Semantic Analysis and Attribute Clustering: Developing A Data Sharing Agreement Ontology

Jane Greenberg, Alice B. Kroeger Professor, Metadata Research Center, College of Computing and Informatics (MRC/CCI), Drexel University
Sam Grabus, Graduate Research Assistant, MRC/CCI, Drexel University
Hongwei Liu, Data Science Intern, MRC/CCI, Drexel University

1. Introduction: Data Sharing in Open and Closed Environments

Open data

Open science, open data, and data sharing stand as remarkable achievements of the early 21st century. Much of the success is tied to the development of a robust cyberinfrastructure comprised of user friendly technology, well-supported open data repositories, metadata standards, and suitable licenses (e.g. creative commons zero). Further, intellectual and policy motivators include federal and institutional data sharing policies, big data, and the growing capacity to support data-driven science, as promoted by Jim Gray's notion of the *Fourth Paradigm*[1]. All of these factors have also helped drive data sharing in communities where data is restricted due to privacy concerns, legal issues, or policy considerations; however, only to a degree.

Closed data

Sharing closed data has progressed, albeit and unsurprisingly, at a much slower pace compared to open data. Significant advances have been made in areas of anonymization and de-identifying personal information[2] to help address regulations and policies (e.g., HIPAA, institutional policies). Even so, the process of sharing data in a closed environment can be cost-prohibitive, and is often delayed or impeded due to legal costs, regulations, and simply fears of risks. Another delay factor is the absence a standard Knowledge Organization Systems (KOS) for addressing common closed data sharing challenges. We are investigating this challenge through the NSF Spoke research project, "A Licensing Model and Ecosystem for Data Sharing"[3] initiative, being pursued as part of the North East Big Data Innovation Hub (NEBDIH)[4].

2. Research Collaboration and Goals

The NSF Spoke research project, "A Licensing Model and Ecosystem for Data Sharing," is being pursued through a collaboration involving Drexel University's Metadata Research Center; Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory (MIT/CSAIL); and Brown University's Computer Science department. We are working toward an environment that will leverage existing de-identification and anonymization developments, and integrate with the DataHub platform[5]. Research goals to be reported on in this workshop will cover work to:

1) Develop a first-phase KOS for facilitating data sharing in environments where the data is not necessarily open and free.
2) Identify KOS classes, sub-classes, and attributes that can help support the automatic minting of component parts of data sharing agreements.

3. Methods and Results

Two studies were conducted based on a collection of 26 data sharing agreements. Study 1 was a keyword extraction and term proximity analysis. Word counts, common words (or phrases) appearing within a selected range of terms, such as "personally identifiable data," "personally identifiable information," "disclosure," "data restriction," and "privacy" were tallied. Study 2 was a content analysis of the licenses, which involved parsing concepts, confirming high-level classes, sub-classes, and attributes. The results include 6 high level classes, 15 mid-level sub-classes, and over 90 lower-level specific attributes, a portion of which can help to expeditiously support the automatic development of data sharing agreements.

4. Workshop Aims

This workshop presentation will provide background context, review our methods, and share results. We seek to engage the KOS community in a discussion to broader participation in this NSF effort, which links to the U.S. NSF Big Data Innovation Hub initiative and has links with global the Research Data Alliance.

Acknowledgements: We acknowledge the support of the National Science Foundation/IIS/BD Spokes/Award #1636788; and thank collaborators Sam Maddden (MIT), Carsten Bining (TU Darmstadt), and Tim Kraska (Brown University), and other individuals who provided us with data sharing agreements.

References

1. Hey, T., Tansley, S., Tolle, K. (2009) *The fourth paradigm: Data-intensive scientific research*. Microsoft Research. Redmond, WA.
2. El Emam, K., et al. (2012). De-identification methods for open health data: The case of the heritage health prize claims dataset. *Journal of Medical Internet Research, 14*(1), e33.
3. *A Licensing Model and Ecosystem for Data Sharing*: http://cci.drexel.edu/mrc/projects/a-licensing-model-and-ecosystem-for-data-sharing/.
4. Northeast Big Data Innovation Hub: http://nebigdatahub.org/.
5. Datahub (2016). *What is datahub?* https://datahub.csail.mit.edu/www/.