Citation:
Qin, J. (2018). A relation typology in knowledge organization systems: Case studies in the research data management domain. In: *Proceedings of the Fifteenth International ISKO Conference, Porto, Portugal, July 9-11, 2018*.

Jian Qin

A Relation Typology in Knowledge Organization Systems: Case Studies in the Research Data Management Domain

Abstract

Relations between concepts (and/or entities, events, and other things) vary depending on the criteria by which relations are defined or viewed. In the domain of research data management, different types of research generate different types of data and terminologies vary between practitioners and basic science researchers even within the same disciplinary domain. Interactions between datasets, between datasets and documentation, and between datasets and computing code can result in different types of relations. This paper employs a framework of analysis to study concept relation types in the research data management domain. By using two cases – one is the GenBank annotation records and the other is the data and artifact collection from a gravitational wave search, this paper demonstrates the types of relations existing in and between datasets, publications, computing codes, and workflows. The analysis and generalization of these relations references the research in AI's knowledge representation and knowledge organization systems (KOS), including both ad hoc subject categories and formal KOS, because in the next AI era, relations as one of the key components of AI applications will be required to function not only as part of KOS for indexing data and publications, but more importantly, to function as codifiable knowledge for machine consumption.

Introduction

Research data management (RDM) is increasingly becoming part of routine work for many libraries and research institutions. Vocabularies for RDM have also emerged to meet the needs of data management and curation. Examples include a taxonomy of data-related terms for the geographical information systems (GIS) (Cole, 2005), the vocabulary for research data repository registration and description (Vierkant et al., 2012), and the DPCVocab framework by Chao et al. (2015). While these endeavors provide useful frameworks for understanding the description of data and curation practices, the complex and entangled relations between concepts in datasets and other research artifacts are still a less known territory waiting for exploration. The recent hype in artificial intelligence (AI)

research and development and the promises that knowledge organization systems (KOS) have toward AI call for a re-examination of relations in KOS.

This paper is an extension of this author's 2002 ISKO article on the evolving paradigm of knowledge representation and organization, in which the author presented two opposite but crossover spectra in knowledge representation and organization, one for pragmatism vs. epistemologism and the other for integration vs. disintegration (Qin, 2002). By using two cases – one is the GenBank annotation records and the other is the data and artifact collection from a gravitational wave search, this paper will demonstrate the types of relations existing in and between datasets, publications, computing codes, and workflows. The analysis and generalization of these relations will reference the research in AI's knowledge representation and KOS, including both *ad hoc* subject categories and formal KOS, because in the next AI era, relations as one of the key components of AI applications will be required to function not only as part of KOS for indexing data and publications, but more importantly, to function as codifiable knowledge for machine consumption.

Framework of Analysis

Concepts and relations between concepts are the backbones of knowledge organization system (KOS). The most common types of relations among terms in thesauri and subject heading lists are broader terms (BT), narrower terms (NT), and related terms (RT), which establish parent-child, part-whole, or associative relations between concepts. Such relations primarily represent the scope of concepts and can be limited in representing many other concept relations beyond concept scope. For example, the Unified Medical Language System (UMLS) defines about 60 relations in five broad categories: R1. physically_related_to, R2. spatically_related_to, R3. functionally_related_to, R4. termporallu_related_to, and R5. conceptually_related_to (UTS, 2017). Another example is the relationships defined by Schema.org for Dataset, which are largely inherited from the CreativeWork class. Major relationships for Dataset (or CreativeWork) include hasPart, isBasedOn, isPartOf, and mentions (Schema.org, 2017).

The content and formats of a KOS are largely determined by the purpose of KOS and the technology available at the time. Thesauri and subject heading lists fit into early computing technology, the purpose of which was to assist in indexing documents. Term-centric KOS worked well to meet the needs for indexing documents, but in the digital era where information and data are highly interconnected, the term-centric approach becomes too limited to fulfill the increasing demand for effective automatic representation of documents and data.

Ontologies as one of the KOS types entail much richer relations for the concepts they cover. In both the UMLS and Schema.org examples, the relations between concepts have

gone far beyond the traditional BT/NT/RT types. Relations between concepts (and/or entities, events, and other things) vary depending on the criteria by which relations are defined or viewed. In the domain of research data management, it is known that different types of research, e.g., experimental, observation, simulation, and survey, generate different types of data and terminologies vary between practitioners and basic science researchers even within the same disciplinary domain. Datasets often come with documentation (some in the form of user guide or manual) and computing code, and are associated with publications. Interactions between datasets, between datasets and documentation, and between datasets and computing code can result in different types of relations. Some of these fall into "research information" domain, a term used to describe everything associated with research process: outputs, grants and projects, equipment, researchers and their affiliations, activities, awards, impact, and so on (Bryant et al., 2017). The complexity of research data and information warrants a closer examination of concept relations in order to provide semantic and technical infrastructures for research data management.

