety Trials (NFVT) system (Thom et al. 1998). These trials aim to provide impartial data regarding cultivar performance to the pastoral industry (Easton et al. 1997). This valuable dataset was subsequently transferred to DairyNZ, where it was incorporated into the DairyNZ Forage Value Index (Chapman et al. 2019). This index plays a vital role in the evaluation of pasture cultivar performance aligned with industry-established standards (Lee et al. 2018, PBRA 2023). The trials consist of replicated studies conducted across the country to assess both current and new cultivars (Chapman et al. 2017). These trials follow a technical protocol, and each trial undergoes an independent audit on an annual basis.

Assessment of dry matter yield is conducted using a rotational management approach. There are summaries available for example for annual, Italian, hybrid and perennial ryegrasses, for all New Zealand trials: upper North Island, lower North Island, upper South Island and lower South Island. Total and seasonal yields are presented. These summaries have been uploaded into AgYields (e.g. dataset www.agyields.co.nz/dataset/447). In this case the temporal (annual/ seasonal yield and dates) information is well described but some challenges were encountered entering summaries referring to location. For instance, which region should be considered when there is an upper South Island yield (e.g. Nelson or Marlborough)? How representative are those summaries in such a large area from a farm level viewpoint? It would be more accurate if NVFT data were added into AgYields as individual sites rather than using regional references.

A systematic and detailed on-farm method to measure pasture growth rates by cutting samples from movable cages using a mower was presented by Radcliffe (1974). This approach was applied in experiments across New Zealand to understand the seasonal growth patterns in grass-clover pastures. Similarly for cereals, researchers, and consultants conduct field surveys to record cereal yields. The samples are collected from representative areas within a field by cutting plants at specific growth stages (e.g. silking, maturity) and measuring biomass and grain yield (PBRA 2023). These surveys provide valuable data on crop performance. By combining harvested area information (ideally following the spatial

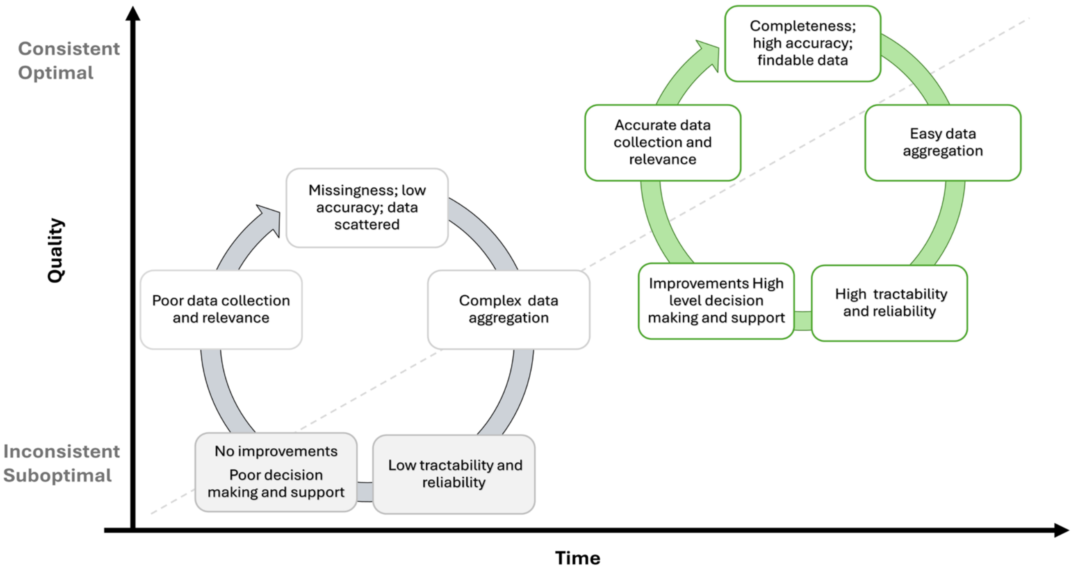
Consistent
Optimal
Quality
Inconsistent
Suboptimal
Time
Missingness; low accuracy; data scattered
Poor data collection and relevance
Complex data aggregation
No improvements Poor decision making and support
Low tractability and reliability
Completeness; high accuracy; findable data
Accurate data collection and relevance
Easy data aggregation
Improvements High level decision making and support
High tractability and reliability

