

A Licensing Model and Ecosystem for Data Sharing



- Alex Bertsch, grad. RA, MIT, Brown University
- 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

IIS/BD Spokes/Award #1636788



DREXEL UNIVERSITY

Metadata
Research Center

College of Computing & Informatics



Overview

1. Motivation for “A Licensing Model and Ecosystem for Data Sharing” (SPOKE work)
2. Northeast hub/Drexel workshop, “Enabling Seamless Data Sharing in Industry and Academia” (<https://doi.org/10.17918/D8159V>)
3. Towards standardized licenses
4. Prototype software platform development
5. Conclusions and next steps

Data sharing advantages

Different Reasons

- More complete picture
- ROI
 - More data
 - More experts
 - Data reuse
- Better Insights into “Big Data”



Open data/open science

DataONE

DFC DataNet
FEDERATION
CONSORTIUM



DREXEL UNIVERSITY

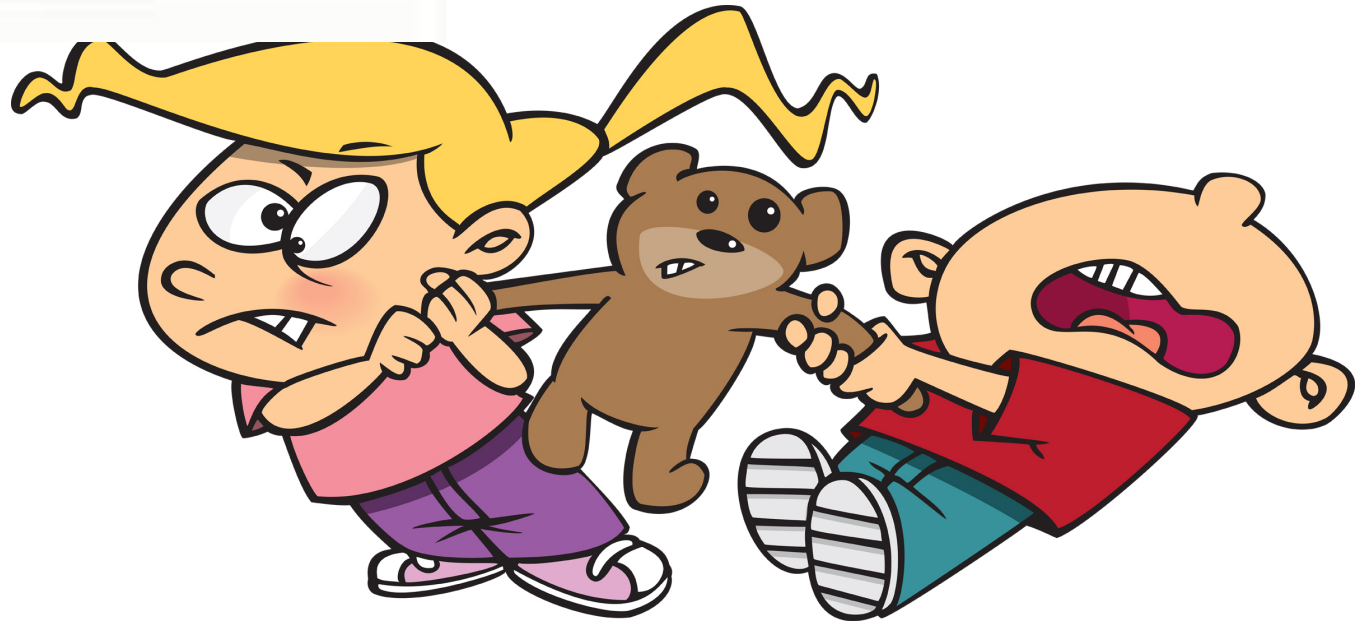
Metadata
Research Center

College of Computing & Informatics



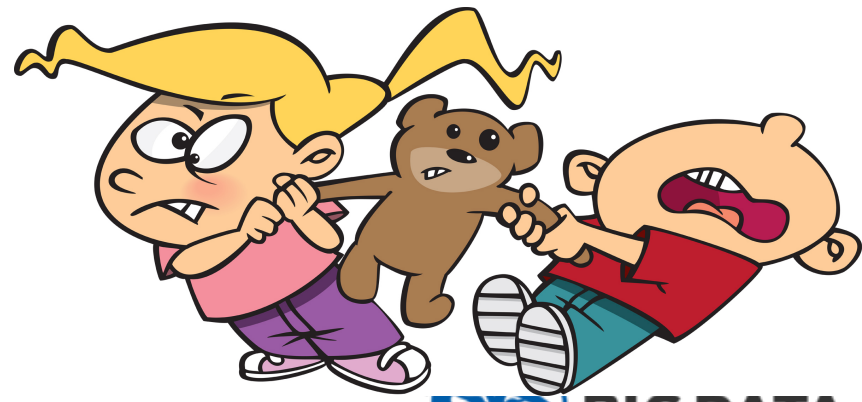
**NORTHEAST
BIG DATA**
INNOVATION HUB

Data sharing barriers





**Licensing, agreements,
rights, privacy, policy,
security, incentive...**



Significant barriers to data sharing,
