A Semantic Data Model for a FAIR Digital Repository of Heterogeneous Agricultural Digital Objects

The Case of the EUREKA FarmBook

Hercules Panoutsopoulos[1], Christopher Brewster[2] and Spyros Fountas[1]

[1]*Department of Natural Resources Management and Agricultural Engineering, Agricultural University of Athens, 75 Iera Odos St, Athens, 11855, Greece*

[2]*Institute of Data Science, Maastricht University, Maastricht, Netherlands, & Data Science Group, TNO, Soesterberg, Netherlands*

Abstract

During the past few years a significant number of agriculture-related research and development projects have been implemented by receiving funding from the European Commission. All these projects have aimed to address specific problems and have produced solutions documented in the digital objects created in their context. However, the uptake of the existing information and knowledge by the concerned stakeholders is not yet adequate. On the other hand, the divergence of the available digital objects in terms of their types and formats poses significant challenges in homogeneously describing them with metadata. The Horizon 2020 EUREKA project aims to make a contribution towards this direction by developing a FAIR digital repository based on a semantic data model. The present document focusses on the design decisions related to it, as well as the rationale for the need to develop FAIR digital repositories based on formal data models.

Keywords

Agriculture, semantic data model, digital repository, FAIR principles, semantic web standards

1. Introduction

In the seven years of implementation of the Horizon 2020 Framework Programme, the European Commission invested nearly one billion euros in projects related to agriculture, forestry, and rural development[1]. These projects have created a large number of digital objects conveying valuable information and knowledge about best and innovative practices. However, the uptake and re-use of these digital objects has not yet been realised to an adequate extent. In addition, most of these digital objects are no longer available after the end of the projects. As a result, there is little potential for the various agricultural stakeholders to have access to the available knowledge for further research and solution development. This is the context in which the EUREKA

IFOW 2021: 2nd Integrated Food Ontology Workshop, held at JOWO 2021: Episode VII The Bolzano Summer of Knowledge, September 11-18, 2021, Bolzano, Italy

hpanoutsopoulos@aua.gr (H. Panoutsopoulos); christopher.brewster@maastrichtuniversity.nl (C. Brewster); sfountas@aua.gr (S. Fountas)

0000-0002-8060-9750 (H. Panoutsopoulos); 0000-0001-6594-9178 (C. Brewster); 0000-0001-9787-6268 (S. Fountas)

[1]For details, the interested reader may refer to https://ec.europa.eu/eip/agriculture/en/about/multi-actor-projects-scientists-and-farmers

project (https://www.h2020eureka.eu/) is being situated. More specifically, EUREKA's goal is to build a meta-repository (i.e., the "FarmBook") of the digital objects created by the EU-funded, Horizon 2020 multi-actor projects. By this way, EUREKA aspires to make a contribution towards ensuring the longevity of the knowledge generated in those "source" projects and therefore, allowing it to be further circulated and used as the basis for future developments. However, in this effort, there are significant challenges to address mostly relating to the heterogeneity of the digital objects to be integrated into the meta-repository.

Semantic web technologies and standards [1], existing for quite many years, provide recommendations for the exchange and interoperability of data in the Linked Open Data (LOD) Cloud. The more recently established FAIR principles [2] have further contributed to the vision of standards-based data sharing. However, what becomes evident from existing practices is that most of the existing repositories are based on ad-hoc data modelling decisions and designs [3]. To enable the access and re-use of the variety of agricultural digital objects, the aim of EUREKA is to develop a FAIR digital repository by drawing upon semantic web standards. To this end, a semantic data model is proposed. This data model has been developed by considering and re-using concepts and relations from existing ontologies. It provides a formal structure aimed to be used for the identification and definition of metadata for describing and annotating the digital objects available from the FarmBook. In this context, the following research questions have been intended to be addressed:

- Which are the main concepts and relations needed to be captured, in our semantic data model, to identify a set of metadata for annotating the breadth of digital objects coming from the Horizon 2020 multi-actor projects?
- Which external ontologies and what components of those need to be considered and re-used for the design of the proposed semantic data model?