This paper uses a framework (Figure 1) to analyze the relations in and between four major types of concepts or entities: data, code, publication, and researcher. The relations will be examined both from the intra-database and between-database perspectives as well as through the lens of association, provenance, discovery, and credit. Association is a generic term used to refer relations between two different concepts or entities, which may be as simple as linking by identifiers, or more complicated ones such as interactive, dependent, causal, or derivative. Provenance refers to relations that will allow researchers to verify and track data and code for quality control and/or reproducibility. Discovery and credit are straightforward and need no further explanation. Each of the four entities in Figure 1 may be examine from these four lens to accurately define the types of relations within and between the entities.

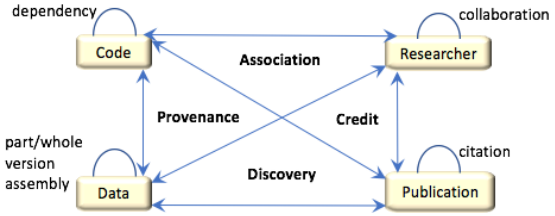


Figure 1. A framework for analyzing relations in research data management

Case Studies

Research data in different disciplinary fields vary greatly and it is likely that not all entities in the analysis framework (Figure 1) will be present in all disciplinary fields, for example, the code entity is not relevant in the repository for curating genome sequences. However, the

framework provides some direction for how to capture the essence of relations in research data management. Another point to make is that the discovery and credit dimensions of relations are straightforward, thus this paper will focus on association and provenance dimensions in the limited space for discussion. The following two cases are from two disciplinary fields remote to one another.

Case One: GenBank Annotations

GenBank is a data repository at the U.S. National Center for Biotechnology Information (NCBI, 2017) for curating genetic sequences. An annotation record in GenBank contains two parts: metadata and the sequence itself. The metadata section has three components: information about the sequence, publications associated with the sequence, and information about sequence data submitter.

As a genetic sequence data repository, GenBank's original (or raw) data are the base for many other databases to link and reference. We can call the relations between two data repositories "between-database" relations. For example, the BioSample database provides the source (biological materials) from which data stored in GenBank are derived, hence is the "source_of" gene sequences in GenBank, or reversely, sequences in GenBank are "derived_from" biomaterials in the BioSample database. ClinVar is another database at NCBI that collects user-contributed variants of known-genes (Figure 2). As such, sequences in GenBank are the "source_of" variants recorded in ClinVar, while ClinVar contains variants of the genes that are stored in GenBank.

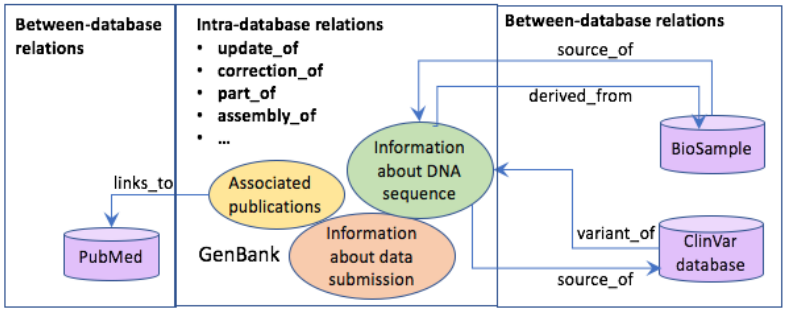


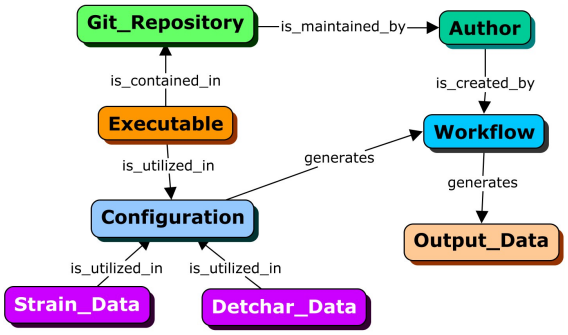
Figure 2. Intra-database and between-database relations in the example of GenBank

Besides between-database relations, GenBank records also have intra-database relations that are mainly between the records and between elements within the same record. It is common in GenBank that a sequence may be updated and resubmitted as a new version, a portion of a gene sequence with a large number of base pairs may be submitted as a separate record, or a record is an assembly of a collection of sequences submitted at different times by different submitters.

The relations at metadata element level requires an ontological approach to model how researchers, data, and publication are related. Relations at this level have been discussed in detail by Qin & Paling (2001) and Zeng & Qin (2016). Although the relations discussed in these publications are not intended for research data, the ways they discuss element level relations can be useful references.

Case Two: Gravitational Wave Research Information

Gravitational Wave (GW) research is at the scale of "big science" not only in the sense of disciplines and collaboration scales but also in the sense of financial investment. As a big science field, the complexity GW data creates many challenges for managing them. The first challenge is how to connect different types of data. The raw GW data are generated by laser interferometers as streams, which will then be calibrated and segmented before researchers analyze them. GW data may be categorized and labeled differently among the scientist depending on who are using the data at which stage of research. For example, recent detection data have been sliced by "Events" and "Bulk", and the input, output, code, documentation, and publications for each event are compiled as one dataset and given a DOI (LIGO, 2017). Figure 3 shows a simplified flow of data and information in a typical GW analysis project.