particularly with **industry**...and other
partners

1. Licensing, agreements

- “Creative commons” does not address need

2. Rights, privacy

- Concerns over sensitive information (e.g., PII)

Significant barriers to data sharing

3. Policy

- Complex regulations governing use of data in different domains

Data lifecycle - living thing

- Do not want to loose control over data downstream
- Has to be updated
- What if data is redacted?

Significant barriers to data sharing

4. Security

- Technical and systematic aspects (~ policy, regulations, confidentiality/rights)

5. Incentives

- Why would someone go to all the effort to share their valuable data?

Still, merit in sharing





Sharing 'restricted' data today

- No sharing without a legal agreement



- Involve lawyers to create individual agreement!

- EDUCATION
- DATA SHARING
- ETHICS & POLICY
- PRIVACY & SECURITY



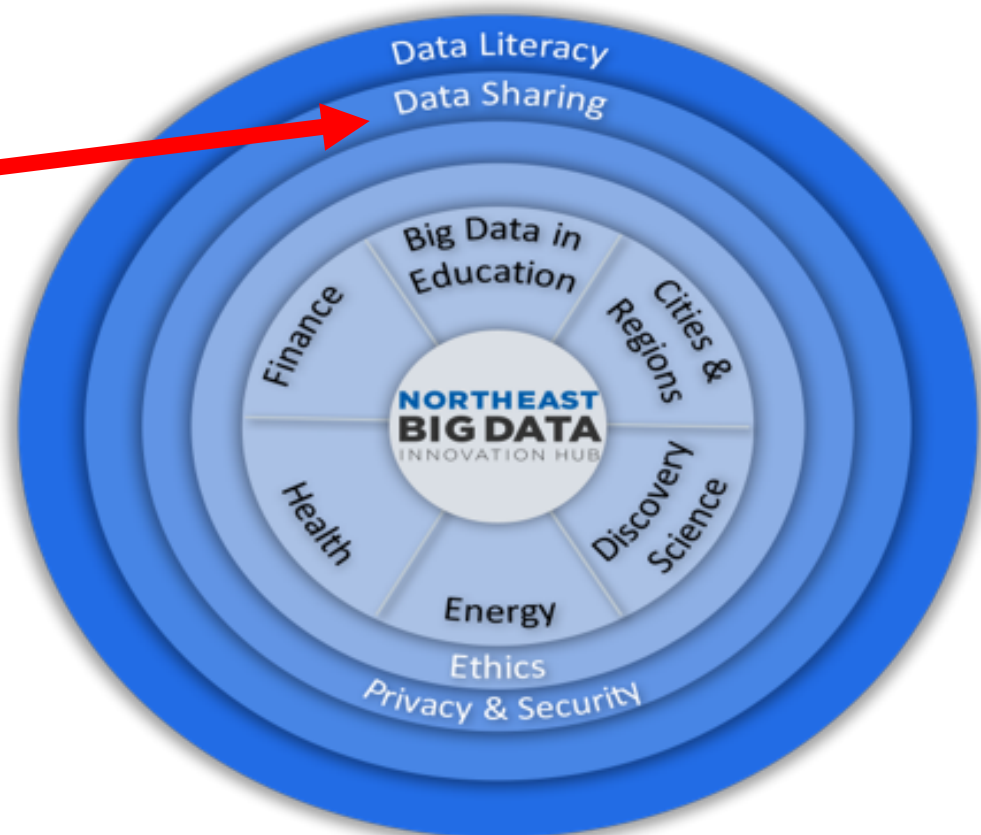
- Health
- Energy
- Cities & Regions
- Finance
- Big Data
Applications in Education
- Discovery
Science & Engineering

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)
 3. Metadata (Search Licenses & Data)
- Principle: Solve the 80% case!

