RESEARCH DATA ALLIANCE

An Ontology of Data Events Based on GBIF Data Papers

Preliminary Findings

Kai Li; Jane Greenberg

Drexel University, Philadelphia PA, USA

DREXEL UNIVERSITY Metadata Research Center College of Computing & Informatics

Introduction

The data paper is an emerging genre of scientific publication that became gradually popular during the past decade. Deeply connected to the idea of data publication (Candela, Castelli, Manghi, & Tani, 2015), a data paper is a "scholarly publication of a searchable metadata document describing a particular online accessible dataset, or a group of datasets, published in accordance to the standard academic practices" (Chavan & Penev, 2011, p. 3). The data paper arguably only aims at describing data objects, which is the most notable difference between this format of scientific publication compared with classic scientific papers. This new type of scientific publication, potentially bound by academic practices and norms, calls for an examination of how it represents the processes in which datasets are created, manipulated and published, and how this pattern is different from what has been identified by empirical laboratory studies, such as the model of information transformation in scientific studies proposed by Knorr (1981):

- Instrumental mode: to decontextualize scientific results from "unnecessary" technical details; and,
- Literary mode: to recontextualize the results into constructed research purposes in scientific writings.

The present study is designed to examine how lifecycles of research data objects are represented in data papers, so that to develop an empirical-driven ontology to express data events and lifecycles described in these papers. Because of the exploratory nature of this study, we selected all the 82 data papers curated by the Global Biodiversity Information Facility (GBIF), which describe datasets that are shared in the GBIF network. A content analysis was conducted to manually identify and classify all data events inscribed in these papers. Our classification was further mapped to an existing data lifecycle. Our preliminary results are presented in this poster, accompanied by a brief discussion of the results as well as the next step of the project.

Methods

In order to pursue the research questions of this project, we selected all the 82 data papers curated by GBIF (originally available at: www.gbif.org/mendeley/data-paper). The list of papers was fetched on July 1, 2017. Due to an upgrade of their website, the original web page is no longer available, but is still accessible through the Wayback Machine service offered by the Internet Archive (https://web.archive.org/web/20170716193637/https://www.gbif.org/mendeley/data-paper).

In each paper, we identified all the sentences that are about data events happened during the preparation of the datasets. We decided to just include sentences from the method section, so that to reduce repetitive contents within the same paper. One difficulty caused by this decision is that the data papers tend to have highly variant formats: even though most papers have one to many sections to dedicate to the method information, these sections may or may not be called "Method" in the papers. To deal with this issue, we took a functional approach to defining these sections: if a section is about the methods to deal with the dataset(s), then it is counted as a method section.

The classification was developed using an inductive approach by one coder. The goal of this scheme is to identify the different types of events in terms of what is achieved in these events. After the initial work was done, the resulting scheme was reviewed by a second coder, via an inter-coder reliability scan, and discrepancies were discussed and revised accordingly. These steps were performed multiple times until both coders felt that the classification reaches saturation.

Results

In total, 533 sentences from the 82 papers are identified as the description of data events. Among these sentences, 44 sentences contain multiple data events. Based on all these sentences, a classification scheme about the nature of the data events was developed. The categories of this scheme and their definitions are displayed in Table 1. The frequencies of each category are summarized in Figure 1.

The categories in our classification were further mapped to the data lifecycle developed by the UK Data Archive ("Research Data Lifecycle," n.d.), which is shown in Figure 2. This lifecycle is selected because of its simplicity, which better suits our needs in this projects. The results of the mapping are shown in Table 2.

