Metadata Solutions and Data Sharing Licensing for Big Data

IEEE Workshop on Big Data Governance and Metadata and Management (BDGMM ’2018)

Jane Greenberg, Alice B Kroger Professor
Drexel University
Sam Grabus, Research Assistant

IIS/BD Spokes/Award #1636788
Overview

1. Questions...

2. Data sharing
   - Set the stage; closed/sensitive data

3. NSF Big Data Innovation Hub
   - “A Licensing Model and Ecosystem for Data Sharing”

4. Implications Big Data Governance and Metadata Management

5. Q&A, discussion
QUESTIONS?
Has anyone here deposited data or shared data for a hackathon?

• Open
• Restricted
• Don’t know…
  • Haven’t but thought about it…
Has anyone here shared research data with a colleague?

I did!!

*It helped me get tenure...*
Has anyone here ever thought...

- WOW, if only I could get that data of...[HEALTH RECORDS] [FOOD PURCHASE/INCOME] I could test that algorithm, conduct seriously robust research that has a real impact

- BUT... I cant because of...
  - Legal issues...
  - Privacy...
  - Policies
QUESTIONS
completed for now...

Data Sharing
Data sharing

• Setting the stage....
Data sharing motivations

1. Data deluge
2. Open science, open source
   • Jim Gray (Microsoft Research) notion of a *Fourth Paradigm*, toward data driven science
3. Local, federal and international policies and mandates
4. Opportunity to solve grand world challenges
How open data on agriculture & nutrition can solve world hunger

07 SEPTEMBER 2016
Give Up Your Data to Cure Disease

By DAVID B. AGUS  FEB. 6, 2016
## Data sharing barriers

<table>
<thead>
<tr>
<th>Policy</th>
<th>Licensing, agreements</th>
<th>Rights, privacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex regulations governing use of data in different domains</td>
<td>“Creative commons” (CC, CC0, etc.) does not address need</td>
<td>Concerns over sensitive information (e.g., PII)</td>
</tr>
<tr>
<td>Data lifecycle – data...living thing</td>
<td>Security</td>
<td></td>
</tr>
<tr>
<td>~ Do not want to loose control over data downstream</td>
<td>Technical and systematic aspects</td>
<td></td>
</tr>
<tr>
<td>~ What if data is redacted?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Why would someone go to all the effort to share their valuable data?</td>
<td></td>
</tr>
</tbody>
</table>
Still, merit in sharing

Involves lawyers to create individual agreement!

No sharing without a legal agreement
Open data

Closed data

Intel-Collaborative Cancer Cloud (CCC) (Dana-Farber, OHSU, Ontario Institute for Cancer Research (OICR)) - data

Collaborative Genomics Cloud (CGC) colocalizing massive genomics datasets) – genomics sharing, identifying cancer causing mutation

FICO score (Fair Isaac Corporation) – credit score, risk

Open data

Dryad

DataONE

DFC DataNet Federation Consortium

Closed data

Intel-Collaborative Cancer Cloud (CCC) (Dana-Farber, OHSU, Ontario Institute for Cancer Research (OICR)) - data

Collaborative Genomics Cloud (CGC) colocalizing massive genomics datasets) – genomics sharing, identifying cancer causing mutation

FICO score (Fair Isaac Corporation) – credit score, risk
Spokes and rings

Co-Chairs
Jane Greenberg, Drexel
Sam Madden, MIT
A Licensing Model and Ecosystem for Data Sharing

1. Licensing Framework / Generator

2. Data-Sharing Platform (Enforce Licenses)
   - DataHub

3. Metadata (Search Licenses and Data)
   - Principle: Solve the 80% case!
A Licensing Model and Ecosystem for Data Sharing

Project Summary

“A Licensing Model and Ecosystem for Data Sharing” is a spokes project led by researchers at Massachusetts Institute of Technology (MIT), Brown University as part of the Northeast Big Data Innovation Hub.