<http://cci.drexel.edu/mrc/projects/a-licensing-model-and-ecosystem-for-data-sharing/>



DREXEL UNIVERSITY

Metadata
Research Center

College of Computing & Informatics

ABOUT

RESEARCH

PUBLICATIONS

PEOPLE

NEWS & EVENTS

SF

[CCI](#) / [Home](#) / [Research](#) /

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, and Drexel 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 inter agreement.

Sharing of data sets can provide tremendous mutual benefits for industry, researchers, and nonprofit organizations. A major obstacle is that data often has restrictions on how it can be used. Beyond open data protocols, many attempts to share relevant data sets between different stakeholders in industry require a large investment to make data sharing possible.

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 creators; 2) Developing a prototype data sharing software platform, ShareDB that will enforce agreement terms and restrictions for the licenses developed, and (3) Identifying 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 that have a 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...

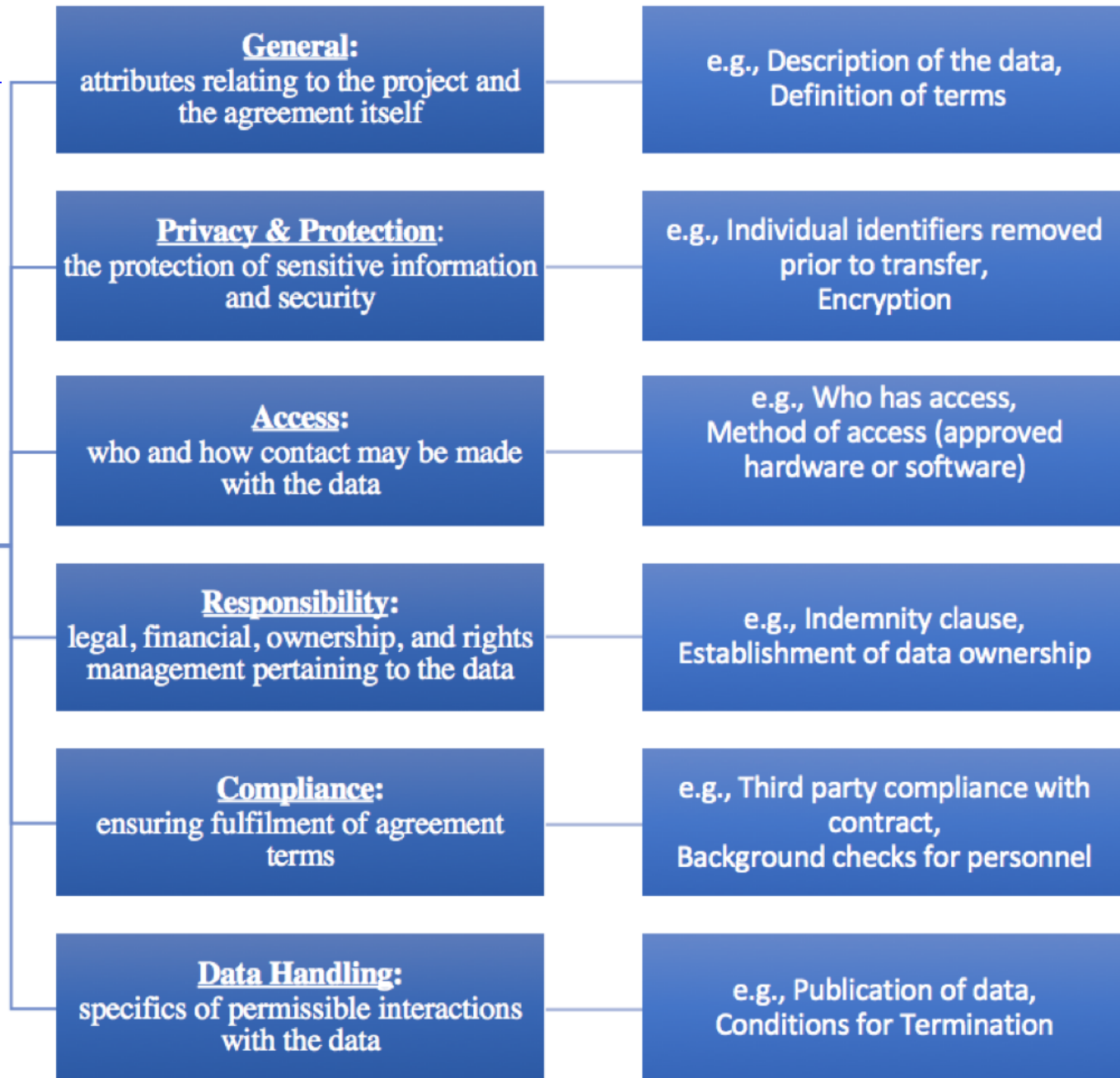
- Collect agreements
- Build a trusted platform
- Good metadata!



