Building a Community-Based FAIR Metadata Schema for Brazilian Agriculture and Livestock Trading Data

Filipi Miranda Soares[1,2,8,*,†], Fernando Elias Corrêa[1,3,†], Luis Ferreira Pires[2], Luiz Olavo Bonino da Silva Santos[2,4,†], Debora Pignatari Drucker[1,5,†], Kelly Rosa Braghetto[1,6,†], Dilvan de Abreu Moreira[1,7,†], Alexandre Cláudio Botazzo Delbem[1,7], Roberto Fray da Silva[1,3], Celso Oviedo da Silva Lopes[1] and Antonio Mauro Saraiva[1,3,8,†]

[1] *University of Sao Paulo, Center for Artificial Intelligence (C4AI), Av. Prof. Lúcio Martins Rodrigues 370, 05508020 São Paulo SP Brazil*
[2] *University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), Drienerlolaan 5, 7522 ME Enschede Netherlands*
[3] *University of Sao Paulo, Institute of Advanced Studies, R. do Anfiteatro 513, 05508060 São Paulo SP Brazil*
[4] *Leiden University Medical Center (LUMC), Albinusdreef 2, 2333 ZA Leiden Netherlands*
[5] *Embrapa Digital Agriculture, Av. Dr. André Tosello 209, 13083-886 Campinas SP Brazil*
[6] *University of Sao Paulo, Institute of Mathematics and Statistics, R. do Matão 1010, 05508090 São Paulo SP Brazil*
[7] *University of Sao Paulo, Institute of Mathematics and Computer Sciences (ICMC), Av. Trab. São Carlense 400, 13566590, São Carlos SP Brazil*
[8] *University of Sao Paulo, Polytechnic School, Av. Prof. Luciano Gualberto 380, 05508010 São Paulo SP Brazil*

Abstract

In this paper, we discuss how we are using metadata schemas and controlled vocabularies to improve interoperability between Brazilian agriculture and livestock trading data providers. A new metadata schema is being created based on a community-based approach. This method relies on knowledge from specialists to define a list of relevant metadata properties for a given domain. In the first step of the research, we extracted metadata from three datasets maintained by three Brazilian public institutions: the Center for Advanced Studies in Applied Economics (Cepea), the Institute of Applied Economic Research (Ipea), and The National Supply Company (Conab). The extracted metadata were the input to the definition of a list of 15 potential metadata properties that specialists are validating.

Keywords

Agriculture, livestock, metadata, trading data, FAIR data principles

1. Introduction

The interoperability of information systems is largely affected by the availability of standards that define common vocabularies and procedures. The large variety of incompatible data practices hinders automated data exchange. Especially in agriculture, interoperability poses a significant challenge due to the multidisciplinary and interdisciplinary nature of the field.

Digital agriculture is characterized by the massive use of technology in crop and livestock production, either by using more efficient and intelligent equipment such as mobile technologies, remote sensing services, and robotics [1], or by the use of data and software technologies, such as databases, adaptive systems, Artificial Intelligence, Big Data, Internet of Things, Virtual and Augmented Reality [1]. Although there are some advances in digital agriculture, more efforts in this

SEMANTICS 2022 EU: 18th International Conference on Semantic Systems, September 13-15, 2022, Vienna, Austria
EMAIL: filipisoares@usp.br (A. 1); fecorrea@usp.br (A. 2)
ORCID: 0000-0002-0674-7960 (A. 1); 0000-0002-2744-6104 (A. 2); 0000-0001-7432-7653 (A. 3); 0000-0002-1164-1351 (A. 4); 0000-0003-4177-1322 (A. 5); 0000-0001-6218-6849 (A. 6); 0000-0002-4801-2225 (A. 7); 0000-0003-1810-1742 (A. 9); 0000-0002-9792-0553 (A. 10); 0000-0002-7004-6042; 0000-0003-2283-1123 (A. 11)

direction are necessary as agriculture is still considered one of the least digitized productive sectors in the world [2].

In this context, the agribusiness sector, which deals with trading, among other things, is demonstrating increasing interest in digital agriculture and is capturing huge financial investments [3]. Thus, an increasing number of scientists worldwide have been committed to collecting, manipulating, and storing agribusiness data, especially data from Brazilian agribusiness, given the high Brazilian agricultural production capacity and the importance of agriculture to the country. Therefore, we believe metadata schemas that can cope with the complexity of agricultural data and have the potential to be standardized can significantly change how we produce and share information on agriculture, especially in Agribusiness.

In this paper, we present an initial effort to build a metadata schema for agricultural trading data to allow data interoperability between three of the main Brazilian data providers on agricultural commodities: the Center for Advanced Studies in Applied Economics (Cepea[2]), the Institute of Applied Economic Research (Ipea[3]), and the National Supply Company (Conab[4]). We address the challenges and issues faced and the strategies we are using to develop the semantic structure of the metadata elements.