In the rest of the article, an overview of the issues related to the adoption of the FAIR principles for developing FAIR digital repositories is provided. Then, our methodology and the semantic data model are presented. Finally, some concluding remarks are provided together with accounts of the steps to be taken next in EUREKA.

2. Data Modelling for FAIR Digital Repositories

The FAIR data principles have been developed as guidelines for good data sharing practices, initially focussing on the publication and re-use of scientific data sets [2]. There has been a surge of interest in applying the FAIR principles to help make data management and the stewardship of research outputs easier, and pave the way towards repositories facilitating FAIR data sharing. Therefore, starting from the concept of "FAIR datasets", it is important to broaden the scope of FAIRness to "FAIR digital objects" and "FAIR digital repositories" [4].

To turn these ideas into reality, we need to establish a FAIR data ecosystem [4]. In such an ecosystem, "FAIR digital repositories" have a pivotal role to play by providing all the necessary mechanisms and tools that will enable the accessibility and re-use of "FAIR digital objects". The main assumption that has underpinned our work is that a "FAIR digital repository" (a digital repository for the housing and delivery of "FAIR digital objects") needs to be built on a data model

designed for digital objects to have explicit metadata. In this context, the re-use of concepts and relations from external, appropriately selected ontologies is considered a best practice in a semantic web paradigm. This approach is particularly important in cases of initiatives that deal with the development of meta-repositories, as in the case of EUREKA. In this application scenario, a well-designed data model, built in the spirit of the FAIR principles, can help towards homogenising the heterogeneity of the divergent types of the digital repositories and objects, and the variety of the designs, standards, and formats that may have been considered. The FAIR principles directly relevant for such a data modelling are I1 (emphasising the need for conceptual designs established upon standards-based, well-defined data models), as well as F2, I2, and R1 (setting out a list of requirements for the metadata to define based on a formal data model).

3. The EUREKA FarmBook's Semantic Data Model

3.1. Methodology

The FarmBook data model's design has been based on the methodology of Noy and McGuiness [5]. Despite the well-acknowledged rigour of contemporary ontology building methods (e.g., KNARM), or the OBO Foundry's principles[2], we have drawn upon a widely-adopted, baseline approach able to be comprehended by non-experts as well, given the EUREKA's target audience (i.e., farmers, foresters, and advisors). The following steps were involved in our model's design:

STEP 1 - Determine the domain and scope of the ontology: The overarching goal of the EUREKA FarmBook's semantic data model is to provide formal descriptions of the concepts, the concept relations, and the properties of concepts needed to be considered for defining the metadata to homogeneously characterise the breadth of digital objects available from the EU-funded, Horizon 2020 multi-actor projects. The design of the data model has been driven by the following, indicative, competency questions:

- What are the subjects (topics) of the digital objects available from project X?
- Which are the geographic locations related to the content of the digital objects of type X?
- What are the digital objects of type X on subject Y, related to the geographic location Z (e.g., the videos on pig farming in oak forests in Spain)?

STEP 2 - Consider re-using existing ontologies: The external ontologies that have been considered, in terms of re-using definitions of concepts, relations, and concept properties, for the design of our semantic data model, are Schema.org, FRAPO, FOAF, and DCMI . These are generic ontologies selected with the aim to help us model basic descriptive features of a digital object relating, among others, to the format in which information is conveyed (text, audio, video, etc.), as well as provenance (e.g., the digital object's creator and source multi-actor project). We are indeed familiar with agriculture-related controlled vocabularies (e.g., AGROVOC) and ontologies (e.g., FoodOn, AgrO), and deeply acknowledge the added value of re-using their components as part of the development of a formal data model. These structures

[2] http://obofoundry.org/principles/fp-000-summary.html

will be considered next in EUREKA for the needs of creating a flexible model of the subject(s) addressed in the various digital objects.

