Forage Data Hub – a platform for sharing valuable datasets for resilience

Ashworth, A.J.[*]; Marshall, L.[†]; Volenec, J.[‡]; Berti, M. [§]; van Santen, E. [‖]; Williams, C. [#]; Gopakumar, V. [††]; Foster, J. [‡‡]; Picasso, V. [#]; Su, J. [†].

[*] USDA-ARS, Poultry Production and Product Safety Research Unit, Fayetteville, AR, 72701, USA

[†] University of Texas at Arlington, Arlington, TX, 76019, USA

[‡] Purdue University, West Lafayette, IN, 47907, USA

[§] North Dakota State University, Fargo, ND, 58105, USA

[‖] University of Florida, Gainesville, FL, 32611, USA

[#] University of Wisconsin-Madison, Madison, WI, 53706, USA

[††] University of Texas at Dallas, Richardson, TX, 75080, USA

[‡‡] Texas A&M AgriLife Research, Beeville, TX 78102, USA

Keywords: database; legacy data; data hub; diverse perennial circular systems; annual forage systems

Abstract. In accord with the necessity to enhance ecosystem services and productivity in food systems, is the increase of data availability at multiple scales and over time. To help meet this need, we discuss the development of a National Forage Data Hub which will act as a platform to curate, share, and analyze data pertaining to forage systems. This centralized hub will leverage existing datasets by bridging multiple sources including forage crop—soil, water, and nutrient availability—yield (animal and crop) potential (and gaps)—climate—management systems at high spatial and temporal resolution enabling system interaction assessments through next-generation analytics. This novel approach to existing datasets will integrate Big Data at the soil-water-plant-animal-climate nexus to advance data storage technology systems for multiple trophic-level research projects.

Introduction

As the pressure to develop ecologically and economically sustainable agricultural systems grows, so too does the necessity to broaden the pool of publicly available agronomic data. A significant proportion of past and present research output has served purpose to fuel at most only a handful of studies; immense potential is lost as these valuable datasets retire or are forgotten. Accumulating and repurposing such data into a central repository will aid in minimizing research redundancy and allow for recent advancements in computing capacity and data sciences to have maximal impact on the field of agriculture. With enough contribution, such a data repository, or data hub, can be leveraged to glean new insight into the key attributes of productive, stable, and resilient agricultural systems. To work towards meeting this end, this paper outlines the creation of a Forage Data Hub, and its subsequent use to informally assess primary productivity and resiliency differences between annual and perennial forage systems.

Methods

The initial data compilation phase of the project consisted of careful consideration and deliberate development of a primary datasheet. The data to be ingested into the Forage Data Hub was expected to be highly variable – in volume, extent, variability, granularity, species, etc. – and thus it was crucial to define a robust set of variables which would accommodate a wide variety of datasets, while at the same time ensuring data interoperability. The Agronomy Ontology (AgrO) provides a collection of semantically organized terms from the agronomy domain which can be used to facilitate the collection, storage, and use of agronomic data, enabling easy interpretation and reuse of the data by humans and machines alike (CGIAR, 2020). As such, it was a valuable resource to consider with respect to the development of the Forage Data Hub's primary datasheet. A set of minimum and preferred data requirements (Table 1) adhering to ontology-based principles were defined, and collaborators began compiling data which met these requirements.

Following several months of the ongoing data compilation process, stock of the data was taken and a sample analysis was performed to verify the value of the data repository. Up until the time of the analysis, the datasheet had accumulated a total of 37,320 data entries. These data were programmatically searched for missing entries in required data fields, and any data with such errors were removed. The remaining 36,603 data entries were subsequently scanned for non-uniformities, misspellings, and other minor errors, which then were corrected automatically using the Python programming language.