Figure 3 Simplified representation of how sustained efforts (over time, x axis) from data collection to analysis reaches an optimal level (quality, y axis). Adapted from Voytek (2016).

criteria specified with yield estimates), total production can be calculated (Statistics New Zealand 2021). To prevent unfairness and inefficiency, researchers and farmers must maximise completeness of their data collection procedures first (Figure 3) to ensure data quality (Chapman 2005) and prevent missing data (Łopucki et al. 2022). A series of fact sheets and video-tutorials have been developed and publicised to promote among the New Zealand farming groups the existing techniques, the basics and standard protocols for data collection (Beef and Lamb New Zealand 2017, AgYields 2023).

Data completeness, accuracy, and consistency are valuable metrics for assessing data integrity and quality (Cong et al. 2007). There are many challenges in ongoing data quality such as: modelling and management, quality control and assurance, analysis, storage and presentation (Chapman 2005). The approach used to handle each one of these issues depends on the application and the level of data quality required for the intended use. The analysis of these Pasture and Crop datasets shows which variables have been well reported and those which are poorly described and therefore potentially compromise the data quality. Schafer (1999) stated that a missing rate of 5% or less is negligible. According to Bennett (2001) a statistical analysis is likely to be biased when more than 10% of data are missing. The amount of missing data is not the single criterion by which one assesses the missing data problem. The missing data mechanisms and patterns have greater impact on research results, for example, than does the proportion of missing data (Dong and Peng 2013).

The absence of a system to find and preserve raw data, combined with their current widespread loss, impedes scientific advancement (Voytek 2016). Therefore, a database such as AgYields is a useful tool to avoid data loss and improve data findability and quality.

Conclusions and future implications

The objective of many sectors depending on data inputs from research is to create a virtual data collection-process and analysis system. Lessons can be gained by the agri-food industry in New Zealand from the social and medical sectors (O'Connor et al. 2022) which rely heavily on research data to improve decision making. Crop datasets tend to be more concise and less complex (mostly monoculture and one sowing date) compared with the pasture datasets which have species mixtures and different sowing dates (autumn, spring sown). This paper has listed the base information using AgYields as an example of a template required to maximize the value and future proof data collected. Information about site altitude, latitude, longitude and water use in irrigation are moderate-poorly reported. Experiments, and on farm data collection can be improved by accurately recording site and date of collection. Referees of peer reviewed journals are encouraged to ensure base data are provided to minimize 'missingness' during data collection and reporting. Missing data limits interpretation, making it difficult to draw meaningful and trustworthy conclusions.

ACKNOWLEDGEMENTS

Funding was provided by Beef + Lamb New Zealand under the research program: “Advancing AgYields to support forage/crop decision making” (Project Number: 21195-3), Kathleen Spragg Research (G-202305-05830), the “Extension tools for Hill Country Futures’ (Project Number: 21126).

REFERENCES

AgYields ND. 2023. FAQs AgYields National Database. Accessed: 20/02/2024. https://www.agyields.co.nz/assets/AgYieldsHelpGuide_Aug2023.pdf

Beef and Lamb New Zealand. 2017. Measuring pasture on hill country. Fact Sheet December: 1-4. Accessed: 15/02/2024 https://beeflambnz.com/knowledge-hub/PDF/measuring-pasture-growth-rates.pdf.

Bennet DA. 2001. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health 25*: 464-469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Carolan L, Smith, FS, Protonotarios V, Schaap B, Broad E, Hardinges J, Al. E. 2015. How can we improve agriculture, food and nutrition with open data? Open Data Institute. Accessed: 10/01/2024.www.godan.info/documents/how-can-we-improve-agriculture-food-and-nutrition-open-data.

Carroll S, Garba I, Figueroa-Rodríguez O, Holbrook J, Lovett R, Materechera S, Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R. 2020. The CARE principles for indigenous data governance. *Data Science Journal 19*. https://doi.org/10.5334/dsj-2020-043

Chapman AD. 2005. Principles of Data Quality. *Report for the Global Biodiversity Information Facility*, Copenhagen: 1-57.