We are addressing data sharing challenges that are too frequently held up due legal matters, policies, privacy concerns, and other challenges that interfere with agreements. Sharing of data sets can provide tremendous mutual benefits for industry, researchers, and nonprofit organizations. A major obstacle is that data often comes with restrictions on how it can be used. Beyond open data protocols, many attempts to share relevant data sets between different stakeholders in industry have largely failed due to a lack of agreement.

We are addressing these challenges by: 1) Creating a licensing model for data that facilitates sharing data that is not necessarily open or free between organizations, 2) Developing a prototype data sharing software platform, ShareDB that will enforce agreement terms and restrictions for the licenses developed, and 3) Collecting and publishing relevant metadata that will accompany the datasets shared under the different licenses, making them easily searchable and interpretable.

“A Licensing Model and Ecosystem for Data Sharing” is also linked with the Northeast Data Sharing Group, comprising of many different stakeholders from widely accepted and usable in many application domains (e.g., health and finance).
Enabling Seamless Data Sharing in Industry and Academia (Fall 2017)

Heard from the trenches…

- Collect agreements
- Build a trusted platform
- Good metadata!
A Licensing Model and Ecosystem for Data Sharing” (NSF Spoke)

• First-phase metadata infrastructure for sharing of restricted data

• System Prototyping
Licenses: First Results
(Sam Grabus, CCI/Drexel)

High-level Categories

- General: attributes relating to the project and the agreement itself
  - e.g., Description of the data, Definition of terms

- Privacy & Protection: the protection of sensitive information and security
  - e.g., Individual identifiers removed prior to transfer, Encryption

- Access: who and how contact may be made with the data
  - e.g., Who has access, Method of access (approved hardware or software)

- Responsibility: legal, financial, ownership, and rights management pertaining to the data
  - e.g., Indemnity clause, Establishment of data ownership

- Compliance: ensuring fulfilment of agreement terms
  - e.g., Third party compliance with contract, Background checks for personnel

- Data Handling: specifics of permissible interactions with the data
  - e.g., Publication of data, Conditions for Termination
<table>
<thead>
<tr>
<th>Privacy &amp; Protection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitive Information</strong></td>
</tr>
<tr>
<td><strong>Regulations</strong></td>
</tr>
<tr>
<td>- Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.)</td>
</tr>
<tr>
<td>- Compliance with federal/state/international data protection laws and regulations</td>
</tr>
<tr>
<td><strong>Privacy</strong></td>
</tr>
<tr>
<td>- Anonymization of data</td>
</tr>
<tr>
<td>- Confidentiality and safeguarding of PII/sensitive data</td>
</tr>
<tr>
<td>- Removal/nondisclosure of company/personnel identification in materials and publications</td>
</tr>
<tr>
<td>- No contact with data subjects</td>
</tr>
<tr>
<td><strong>Security</strong></td>
</tr>
<tr>
<td>- Sharing non-confidential data</td>
</tr>
<tr>
<td>- Password protection/authentication of files</td>
</tr>
<tr>
<td>- Encryption</td>
</tr>
</tbody>
</table>
NLTK – parsing terms

• Set maximum keywords length: 5
  List top 1/5 of all the keywords

Result:
Keyword: research studies involving human subjects,
  score: 20.4583333333
Keyword: district assigned student identification numbers,
  score: 18.8387650086
Keyword: includes personally identifiable student information,
  score: 17.6168132942
Keyword: district initiated data research projects,
  score: 14.8577044025
Keyword: support effective instructional practices,
  score: 13.0
Keyword: personally identifiable information shared,
  score: 11.3440860215
Keyword: disclose personally identifiable information,
  score: 11.1440860215
Keyword: policy initiatives focused,
  score: 9.0
Keyword: informing education policies,
  score: 9.0
Sample 32 agreements – now, 70+
**Sentence with highest scores:**

<table>
<thead>
<tr>
<th>privacy</th>
<th>protection</th>
<th>set</th>
</tr>
</thead>
<tbody>
<tr>
<td>applicable</td>
<td>privacy</td>
<td>law</td>
</tr>
<tr>
<td>privacy</td>
<td>rule</td>
<td>standard</td>
</tr>
<tr>
<td>definition</td>
<td>set</td>
<td>privacy</td>
</tr>
<tr>
<td>data</td>
<td>security</td>
<td>protection</td>
</tr>
</tbody>
</table>