Note:
1. Git_Repository: scientists use GitHub to store their scripts or executables.
2. Configuration: a file in which everything needed: data, parameters, executables, and output location for running the analysis is listed.

Figure 3. Gravitational Wave research information overview

It is obvious that entities in Figure 3 are associated with one another in different ways. For example, within the *Executables* entity, it is common that the execution of one piece of code is dependent on certain software programs or a particular version of the software to work, hence code A is *dependent_on* code B would be the dependency relation (Figure 1) within the code entity, while *Executable* is maintained in GitHub and referenced in *Configuration*.

As GW research is highly data and computationally intensive, provenance relations play a significant role in making sure data, code, and workflows can be tracked all the way to the starting point, should any of them need to be verified for reuse or quality control purposes.

The relation types for provenance focus on who designed the configuration, wrote the code, and compiled the workflows, which data segments were selected from where, and where the output was sent. In a research environment where everything is done through writing programming code, such relation types will also require to be programmable for automatic capture during the whole computational process.

Discussion and Conclusion

The two (very brief) case studies offer us a number of lessons. First, there is not a single universally applicable relation typology for all disciplines when it comes to RDM. Each discipline has its own idiosyncrasies and needs domain analysis at several levels together with a thorough understanding of the domain as well as scientific and technical requirements for managing the data to enable discovery and use. Second, A relation typology is not a simple task to identify a list of labels or terms, but rather, it is a challenging task requiring a comprehensive set of knowledge and skills and a team of different talents. Relation identification is not only a process of identifying what concepts and entities there are, but also how they are related within the database, across databases, and throughout the whole research lifecycle. Finally, the ability to generalize relation types for a research domain marks how mature and effective the RDM practice is in that domain, because the existence of a relation vocabulary signals the depth of research and development in that domain. The relation vocabulary in UMLS is an excellent example for this point.

What should an RDM relation typology be exactly? The answer is that it depends. The framework proposed in this paper offers a starting point to identify and map out relation vocabularies for RDM, no matter for which domain you are defining and the relations. If we take association, provenance, discovery, and credit as the four dimensions of RDM relation typology, then the levels – cross-element, intra-database, cross-database – will be a way to cross-cut each dimension for getting a comprehensive picture of the relations in that domain. Some relation types can be prescribed for any domain, e.g., the classical relation types such as hierarchical, sibling, part-whole, but many more cannot be prescribed by a universally agreed list of relations in specialized disciplinary fields. The importance of relation typology will only increase as more disciplinary fields become data-intensive.

Relations have been an important area but not a popular topic of KOS research. This paper addresses the need for relation typology study and argues its importance in research data management in particular. The framework for relation type analysis provides some starting point for conducting domain analysis for identify relation types for research data management.

References

Bryant, R., Clements, A., Feltes, C., Groenewegen, D., Huggard, S., Mercer, H., Missingham, R., Oxnam, M., Rauh, A., & Wright, J. (2017). Research Information Management: Defining RIM and the Library's Role. Dublin, OH: OCLC Research. doi:10.25333/C3NK88

Chao, T. C., Cragin, M. H. and Palmer, C. L. (2015). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. Journal of Association for Information Science and Technology, 66: 616–633. doi:10.1002/asi.23184

Cole, F. (2005). The discourse of data: exploring data-related vocabularies in geographic information systems description. Journal of Information Science, 31(1): 44-56. DOI: 10.1177/0165551505049258

LIGO. (2017). LIGO Open Science Center. https://losc.ligo.org/start/. Accessed 31 January 2018.

NCBI. (2017). GenBank overview. https://www.ncbi.nlm.nih.gov/genbank/ Accessed 31 January 2018.

Qin, J. (2002). Evolving paradigms of knowledge representation: a comparative study of classification, XML/DTD, and ontology. In: Proceedings of the Seventh International Society for Knowledge Organization Conference, July 10-12, 2002, Granada, Spain, 465-471. Würzburg, Germany: Ergon.

Qin, J. & Paling, S. (2001). Converting a controlled vocabulary into an ontology: the case of GEM. Information Research: An International Electronic Journal, 6(2): http://InformationR.net/ir/6-2/paper94.html (January). Accessed 31 January 2018.

Schema.org. (2017). Dataset. Version 3.2. http://schema.org/Dataset

UTS. (2017). UMLS Terminology Services Metathesaurus Browser. Semantic Network 2017AA release. https://uts.nlm.nih.gov/metathesaurus.html

Vierkant, P., Spier, S., Rücknagel, J., Gundlach, J., Fichtmüller, D., Pampel, H., Kindling, M., Kirchhoff, A., Göbelbecker, H.-J., Klump, J., Bertelmann, R., Schirmbacher, P., Scholze, F. (2012). Vocabulary for the registration and description of research data repositories. http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:76875/component/escidoc:76874/re3data_vocabulary_v2-0.pdf

Zeng, M.L. & Qin, J. (2016). Metadata. 2nd ed. Chicago: ALA-Neal-Schuman.