The paper is further structured as follows: Section 2 gives the background of our work, Section 3 discusses the methodology we have been following, Section 4 presents our results so far, and Section 5 gives our conclusions.

2. Background: metadata schemas and metadata records

Metadata is data about data [4][5]. This definition, however, does not say much since it does not define which aspects of the data are represented in the metadata. As pointed out by Coyle [6], metadata are more complex than just data: they are human-made artifacts created with a purpose and to perform a function in data description. Mayernik [7] presented a literature review that included many definitions for metadata. Among them is the definition of Greenberg [4], which considered metadata as "data attributes that describe, provide context, indicate the quality, or document other object (or data) characteristics", and the definition of Smiraglia [8], which defined metadata as "structured descriptors of information resources, designed to promote information retrieval". From these definitions, we can consider metadata as attributes or properties that describe characteristics of data resources for information retrieval and to allow the correct interpretation and use of data, amongst other purposes.

Metadata and data are often registered in digital records as property-value pairs. A set of metadata is used to represent a subject, which is an entity from the real world that one wants to describe. Thereby, metadata can be represented as triples — subject-predicate-object [5]. A set of triples forms a graph, where the subjects and objects are the nodes, and the predicates are the edges.

Metadata is usually part of a metadata schema. The metadata properties in a metadata schema have two components [9]: *semantic*, i.e., the definition of the meaning of the term that can be used as a metadata field, and *syntax*, i.e., the recommendation or specification of the format of data (codification) that should be assigned to a metadata field.

With the advances of the Semantic Web and linked data, new technologies have emerged to improve the use of metadata, especially by machines. One of the most prominent technologies is the Resource Description Framework (RDF), which associates Unique Resource Identifiers (URI) to terms assigned as metadata elements. To be used in RDF parameters, namespaces have been assigned to metadata schemas to preserve the context of the metadata elements in the multiple environments in which they can be used.

In the Agriculture domain, the Food and Agriculture Organization of the United Nations (FAO) has developed the Agricultural Metadata Element Set (AGMES), which is a metadata schema that describes a variety of digital resources for agriculture [10]. Many other communities have been working on

[2] Available from: https://www.cepea.esalq.usp.br/br, last accessed 2022/07/04.
[3] Available from: http://www.ipeadata.gov.br/Default.aspx, last accessed 2022/07/04.
[4] Available from: https://sisdep.conab.gov.br/precosiagroweb/, last accessed 2022/07/04.

developing ontologies and other kinds of controlled vocabularies for agriculture, with more than a hundred available on AgroPortal[5].

3. Methodology

We started by analyzing three of the largest Brazilian agricultural commodities datasets to understand which metadata they apply to their data. These three datasets were selected *a priori* based on the solid scientific reputation and recognized contribution to the Brazilian Community of the organizations behind these datasets, namely Cepea, Conab, and Ipea. These datasets contain essential statistical data from Brazil's agricultural trades. Within these data, we decided to set the scope on time-series data of market prices since all three datasets contain this kind of data.

Through this consultation, we extracted descriptive information that was available along with the data on the websites of these data providers. A specialist then validated the information obtained.

From the analysis carried out, three new processes were identified: (i) definition of the list of common descriptive words used to fulfill the metadata fields, such as product names (e.g. cotton, cattle, rice, soybeans), units of measure (e.g. box of 15kg), locations, among others; (ii) harmonization of the data for the product name metadata field, by adopting the controlled vocabulary Agrotermos, an agriculture thesaurus that has the largest base of agriculture terminology in Brazilian Portuguese, and is a terminology provider for Agrovoc, the FAO multilingual thesaurus; (iii) based on the most representative descriptors common within three datasets, selection of the first list of 15 potential metadata properties for agricultural commodities. In this step, we gave names and definitions to these metadata properties.

We are now querying the community of practice about this first set of metadata properties to validate the terms and definitions and suggest other terms that could be significant to the domain. For this purpose, specialists from nine Brazilian organizations participating in the 5th Brazilian Open Government Action Plan[6] were asked to annotate sample datasets with the 15 metadata properties. In addition, they will answer a System Usability Scale (SUS) questionnaire adapted from the model presented by Tullis & Albert [11], originally developed to evaluate information systems' interfaces.