**Frequency from the most to the least:**

- [ ]
- [ ]
- [ ]
- [ ]
- [ ]
- [ ]
- [ ]
- [ ]
- [ ]
- [ ]

- [ ]
Goal: Licensing Framework

Standard terms for researchers/data providers, lawyers, and compliance teams

- ✔ Controlled access
- ✗ Tracking of access
- ✔ Usage rights (e.g., publication, copying)
- ✗ Duration of use
- ✔ Warrantees of correctness/completeness/availability
- ✗ Other requirements
Is this possible: Technology ✧ Sharing Agreements

**Technical**
- Access control & rights management

**Expiration**
- Logging & auditing
- Provenance/Finger printing
- De-identification
- “Noising”
- Aggregation

**Agreement Clauses**
- Controlled access (who & where)
- Tracking of access
- Usage rights (e.g., publication, copying)

**Duration of use**
- Warrantees of correctness/completeness/availability
- Other requirements
Is this possible: Technology ⊗ Sharing Agreements

**Technical**
- Access control & rights management
- Expiration
- Logging & auditing
- **Provenance/Finger printing**
- De-identification
- “Noising”
- Aggregation

**Agreement Clauses**
- Controlled access (who & where)
- Tracking of access
- **Usage rights** (e.g., publication, copying)
- Duration of use
- Warrantees of correctness/completeness/availability
- Other requirements
Guide to using ShareDB: Privacy Profiles

To create a new Privacy Profile and specify controls over your data set select ‘Create New Privacy Profile’

To browse existing Privacy Profiles (made by you or other users) and add one to this data set select 'Add Existing Privacy Profile' and click on the desired Privacy Profile.

Add Privacy Profiles

Create or change data privacy specifications for your data sets.

Create New Privacy Profile

Add Existing Privacy Profile
Health Insurance Portability and Accountability Act

Guide to using ShareDB: Privacy Profiles

Select desired privacy and security settings for your dataset. Once the Profile is created, it can be applied to any dataset.

Create new Privacy Profile for: testdata

Privacy Profile Name:

Regulations

- [ ] HIPAA
- [ ] FERPA

Privacy

- [ ] PII Anonymized or Removed
- [ ] PII Anonymized
- [ ] PII Removed

Reidentification

- [ ] Use K-Anonymity

K-size: Bucket Size for K
<table>
<thead>
<tr>
<th>IDENTIFICATION</th>
<th>FIRST_NAME</th>
<th>LAST_NAME</th>
<th>ADDRESS</th>
<th>PHONE_NUM</th>
<th>GENDER</th>
<th>SPECIES</th>
<th>RANDOM_SURVEY_ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sam</td>
<td>Grabus</td>
<td>123 Sesame Street, Philadelphia, PA</td>
<td>867-5309</td>
<td>Female</td>
<td>Human</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Jane</td>
<td>Greenberg</td>
<td>3141 Chestnut St, Philadelphia, PA 19104</td>
<td>555-5555</td>
<td>Female</td>
<td>Human</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Kingman</td>
<td>Grabus</td>
<td>123 Sesame Street, Philadelphia, PA</td>
<td>867-5309</td>
<td>Male</td>
<td>Dog</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Ted</td>
<td>Wark</td>
<td>103 Fayette St, Conshohocken, PA</td>
<td>123-5555</td>
<td>Male</td>
<td>Human</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>Morgi</td>
<td>Wark</td>
<td>103 Fayette St, Conshohocken, PA</td>
<td>123-5555</td>
<td>Male</td>
<td>Dog</td>
<td>No</td>
</tr>
<tr>
<td>IDENTIFICATION</td>
<td>FIRST_NAME</td>
<td>LAST_NAME</td>
<td>ADDRESS</td>
<td>PHONE</td>
<td>GENDER</td>
<td>ANIMAL</td>
<td>OWNERSHIP</td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>-----------</td>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>-----------</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Kingman</td>
<td>Grabus</td>
<td>123 Sesame Street, Philadelphia, PA</td>
<td>867-5309</td>
<td>Male</td>
<td>Dog</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Ted</td>
<td>Wark</td>
<td>103 Fayette St, Conshohocken, PA</td>
<td>123-5555</td>
<td>Male</td>
<td>Human</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>Morgi</td>
<td>Wark</td>
<td>103 Fayette St, Conshohocken, PA</td>
<td>123-5555</td>
<td>Male</td>
<td>Dog</td>
<td>No</td>
</tr>
</tbody>
</table>
Apply Privacy Profile To Tables
Profile name: hipaa pii removed
params: None