STEP 3 - Define the classes and class hierarchy and STEP 4 - Define the properties of the classes: The details of the definition of the model classes, their hierarchy, and the class properties are presented in the next section.

3.2. Semantic Data Model Description

The FarmBook's semantic data model is shown in Figure 1 below. The digital object available by a multi-actor project, the project providing it, its creator, and the geographic location mentioned in its content are key concepts needed to be captured in the model. A digital object's subject may be a term defined in (and thus, imported from) agriculture-related controlled vocabularies (e.g., AGROVOC) or ontologies (e.g., FoodOn, AgrO), and is denoted with a rectangular shape following a modelling convention similar to that in FoodOn. We consider more than one such structures because of the divergence in the "source" projects' focus.

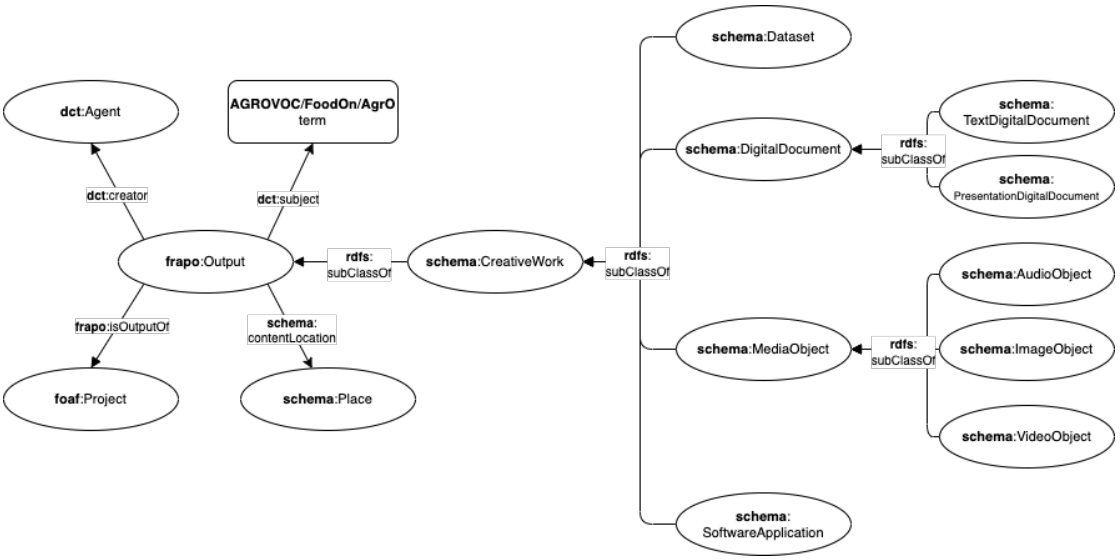

Figure 1: The EUREKA FarmBook's semantic data model.

Table 1 shows the alignment between the concepts captured in our model and the components re-used from existing ontologies. Table 2 provides information about the relations captured in our model and the ontologies they have been imported from.

To describe the full extent of the digital object categories needed to be considered for the FarmBook's design, a set of relevant concepts, organised in appropriate class hierarchies, have been imported into our semantic data model from the Schema.org ontology. These concepts are: (i) TextDigitalDocument and PresentationDigitalDocument (being subclasses of the DigitalDocument class); (ii) AudioObject, ImageObject, and VideoObject (being subclasses of the MediaObject class); (iii) SoftwareApplication; and (iv) Dataset. DigitalDocument, MediaObject,

Table 1
Concepts needed to be modelled and their links to concepts from external ontologies

Concept needed to be modelled	Concept description	Ontology considered	Component re-used
Digital object	Digital object housed in FarmBook	FRAPO	Output
Project	The digital object's source project	FOAF	Project
Creator	The digital object's creator(s)	DCMI	Agent
Geographic location	The geographic location(s) mentioned in the digital object	Schema.org	Place

Table 2
Relations needed to be modelled and their links to relations from external ontologies

Relation needed to be modelled	Ontology considered	Relation re-used
An Output is created by Agent(s)	DCMI	creator
An Output comes from Project	FRAPO	isOutputOf
An Output has a subject	DCMI	subject
An Output relates to Place(s)	Schema.org	contentLocation

SoftwareApplication and Dataset are all subclasses of the CreativeWork class, which has, in turn, been modelled as a subclass of the frapo:Output class.

