A Licensing Model and Ecosystem for Data Sharing

Sam Madden¹, Jane Greenberg² ³, Carsten Binnig⁴, Tim Kraska⁴, Danny Weitzner¹, & Sam Grabus² ³

¹ MIT, ² Metadata Research Center, ³ Drexel University, ⁴ Brown University

Summary

A part of the NSF Big Data regional innovation hub program, the Northeast hub, is addressing key data sharing challenges by:

- Creating a licensing model for data that facilitates sharing data that is not necessarily open or free between different organizations,
- Developing a prototype data sharing software platform, ShareDB, which will enforce the terms and restrictions of the developed licenses, and
- Developing and integrating relevant metadata that will accompany the datasets shared under the different licenses, making them easily searchable and interpretable.

To ensure that the developed tools and licenses are useful, the project will form the Northeast Data Sharing Group, comprised of many different stakeholders to make the licensing model widely accepted and usable in many application domains (e.g., health and finance).

Timeline

- Year 1: Requirements gathering. Initial requirements gathering workshop
- Year 2: Version 0.1 first draft version of the licensing model. Present model at workshop for suggestions. Identify stakeholders who commit to using the licensing model
- Year 3: Transition North East Data Sharing Consortium into a non-profit organization

Rationale

Sharing of data sets can provide tremendous mutual benefits for industry, researchers, and nonprofit organizations. A major obstacle is that data often comes with prohibitive restrictions on how it can be used (e.g., requiring the enforcement of legal terms or other policies, handling data privacy issues, etc.). Additionally, many attempts to share relevant data sets between different stakeholders in industry and academia fail or require a large investment to make data sharing possible.

Key Components

1. Data-sharing Licensing Framework/Generator
2. Data-Sharing Platform (enforce licenses)
3. Metadata (Search Licenses & Data)

Licensing Framework:
Creating a set of possible options that can be easily composed into a standardized data sharing agreement for different domains.

Data-Sharing Platform:
Develop a prototype software system for data sharing, which seamlessly enforces the restrictions stated in the developed licenses.

Metadata:
Develop a metadata scheme which leverages the best-of-breed from the vast amount of existing metadata standards.

References


GOAL: LICENSING FRAMEWORK

Standard terms that researchers, lawyers, and compliance teams conform with:
- Controlled access
- Tracking of access
- Usage rights (e.g., publication, copying)
- Duration of use
- Warranties of correctness/completeness/availability
- Other requirements and regulations

GOAL: HOSTED DATA-SHARING PLATFORM

ShareDB

A Licensing Model and Ecosystem for Data Sharing is supported by NSF award: 1636788

Current Progress @ Brown, Drexel, & MIT

- Gathering examples of data sharing licenses
- Parsing essential attributes
- Natural language processing, term clustering and categorization
- Initial data sharing platform (DataHub)
- Added support for access control & authorization
- Exploring research issues related to anonymization & de-identification of PII

A Licensing Model and Ecosystem for Data Sharing is supported by NSF award: 1636788

In collaboration with:

MIT CSAIL
Metadata Research Center
Brown Computer Science