With the aim of ensuring an honest comparison between the resiliency of annual and perennial forage systems, only a subset of the remaining data was selected for analysis. The set of harvest years for which the current dataset had entries from both annual systems and perennial systems was determined, and only data reporting yields from harvests which took place during one of these years were selected. Such a selection is meant to aid in restricting calendar year dependent confounds from skewing results towards whichever group includes data from the more climatically extreme year. After the data were split into two groups – one consisting of data from perennial systems, and the other consisting of data from annual systems – the mean dry matter yield across all years and locations was computed for each group in order to assess baseline differences in primary productivity.

Table 1. Required and recommended data and metadata describing forage data collection and processing.

Forage data parameter	Metadata description	Status
Location	Laditude and longitude (NDAD 83)	Required
Perennial/Annual/Biannual designation	Binary dropdown	Required
Mixture#	Select 1-4 species dropdown†	Required
Planting_Date	Planting date (year)	Required
Harvest_Date_Year	Date harvested (year)	Required
Treatment	Treatments tested corresponding to yield reported (select drop down treatment¶)	Required
Replicate #	Replicate # (if mean data, replication=# of observations comprising mean)	Required
Biomass Yield	Dry matter yield in units (Mg, ha)	Required
Least Significant Difference	Least significant difference at 5% probability	Required
Irrigation_Volume	Irrigation volume (cm, yr) applied during the experimental year. If none, "0"	Required
NPK_Rate Applied	Total annual rate, kg, ha. Add 'N/A' if rate is unknown.	Required
Data_Type	Yield reported either plot level or mean	Required
Livestock	Livestock present, yes or no	Required
Grain_Yield	Grain yield Mg ha^{-1} (add moisture % in metadata)	Preferred
Variety	If a mixture, then add cultivar(s) for corresponding species.	Preferred
Harvest#	Harvest number within year corresponding to reported yield.	Preferred
Soil_Series	Offical soil series/mapping unit	Preferred
CV	Coeffient of variation (CV)	Preferred
Relative_Proportion_Species	Relative proportion of each species (in a mixture), %.	Preferred
Cutting_Height	Forage cutting height, cm	Preferred
Plot_Area	Experimental unit size, m2	Preferred
Harvested_Area	L x W, m2	Preferred
Precipitation	Corresponding to planting-harvest period reported, total, cm. (NOAA station in metadata).	Preferred
Temprature	Corresponding to planting-harvest period reported, average, C. (NOAA station in metadata).	Preferred
Forage quality (CP, ADF, and NDF)*	Oven dry matter, g/kg, (provide methods in metadata).	Preferred
Soil properties (C, N, P, K, nitrate, pH, BD)††	N-P-K, kg, ha; (provide methods in metadata along with sampling depth corresponding to data).	Preferred
Biotic_Pest_Control_Total#	Biotic pest control, list common chemical name in metadata (g ai ha^{-1}) and # of chemicals applied during harvest period in datasheet	Preferred
Grazer_Density	If yes to livestock present, then (animal unit per ha)	Preferred
Grazing_Duration	Animal grazing days per experimental year	Preferred
Stocking_Method	continous, rotational, mob grazing	Preferred
Ruminant_type	Ruminant type (cattle, sheep, other)	Preferred
Weight_Gain	Live weight gain kg, ha^{-1}	Preferred
Additional_Data	Additional response data (yield, carbon, etc.)	Preferred

¶ options include: fertility trial (N, P, or K), variety trial, mixture/intercrop trial, grazing, pesticide, cutting height, cover crop, irrigation, "other", and "N/A". Users are prompted to provide additional treatment data in the 'metadata' tab.
† Based on the assumption that >4 species in a diverse mixture will result in dominant species accounting for at least 50% biomass (Ashworth et al., 2018). Required to list dominant species in mixture 1-4.
* CP=crude protein, ADF=acid detergent fiber, and NDF=neutral detergent fiber.
†† C=carbon; N=nitrogen; P=phosphorus; K=potassium; BD=bulk density.