4. Results

The complete list of metadata is under review, so the first version of the metadata schema is not available on RDF or any other Semantic Web format yet. The 15 metadata properties we used to illustrate our approach and their definitions are the following:

- *Product group*: types of products. The best practice is to use a controlled vocabulary such as Agrotermos. Examples: grain, vegetable, meat.
- *Theme*: main topic investigated in the statistical operation. The best practice is to use a controlled vocabulary. Examples: regional average price, indicator, index, cost, technical or zootechnical coefficient.
- *Product name*: name of the agricultural or livestock product. The best practice is to use a controlled vocabulary such as Agrotermos. Examples: soy, corn, cattle.
- *Verbatim name*: natural language name given to the data series in the original dataset.
- *Publisher*: entity responsible for making the resource available[7].
- *Creator*: entity responsible for creating the resource[8].
- *References*: related resources that are referenced, cited, or otherwise pointed to by the described resource[9].
- *Data type*: kind of the data that make up the record. The best practice is to use a controlled vocabulary. Examples: multivariate, univariate, synthetic, sequential, time series, text.

[5] Available from: http://agroportal.lirmm.fr/, last accessed 2022/09/02.
[6] Available from: https://wiki.rnp.br/pages/viewpage.action?pageId=155657461, last accessed 2022/07/04.
[7] Available from: http://purl.org/dc/terms/publisher, last accessed 2022/07/04.
[8] Available from: http://purl.org/dc/terms/creator, last accessed 2022/07/04.
[9] Available from: http://purl.org/dc/terms/references, last accessed 2022/07/04.

- *Frequency*: temporal frequency of data publication. Examples: daily, monthly, bimonthly, yearly.
- *Unity*: measure or quantity of the product. Examples: 'arroba' (Brazilian unit of weight, equivalent to 15kg, represented by the symbol @); a bag of 30kg.
- *Location*: place or region to which the data series refers. The best practice is to use geographic coordinates (decimal latitude and longitude) or insert names according to Geonames[10].
- *Start of reference period*: date of publication of the first dataset of a data series. The best-recommended practice is to adopt date encoding schemes such as ISO 8601 or ABNT NBR 5892.
- *End of reference period*: date of publication of the last dataset of a data series. The best-recommended practice is to adopt date encoding schemes such as ISO 8601 or ABNT NBR 5892. In the case of a current series, this field must be left empty.
- *License*: A legal document that defines the policy to do something with the resource[11]. The best-recommended practice is to adopt a URI to indicate the license.
- *Methodology*: summary of the methods used to generate the dataset. The best practice is to indicate the resource URI published in an open Web access format[12].

The *Product group* property has been defined to group data records into broader product categories, while the *Product name* property specifies the subgroup of the product. In a metadata record about carrots, for example, the *Product group* would be *vegetable*, and the product name would be *carrot*. We are using the Agrotermos thesaurus to standardize data for product group and name properties. However, although Agrotermos has a huge base of agriculture terminology, some important terms from the trade's domain are still missing, such as, for example, 'boi gordo' (Portuguese for fed cattle), which is a very popular term to designate a specific kind of cattle in the Brazilian trade market. Therefore, we are working with the Agrotermos Developing Team[13] at the Brazilian Agricultural Research Corporation (Embrapa) to include missing terms in Agrotermos.

Besides the controlled vocabulary fields, the *verbatim name* property allows the name given in the original dataset to be preserved. The *Theme* property indicates the kind of trade indicators, which is deeply related to the *methodology* property since each kind of trade indicator follows a possibly different method.

Publisher and *Creator* were imported from the Dublin Core schema. The *Publisher* property can be used to refer to the institution that published the datasets (e.g., Cepea, Ipea, Conab). In contrast, the *Creator* property refers more specifically to the person or department within those institutions responsible for generating the datasets. The *References* property was also imported from Dublin Core and gives the link to the address of the dataset on the Web. The *License* property was another Dublin Core term we incorporated into our metadata schema. It describes the license of use of the dataset given by the data provider.

Other properties, i.e., *Data type*, *Frequency*, *Unity*, *Location*, *Start of reference period*, and *End of reference period* are often variants of terms already present in generic metadata schemas, such as Schema.org[14], Wikidata properties[15], and the Data Catalog Vocabulary (DCAT[16]), so in a later step of the research, we will identify and sort out which of these generic metadata vocabularies should be referred to in our metadata schema.

The metadata schema presented in this paper has been named *Agriculture and Livestock Metadata Elements Set* (Almes), or Almes Core (short name). The name hints at a broader scope than the metadata schema presented in this paper, which only focuses on trading data. This is because we intend to extend the metadata schema to cover other sub-disciplines of agriculture and livestock in the upcoming years of the project.