Base Tables

testdata

License applied
Apply Profile

Preview 'testdata_privacy_profile_6' privacy settings
<table>
<thead>
<tr>
<th>gender</th>
<th>random_survey_answer</th>
<th>identification</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Yes</td>
<td>1</td>
<td>Human</td>
</tr>
<tr>
<td>Female</td>
<td>No</td>
<td>2</td>
<td>Human</td>
</tr>
<tr>
<td>Male</td>
<td>Yes</td>
<td>3</td>
<td>Dog</td>
</tr>
<tr>
<td>Male</td>
<td>Yes</td>
<td>4</td>
<td>Human</td>
</tr>
<tr>
<td>Male</td>
<td>No</td>
<td>5</td>
<td>Dog</td>
</tr>
</tbody>
</table>
Overview

1. Questions...

2. Data sharing
   • Set the stage; closed data

3. NSF Big Data Innovation Hub
   • “A Licensing Model and Ecosystem for Data Sharing”

4. Implications Big Data Governance and Metadata Management

5. Q&A, discussion
### How Standards Proliferate:

(See: AC chargers, character encodings, instant messaging, etc.)

<table>
<thead>
<tr>
<th>Situation:</th>
<th>Soon:</th>
</tr>
</thead>
<tbody>
<tr>
<td>There are 14 competing standards.</td>
<td>We need to develop one universal standard that covers everyone's use cases.</td>
</tr>
</tbody>
</table>

**Why reinvent the wheel when you don't have to?**
### Lay of the land: Agent, access/rights, + workflow

<table>
<thead>
<tr>
<th>REQUIREMENTS</th>
<th>EXAMPLE METADATA STANDARDS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DATA PUBLICATION, DOMAIN DISCOVERY</strong></td>
<td></td>
</tr>
<tr>
<td>Persistent Identifiers</td>
<td>Product (Schema.org), DOI (Digital Object Identifiers), Handle system, OAIS (Open Archival Information System)</td>
</tr>
<tr>
<td>Domain specific schemes</td>
<td>Schema.org, RDA metadata directory or other resources</td>
</tr>
<tr>
<td><strong>IDENTIFICATION/DESCRIPTION</strong></td>
<td></td>
</tr>
<tr>
<td>Personal Identifiable Information</td>
<td>Person (Schema.org) vCard (Virtual Business Card), VIAF (Virtual International Authority File), ORCID (Open Researcher and Contributor ID)</td>
</tr>
<tr>
<td>Organization profile</td>
<td>Organization (Schema.org), ORCID, NAF (Name Authority File), EAC (Encoded Archival Context) for Organizational Bodies</td>
</tr>
<tr>
<td>Attribution</td>
<td>Same as PII</td>
</tr>
<tr>
<td><strong>LICENSING AND USE</strong></td>
<td></td>
</tr>
<tr>
<td>Restriction on Use</td>
<td>Embargos and Leases (Project HYDRA), PCDM (Portland Common Data Model: Rights Extension), METS, PREMIS (Preservation Metadata Data Dictionary)</td>
</tr>
<tr>
<td>Training/user requirements</td>
<td>Technical metadata, operational (see ‘Technical Format’ and ‘Restriction on Use’)</td>
</tr>
<tr>
<td>Technical format</td>
<td>Accessibility (Schema.org), W3C MS Global Access for All (AfA) Information Model Data Element Specification, PREMIS</td>
</tr>
<tr>
<td>Privacy</td>
<td>EHR (Electronic Health Records)</td>
</tr>
<tr>
<td><strong>LIFE-CYCLE MANAGEMENT</strong></td>
<td></td>
</tr>
<tr>
<td>Workflow</td>
<td>Protocols found via scientific research, such as Taverna and Kepler will aid this work.</td>
</tr>
<tr>
<td>Provenance</td>
<td>PROV-Model (Provenance Model, W3C), PREMIS</td>
</tr>
<tr>
<td>Accountability/Authenticity</td>
<td>PREMIS</td>
</tr>
</tbody>
</table>
Just a few... existing metadata and rights standards