Licenses: First Results

(Sam Grabus:
smg383@drexel.edu)

High-level Categories



Privacy & Protection

Sensitive Information

<i>Regulations</i>	<i>Preparing data</i>	<i>Access</i>
<ul style="list-style-type: none"> • Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.) • Compliance with federal/state/international data protection laws and regulations 	<ul style="list-style-type: none"> • Identification of confidential/special categories of information (e.g., <u>pii</u>, proprietary) • Individual identifiers removed/anonymized prior to transfer 	<ul style="list-style-type: none"> • Who has access to <u>pii</u>/confidential data • Who has access to proprietary information
<i>Privacy</i>	<i>Avoiding re-identification</i>	<i>Exceptions</i>
<ul style="list-style-type: none"> • Anonymization of data • Confidentiality and safeguarding of PII/sensitive data • Removal/nondisclosure of company/personnel identification in materials and publications • No contact with data subjects 	<ul style="list-style-type: none"> • No direct/indirect re-identification • Statistical cell size (how many people, in aggregated form, can be released in groups) • Merging data with other sets (e.g., allowed with aggregated data—not in any way that will re-identify) 	<ul style="list-style-type: none"> • Exceptions to confidentiality • Conditions of proprietary information disclosure • Conditions of <u>pii</u> disclosure (who, what, and for what purpose?) • Limitations on obligations if data becomes public • Limitations on obligations if data is already known prior to agreement • Limitations on obligations if data given by 3rd party without restriction
<i>Security</i>		
<ul style="list-style-type: none"> • Sharing non-confidential data • Password protection/authentication of files • Encryption 	<ul style="list-style-type: none"> • Security training for involved personnel • Establishing infrastructure to safeguard confidential data 	

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

System brainstorming, building on DBHub

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
- Warrantees of correctness/completeness/availability
- Other requirements

Is this possible: Technology \bowtie 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 \bowtie 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 \bowtie 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 \bowtie 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