[10] Available from: http://www.geonames.org/, last accessed 2022/09/03.
[11] Available from: http://purl.org/dc/terms/license, last accessed 2022/07/04.
[12] Available from: https://metadados.ibge.gov.br/consulta/glossario/estatistico, last accessed 2022/07/04.
[13] The full name of the Team in Portuguese: Comissão Permanente de Trabalho em Vocabulários Controlados, Agroterminologias e Agrossemântica da Embrapa (GTermos).
[14] Available from: https://schema.org/, last accessed 2022/07/04.
[15] Available from: https://www.wikidata.org/wiki/Wikidata:List_of_properties, last accessed 2022/07/04.
[16] Available from: https://www.w3.org/TR/vocab-dcat-2/, last accessed 2022/07/04.

5. Final considerations

We expect the metadata schema we are developing will improve the interoperability within agriculture and livestock trading data providers in Brazil and that it will also be useful for similar purposes in other countries after being extended. We also expect that the methods to make our metadata schema FAIR applied in this research can be reused by other initiatives with similar purposes.

In the following steps, we will review the metadata properties based on suggestions from the community of practice. After that, we move forward to publish the metadata schema and all the documentation needed so that the community can use it. The FAIR principles will be applied in the publication process to make the metadata schema findable on the web, accessible without restrictions, interoperable with other metadata schemas, and reusable by humans and machines [12]. The FAIR Data Points [13] will be implemented to enforce the FAIRness of the data repositories containing harmonized datasets described with the Almes Core properties.

Agriculture is a vast subject, so we expect to address other sub-disciplines to create other metadata sets in the future. Given the country's considerable biodiversity and climate conditions, Brazilian agriculture has many peculiarities, so we expect to develop controlled vocabularies in Portuguese (with translations to English) that can better suit the needs of the Brazilian agriculture and livestock data community.

6. Acknowledgments

FMS thanks the São Paulo Research Foundation (FAPESP) for the research grants n. 2021/15125-0 and 2022/08385-8. The authors thank the Center for Artificial Intelligence (C4AI), a joint initiative of USP, IBM and FAPESP (process n. 2019/07665-4), for the technical and/or financial support to their research. AMS thanks the Brazilian National Council for Scientific and Technological Development (CNPq) for the research grant n. 312605/2018-8.

7. References

[1] Shamin, A., Frolova, O., Makarychev, V., Yashkova, N., Kornilova, L., Akimov, A. Digital transformation of agricultural industry. IOP Conference Series: Earth and Environmental Science, 346(1), (2019). https://doi.org/10.1088/1755-1315/346/1/012029

[2] Saraiva, A. M., Costa, W. F., Xavier, F., Albertini, B. de C., Pfeifer, R. A. C., Júnior, M. A., Simplício, A. K. V. Dados na Agricultura Digital: ciclo, padronização, qualidade, compartilhamento e segurança. In Daniel Marçal Queiroz, D. S. M. Valente, F. de A. C. Pinto, & A. Borém (eds.).: Agricultura Digital. UFV, Viçosa, pp. 308–325 (2020).

[3] Duncan, E., Rotz, S., Magnan, A., Bronson, K. Disciplining land through data: The role of agricultural technologies in farmland assetisation. Sociologia Ruralis 62(2), 231–249 (2022). https://doi.org/10.1111/SORU.12369

[4] Greenberg, J. Understanding Metadata and Metadata Schemes. Cataloging & Classification Quarterly, 40(3-4), 17–36 (2005). https://doi.org/10.1300/J104V40N03_02

[5] Pomerantz, J. Metadata. The MIT Press, Cambridge (2015).

[6] Coyle, K. Understanding Metadata and Its Purpose. The Journal of Academic Librarianship, 31(2), 160–163 (2005). https://doi.org/10.1016/J.ACALIB.2004.12.010

[7] Mayernik, M. S. Metadata. Knowledge Organization, 47(8), 696–713 (2020). https://doi.org/10.5771/0943-7444-2020-8-696

[8] Smiraglia, R. P. Metadata: a cataloger's primer. Routledge, Abingdon-on-Thames (2005).

[9] Chan, L. M., Zeng, M. L. Metadata Interoperability and Standardization - A Study of Methodology Part I. D-Lib Magazine, 12(6), (2006). https://doi.org/10.1045/june2006-chan

[10] Food and Agriculture Organization of the United Nations (FAO). AgMES 1.1 Namespace Specification (2010). http://aims.fao.org/standards/agmes/namespace-specification, last accessed 2022/07/04.

[11] Tullis, T., Albert, B. Measuring the User Experience. 2nd ed. Elsevier, Amsterdam, (2013).

[12] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 1–9 (2016). https://doi.org/10.1038/sdata.2016.18

[13] Bonino, L., Kaliyaperumal, R., Kees Burger, A., Kuzniar, A. Gavai. FAIR Data Point Specification (2021), https://github.com/FAIRDataTeam/FAIRDataPoint/wiki, last accessed 2022/07/04.