Chapman DF, Bryant JR, Olayemi ME, Edwards GR, Thorrold BS, McMillan WH, Kerr GA, Judson G, Cookson T, Moorhead A, Norriss M. 2017. An economically based evaluation index for perennial and short-term ryegrasses in New Zealand dairy farm systems. *Grass and Forage Science 72*: 1-21. https://doi.org/10.1111/gfs.12213

Chapman D, Cosgrove G, Kuhn-Sherlock B, Stevens D, Lee J, Rossi L. 2019. Scaling issues in the interpretation of dry matter yield differences among perennial ryegrass (*Lolium perenne*) cultivars. *Journal of New Zealand Grasslands 81*: 209-216. https://doi.org/10.33584/jnzg.2019.81.410

Cong G, Fan W, Geerts F, Jia X, Ma S. 2007. Improving data quality: Consistency and accuracy. *Proceedings of the 33rd international conference on very large data bases*: 315-326.

Daft, RL and Lengel, RH. 1986 Organizational information requirements, media richness and structural design. *Management Science*: *32* (5), 554-571. http://dx.doi.org/10.1287/mnsc.32.5.554

Dong Y, Peng CY. 2013. Principled missing data methods for researchers. *SpringerPlus 2*: 1-17. https://doi.org/10.1186/2193-1801-2-222

Easton S, Baird D, Baxter G, Cameron N, Hainsworth R, Johnston C, Kerr G, Lyons T, Mccabe R, Nichol W, Norriss M. 1997. Annual and hybrid ryegrass cultivars in New Zealand. *Proceedings of the New Zealand Grassland Association 59*: 239-244. https://doi.org/10.33584/jnzg.1997.59.2248

FAO. 2021. Land use statistics and indicators Global, regional and country trends 1990-2019. *Faostat Analytical brief 28*: 1-14. Accessed:11/01/2024. https://www.fao.org/3/cb6033en/cb6033en.pdf

Ferris L, Rahman Z. 2017. Responsible data in agriculture. *F1000Research 2*: 1306. https://doi.org/10.7490/f1000research.1114555.1

Gandar PW, Kerr JP. 1980. The efficacy of agronomic research in New Zealand. *Proceedings of the Agronomy Society New Zealand. 10*: 87-92.

Guerra-García C, Nikiforova A, Jiménez S, Perez-Gonzalez HG, Ramírez-Torres M, Ontañon-García L. 2023. ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design. *Data & Knowledge Engineering 145*: 102152. https://doi.org/10.1016/j.datak.2023.102152

Hendy J, Ausseil AG, Bain I, Blanc É, Fleming D, Gibbs J, Hall A, Herzig A, Kavanagh P, Kerr S, Leining C. 2018. Land-use modelling in New Zealand: current practice and future needs. Motu Working Paper 18-16 *Motu Economic and Public Policy Research*: 1-70.

Lee JM, Chapman DF, Wims CM, Griffiths WM, Popay AJ, Wilson DJ, Bell NL. 2018. Implications of grass–clover interactions in dairy pastures for forage value indexing systems. 2. Waikato. *New Zealand Journal of Agricultural Research 61*: 147-173. Taylor & Francis. https://doi.org/10.1080/00288233.2017.1394330

Liu YN, Li JZ, Zou ZN. 2016. Determining the real data completeness of a relational dataset. *Journal of Computer Science and Technology 31*: 720-740. https://doi.org/10.1007/s11390-016-1659-x

Łopucki R, Kiersztyn A, Pitucha G, Kitowski I. 2022. Handling missing data in ecological studies: Ignoring gaps in the dataset can distort the inference. *Ecological Modelling 468*: 109964. https://doi.org/10.1016/j.ecolmodel.2022.109964

Malaverri JEG, Medeiros CB. 2012. Data Quality in Agriculture Applications. *Proceedings XIII*

GEOINFO. Campos do Jordao. Accessed: 11/01/2024. http://mtc-m16c.sid.inpe.br/col/sid.inpe.br/mtc-m16c/2015/12.17.17.09/doc/proceedings2012_p14.pdf.

Moot DJ, Griffiths WM, Chapman DF, Dodd MB, Teixeira CSP. 2021. AgYields - a national database for collation of past, present and future pasture and crop yield data. *New Zealand Grasslands Association 83*: 15-22. https://doi.org/10.33584/jnzg.2021.83.3512

PBRA. 2023. New Zealand Plant Breeders Research Association (PBRA). Accessed:28/10/2023. https://www.pbra.co.nz/.