To measure the differences in resiliency between the two groups, a method adapted from *Resilience, Stability, and Productivity of Alfalfa Cultivars in Rainfed Regions of North America* (Picasso et al. 2019) was used. A *crisis year* – the year which had the minimum mean yield across both groups and all locations – was identified from the set of years in consideration, and the mean dry matter yield for this year was computed for each group. The remaining years were designated as *normal years*, and were used to establish an expected dry

matter yield for each group under normal conditions – the group's *normal production*. The unitless *resilience ratio*, defined as the ratio of the mean yield in the crisis year to the mean yield across all normal years, was then calculated for each group and compared to assess differences between the resiliency of annual and perennial forage systems.

Results and Discussion

Of the 36,603 data entries in the initial data compilation for the Forage Data Hub, 14,225 satisfied the data selection requirements outlined in the previous section. This subset included data from 93 distinct locations across the United States, including locations in 16 different states. The included years and the number of data entries of each type reported from each year are listed in Table 2.

Table 2. Number of data entries in years with data from both groups.

Year	1991	1992	1993	1994	2001	2006	2008	2009	2010	2011	2012	2016	2017	2019
Annuals	8	48	8	128	20	12	124	180	157	338	162	20	9	11
Perennials	823	1113	238	66	2128	1375	1036	1201	1385	1368	1273	486	299	239

Out of this set of fourteen years, 1992 had the lowest mean dry matter yield across all locations, and was thus designated as the crisis year for this study. Results from this preliminary analysis appear to concur with findings from similar studies (Sanford et al. 2021) and suggest that perennial forage systems are, on average, far more resilient to extreme circumstances than their annual counterparts. On average, perennial systems maintained just over 50% of their normal production during the crisis year, while annual systems maintained only about 14% of their normal production.

Table 3. Primary productivity and resiliency differences between annual and perennial systems. The resilience ratio is the yield in the crisis year divided by the average of the yields in normal years.

Group	Mean Yield in Normal Years (Mg ha^{-1})	Mean Yield in Crisis Year (Mg ha^{-1})	Resilience Ratio
Annuals	7.2	1.0	0.14
Perennials	11.7	6.0	0.52

Though the Forage Data Hub is still in its infancy, this simple, informal analysis demonstrates the potential of a central data hub for forage systems to be used to inform risk management decisions and validate local findings on a wider scale. As the infrastructure of the Forage Data Hub grows and data continues to accumulate, there will be opportunity for more in-depth and sophisticated analyses, allowing for deeper insights into the functionality of grasslands.

Conclusions

The concurrence of the findings from our preliminary analysis with previous results from other research articles lends itself in favor of the longer-term success of the Forage Data Hub and similar central data repository systems. Beyond being used for formal research, such systems, if properly maintained, may be able serve as data sources to inform any number of agronomic decisions. Commercial producers, independent producers, and the Earth's ecosystem alike, all stand to benefit from wider agronomic data availability and standardization.

Future plans include the development of a web-based system for submitting, accessing, and visualizing data, as well as further automation of data management and integration of a greater variety of data (e.g. climate data, soil data) into the hub. As these plans are realized and data continues to accumulate, so too will the potential to harness the immense capabilities of modern data science in favor of the field of agriculture.

References

CGIAR. (2020). Responsible data guidelines: Managing privacy and personally identifiable information in the research project data lifecycle. CGIAR Platform for Big Data in Agriculture. https://bigdata.cgiar.org/responsible-data-guidelines/

Picasso, Valentin D., Casler, Michael D., and Undersander, Dan. 2019. Resilience, Stability, and Productivity of Alfalfa Cultivars in Rainfed Regions of North America. Crop Science, vol. 59, no. 2, 2019, pp. 800–810., https://doi.org/10.2135/cropsci2018.06.0372.

Sanford, G. R., Jackson, R. D., Booth, E. G., Hedtcke, J. L., & Picasso, V. (2021). Perenniality and Diversity Drive output stability and resilience in a 26-year Cropping Systems Experiment. *Field Crops Research*, *263*, 108071. https://doi.org/10.1016/j.fcr.2021.108071