- Rights statements.org:  

- Mets:  
  [http://www.loc.gov/standards/rights/METSRights.xsd](http://www.loc.gov/standards/rights/METSRights.xsd)  
  (rights declaration extension schema)

- Open Digital Rights Language (ODRL):  
  [https://www.w3.org/TR/odrl/](https://www.w3.org/TR/odrl/),  
  [https://www.w3.org/ns/odrl/2/](https://www.w3.org/ns/odrl/2/)

- ONIX-PL for licensing terms:  
Connecting with Initiatives

• Research Data Alliance
  • Legal interoperability Interest Group
  • RDA/NISO Privacy Task Group
  • RDA Metadata IG, WG (Metadata Standards Directory WG, Metadata Standards Catalog WG)

• Datasets licensing project: https://datasetlicencing.wordpress.com/

• Dataverse data tags project

• Linked Content Coalition—LCC Rights Reference Model as part of the LCC Framework: http://www.linkedcontentcoalition.org
FRAMEWORKS

https://www.force11.org/group/fairgroup/fairprinciples

• **FINDABLE:**
  
  F1. (meta)data are assigned a globally unique and eternally persistent identifier.
  F2. data are described with rich metadata.
  F3. (meta)data are registered or indexed in a searchable resource.
  F4. metadata specify the data identifier.

• **ACCESSIBLE:**
  
  A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
  A1.1 the protocol is open, free, and universally implementable.
  A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
  A2 metadata are accessible, even when the data are no longer available.

• **INTEROPERABLE:**
  
  I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  I2. (meta)data use vocabularies that follow FAIR principles.
  I3. (meta)data include qualified references to other (meta)data.

• **RE-USABLE:**
  
  R1. meta(data) have a plurality of accurate and relevant attributes.
  R1.1. (meta)data are released with a clear and accessible data usage license.
  R1.2. (meta)data are associated with their provenance.
  R1.3. (meta)data meet domain-relevant community standards.
On the metadata front - implications

- Never a one size fits all
- Do not want to reinvent the wheel, but seek to improve it
- Metadata longevity; data life-cycle mgmt.
  - Metadata governance hand-in-hand with BDGMM
- Machine readability for automating the life-cycle and processes
By agreeing and submitting this license, you (the author(s) or copyright owner) grant to Drexel University Libraries the non-exclusive right to reproduce, translate (as defined below), and/or distribute your submission (including the abstract) in print and electronic format and in any medium.
Conclusions and next steps

- Work underway, a lot of heavy lifting...
  - Mining licenses shows great diversity, but similarities
  - Metadata expertise
- Infrastructure to build on assisted with prototyping
- Continue to collect licenses
- Community building and connecting, IEEE-BDGMM, RDA – Research Data Alliance
Team members

- Sam Madden, Lead PI, Massachusetts Institute of Technology
- Carsten Binnig, PI, Brown University
- Sam Grabus, grad. RA, Drexel University
- Jane Greenberg, PI, Drexel University
- Hongwei Lu, grad. RA, Drexel University
- Famien Koko, grad. RA, MIT
- Tim Kraska, PI, Brown University
- Danny Weitzner, PI, MIT