Percival NS. 1980. Cool-season growth responses of kikuyu grass and ryegrass to gibberellic acid. *New Zealand Journal of Agricultural Research 23*: 97-102. https://doi.org/10.1080/00288233.1980.10417851

Radcliffe JE. 1974. Seasonal distribution of pasture production in New Zealand I. Methods of measurement. *New Zealand Journal of Experimental Agriculture 2*: 337-340. https://doi.org/10.1080/03015521.1974.10427692

Runck B, Jogleka A, Silverstein KA, Chan-Kang C, Pardey PG, Wilgenbusch JC. 2022. Digital agriculture platforms: Driving data-enabled agricultural innovation in a world fraught with privacy and security concerns. *Agronomy Journal 114*: 2635-2643. https://doi.org/10.1002/agj2.20873

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*: p.e0274677. https://doi.org/10.1371/journal.pone.0274677

Schafer JL. 1999. Multiple imputation: a primer: 8(1); pp.3-15. *Statistical Methods in Medical Research 8*: 3-15. https://doi.org/10.1177/096228029900800102

Sierra JC, Cerón-Souza I, Avellaneda YA, Muñoz EAM, Martínez JDJ V. 2023. Phenotypic variation of Kikuyu grass (*Cenchrus clandestinus*) across livestock production farms in Colombian highlands is explained by management and environment rather than genetic diversity. *Crop and Pasture Science*. https://doi.org/10.1071/CP22360

Statistics New Zealand. 2021. 2020 Agricultural data for Northland Regional Council. Accessed:28/01/2024. http://infoshare.stats.govt.nz/infoshare/Default.aspx.

Teixeira CSP, Gee TM, Hawke MF, Moot DJ. 2023. Pasture production: a compilation of historical datasets from farms in Bay of Plenty. *Journal of New Zealand Grasslands 85*: 17-28. https://doi.org/10.33584/jnzg.2024.85.3600

Teixeira C. SP, Olykan ST, Moot DJ. 2023a. A review of pasture yields and growth rates in Northland. *New Zealand Journal of Agricultural Research*: 1-20. https://doi.org/10.1080/00288233.2023.2194027

Thom; ER, Waugh CD, McCabe RJ. 1998. Growth and persistence of perennial and hybrid ryegrasses when grazed by dairy cows in the central Waikato. *New Zealand Journal of Agricultural Research 41*: 477-486. https://doi.org/10.1080/00288233.1998.9513331

Voytek B. 2016. The virtuous cycle of a data ecosystem *PLoS Computational Biology 12*: e1.005037. https://doi.org/10.1371/journal.pcbi.1005037

Weersink A, Fraser E, Pannell D, Duncan E, and Rotz S. 2018. Opportunities and challenges for big data in agricultural and environmental analysis. *Annual Review of Resource Economics 10*: 19-37. https://doi.org/10.1146/annurev-resource-100516-053654

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. 2016.The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*: 1-9. https://doi.org/10.1038/sdata.2016.18

Yi J, Cohen S, Rehkamp S, Gomez MI, Ge H, Jaromczyk J. 2021. Strengthening our understanding of the food energy water nexus through completing suppressed datasets. *Agricultural & Applied Economics Association Annual Meeting*, Austin, TX, August 1 – August 3.

Supplement 1 The base information required on the AgYields National Database data entry mode.

Reference details

Published/unpublished*	Title*	Author(s)*	Journal	Pub Year*	Volume	Pages	Url	DOI	Description/Notes

Any additional information related to the study (indicate Yes (Y))

Met Files	Photos	Soil Water Data	Raw Data

Site details and attributes

Region*	Location Name	Site Name*	Latitude	Longitude	Altitude	Soil Types	Is it pasture/crop?*	Is it resident/sown?	Sowing Date	Harvest Method

Irrigation Level	kgNyr	kgPyr	kgKyr	kgSyr	kgLimeyr	Defoliation Method	Dominant Species	Cultivar	Flowering Date	Additional Species	Cultivars

Experiment details and datagrid

Experiment Name	Measurement Unit	What yield are you recording? *	Data Source	Start Date	End Date	DM Yield	Grain Yield	Growth Rate	Annual Yield

Note: The Irrigation, fertilizer and Species information can be either a site attribute or experimental treatment. * Compulsory entry.

Metadata available at: https://www.agyields.co.nz/assets/AgYieldsHelpGuide_Aug2024.pdf