This is a decision based on the fact that there are project outputs not considered as "creative work" (e.g., the network of partners established in the project). This hierarchy is illustrated in Figure 1 by viewing the diagram from right to left (i.e., moving from a lower to a higher level of abstraction in regard to the definition of our model's classes). The import of specific OWL classes and properties (namely, owl:Restriction, owl:Class, owl:onProperty, and owl:minCardinality), has allowed to infuse "heavier" semantics into our model and explicitly define cardinality restrictions over specific relations (e.g., the dct:creator relation as shown in the code snippet below).

By capturing a number of digital object properties, modelled with the help of the schema: {description, keywords, inLanguage, potentialAction, url, fileSize, dateCreated, license} and the dct: {title, format, type} properties, as well as properties of the "source" project modelled with the help of the schema: {name, alternateName, url} properties, it has been made feasible to assemble the FarmBook's semantic data model and, finally, come up with a set of metadata considered both rich and flexible enough (including administrative, structural, and provenance information) to effectively cover the needs of the EUREKA FarmBook's target end-users (namely, farmers, foresters, and advisors).

4. Discussion and Future Perspectives

Despite the considerable investments of the European Commission in agriculture-related research and development projects, the concerned stakeholders cannot easily find and access relevant information and knowledge. In addition, there is a need for good data sharing practices.

The broad spectrum of formats in which data and information is conveyed makes imperative a shift of focus from "FAIR datasets" to “FAIR digital objects” and “FAIR repositories”. The EUREKA FarmBook’s data model, designed upon well-acknowledged semantic web standards, is an initiative that paves the way towards the development of FAIR repositories of heterogeneous, agricultural digital objects. The semantic data model proposed will be used to define a set of metadata for the annotation of the digital objects, currently available and collected from the various multi-actor projects, which will be housed in the FarmBook. Moreover, it is also intended to serve as a reference model, adopted by future projects, for the description of the variety of the data and information that will be generated in their context and become available through their own, custom repositories. This will facilitate a uniform, standards-based approach in the characterisation of agricultural digital objects and enable their broader uptake and re-use.

The next steps in EUREKA involve the development of a graph-based structure for modelling the subjects addressed in the FarmBook’s digital objects, based on the re-use of components of agricultural ontologies (e.g., FoodOn, AgrO) and controlled vocabularies (e.g., AGROVOC). Furthermore, the procedures for ingesting digital objects in the FarmBook will be established. This will allow to test in practice the utility of our model and the metadata set derived from it.

Acknowledgments

The work presented in this documented has taken place in the Horizon 2020 EUREKA project receiving funding from the EU under the No 862790 Grant Agreement.

References

[1] C. Bizer, T. Heath, T. Burners-Lee, Linked data - the story so far, in: A. P. Sheth (Ed.), Semantic Services, Interoperability and Web Applications: Emerging Concepts, IGI Global, Hershey, 2011, pp. 205–227. doi:10.4018/978-1-60960-593-3.

[2] M. D. Wilkinson, M. Dumontier, ..., B. Mons, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 1–9. doi:10.1038/sdata.2016.18.

[3] A. Dunning, M. de Smaele, J. Böhmer, Are the fair data principles fair?, International Journal of Digital Curation 12 (2017) 177–195. doi:10.2218/ijdc.v12i2.567.

[4] E. Commission, Turning fair into reality: Final report and action plan from the european commission expert group on fair data, 2018. URL: https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf.

[5] N. F. Noy, D. L. McGuinness, Ontology development 101: A guide to creating your first ontology, 2001. URL: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.