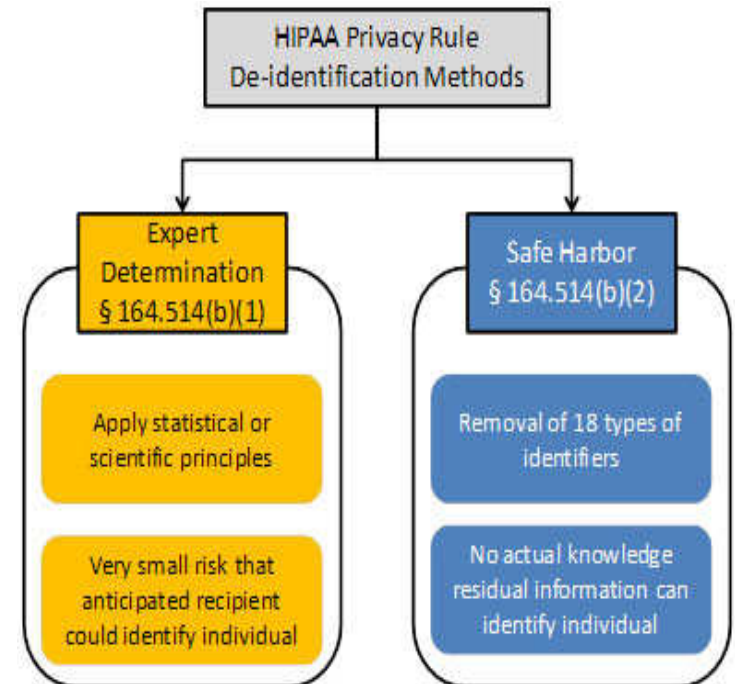
Other requirements

Platform: First Results

- De-identification is a major obstacle for data sharing (e.g., HIPAA, FERPA, ...)
- Interactive

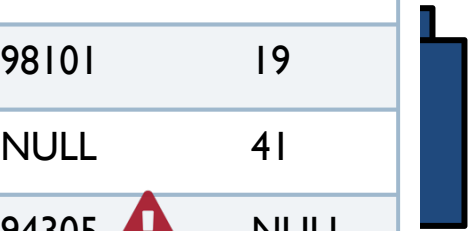
De-identification tool

- Detect sensitive columns (rule catalog, user-defined, machine learning, ...)
- Automatically de-identify

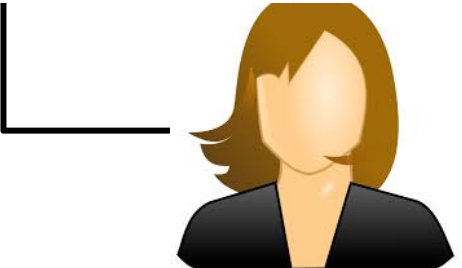


HIPAA: Interactive DE-identification

Id	Name	Street	City	State	P-Code	Age
1	J Smith	123 University Ave	Seattle	Washington	98106	42
2	Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
3	Bob Wilson	345 Broadway	Seattle	Washington	98101	19
4	M Jones	245 Third Street	Redmond	NULL	98052	299
5	Robert Wilson	345 Broadway St	Seattle	WA	98101	19
6	James Smith	123 Univ Ave	Seatle	WA	NULL	41
7	JWidom	123 University Ave	Palo Alto	CA	94305	NULL
...



a



data owner

[johndoe](#) / [test](#) /

Manage Repository Licenses

Add License

License Name	Applied To Tables	Manage Table Application
FERPA Data Anonymized	Applied to All Tables ✓	Manage
HIPAA PII Removed	Not Applied to All Tables ✗	Manage

[About](#)

[Documentation](#)

[GitHub Repo](#)

[API](#)

again

License not applied ✖

Apply To Table

changed

License not applied ✖

Apply To Table

Collaborators

✖ user1

✖ user2

Add Collaborators

Permissions for repo database tables:

- select
- update
- insert
- delete
- truncate
- references
- trigger

Permissions for repo files:

- read
- write

Add

Conclusions

- Work underway, a lot of heavy lifting...
- Infrastructure to build on
- Metadata expertise
- Mining licenses shows great diversity
- Community building and connecting

<https://cci.drexel.edu/ShareBigData>



Share Big Data

Introduction

The Northeast Hub Data Sharing Ring facilitates the exchange of solutions to advance (and benefit others). As a community, we seek to address key data sharing challenges relating to education about data sharing benefits.

- Successful agreements
- Share your case
- Links to licenses

April 2017, at 19:53.

[Privacy policy](#) [About ShareBigData](#) [Disclaimers](#)

[Home](#)
[People](#)

[Big Data](#)
[Sharing Big Data 101](#)
[Examples](#)
[Use cases](#)
[Licenses & Metadata](#)

[Tools](#)
[What links here](#)
[Related changes](#)
[Special pages](#)
[Printable version](#)

Final comment - Next Steps

- Data Sharing Spoke Workshop (Spring 2018)
 - Workshop agenda, slides:
<http://cci.drexel.edu/mrc/news/2016-11-bigdatahubworkshop/>
 - Final report: *Enabling Seamless Data Sharing in Industry and Academia* is at: <https://doi.org/10.17918/D8159V>.
- Collect more agreements and create license framework 0.1 (Grabus, Sam smg383@drexel.edu)
- Extend tool support, continue prototyping
- IRB/RDA connection, metadata check-list