Category	Definition
Data analysis	The computational works based on the dataset that lead to the creation of new variables.
Data classification	The identification of samples based on taxonomies or classification schemes.
Data collection	The processes in which observations are collected and inputted into the dataset.
Data formatting	The change of the file format of the dataset.
Data identification	The assigning of identifiers to observations in the dataset.
Data manipulation	The change of data values based on a different view, scale, thesaurus, or specification.
Data removal	The deletion of observations in the dataset.
Data sampling	The creation of new samples within the dataset.
Data sharing	The publication of the dataset in any public channels.
Data validation	The checking and correcting of data points for quality control purposes.
Data visualization	The creation of visual representations based on the data values.
Databasing	The creation of the architecture of the database.
Georeferencing	The association with a map (and physical geological names) to spatial locations, such as the transformation between coordinates (such as UPS, UTM, and MGRS) to city names, or vice versa.
Metadata creation	The creation of descriptive metadata for the dataset.
Metadata formatting	The change of the file structure based on metadata standards.

Table 1: Categories and their definitions of our classification scheme

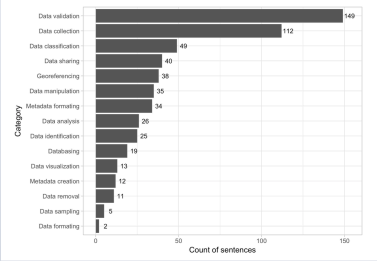

Figure 1: Frequencies of each category in our sample

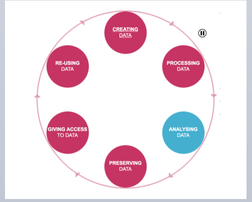

Figure 2: Data lifecycle of the UK Data Archive ("Research Data Lifecycle," n.d.)

Task	Creating	Processing	Analyzing	Preserving	Giving access to	Reusing
Data analysis			Y			
Data classification	Y					
Data collection	Y	Y				
Data formatting				Y		
Data identification		Y				
Data manipulation			Y			
Data removal		Y				
Data sampling		Y				
Data sharing					Y	
Data validation		Y				
Data visualization			Y			
Databasing	Y					
Georeferencing	Y					
Metadata creation	Y			Y		
Metadata formatting				Y		

Table 2: Mapping between our classification and UK Data Archive Lifecycle

Conclusion and Future Work

It is not surprising that all the data events identified in this study are concentrated on the first five of the six steps of the UK Data Archive Lifecycle. The missing of tasks related to data reuse is consistent with the nature of the data paper, which is to describe how datasets come into being, rather than how they are reused by other studies. This finding also proves the validity of the UK Data Archive Lifecycle in terms of its coverage of data tasks that are included in the current practice of the data paper.

The ultimate goal of this study is to create a more empirical-driven ontology about how datasets are being collected, manipulated, and shared in scientific activities. Our preliminary findings raise questions to the linearity of tasks in the existing data lifecycles. Even though not fully explored, it is clearly that a lot of the data events that we found belong to different stages of the data lifecycle. Another evidence of this point is that a few dozens of sentences contain more than one data events; many if not most of these events belong to different stages of the lifecycle as well. These inconsistencies clearly call for future studies about the relationship between data events and a more empirical examination of their linearity.

The work presented in this poster is preliminary. Next steps of this project include a more extensive review of our coding and tracking the seriality of events based on the descriptions of data papers. The classification scheme, while having some secondary analysis, requires collective verification by professionals in the broad knowledge domain of ecology, and to be tested on a larger sample of data papers, or potentially scientific papers. We also seek to track how data events are described to happen in the timeline within data papers before we can reach stronger conclusions concerning the lifecycle of research datasets in the real contexts of this genre of publication.

Reference

- Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, *66*(9), 1747–1762.
- Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(15), 1.
- Knorr, K. D. (1981). The Manufacture of Knowledge An Essay on the Constructivist and Contextual Nature of Science. Retrieved from https://philpapers.org/rec/KNOTMO-2
- Research Data Lifecycle. (n.d.). Retrieved September 11, 2017, from http://www.data-archive.ac.uk/create-manage/life-cycle
- Spencer, S. (2012). *What is DDI.*

Sponsored in part through RDA Data Share fellowship grant from the Alfred P. Sloan Foundation, #G-2014-13746.

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com