Share BIG DATA

A Licensing Model and Ecosystem for Data Sharing

Board on Research Data and Information/US CODATA
International Coordination for Science Data Infrastructure
November 1, 2017

Jane Greenberg, Alice B Kroger Professor
Drexel University

IIS/BD Spokes/Award #1636788

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

NSF

NORTHEAST BIG DATA INNOVATION HUB

Share BIG DATA

Team members

- 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

NSF

MIT CSAIL

DREXEL UNIVERSITY
Metadata
Research Center
College of Computing & Informatics

BROWN
Computer Science

NORTHEAST
BIG DATA
INNOVATION HUB

Overview

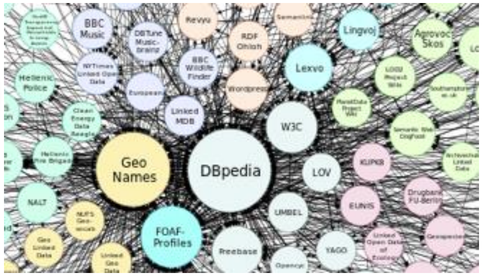
1. Data Sharing: Open Environments
   - Lots and lots of good resources
2. Closed Environments
   - “A Licensing Model and Ecosystem for Data Sharing” (NSF Spoke)
     - First-phase KOS for sharing of restricted data
     - Prototyping
3. Conclusions and next steps

DREXEL UNIVERSITY
Metadata
Research Center
College of Computing & Informatics

Open data

Yeah!

DRYAD



DataONE

DFC DataNet FEDERATION CONSORTIUM

Closed data

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

Collaborative Genomics Cloud (CGC)colocalizing massive genomics datasets)

FICO®

FICO score (Fair Isaac Corporation)

Data sharing barriers

Policy

- Complex regulations governing use of data in different domains
- Data lifecycle – data…living thing

  ~ Do not want to loose control over data downstream

  ~ What if data is redacted?

Licensing, agreements

“Creative commons” (CC) does not address need

Security

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

Rights, privacy

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

Incentives

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

Still, merit in sharing

No sharing without a legal agreement

Involves lawyers to create individual agreement!

A Licensing Model and Ecosystem for Data Sharing

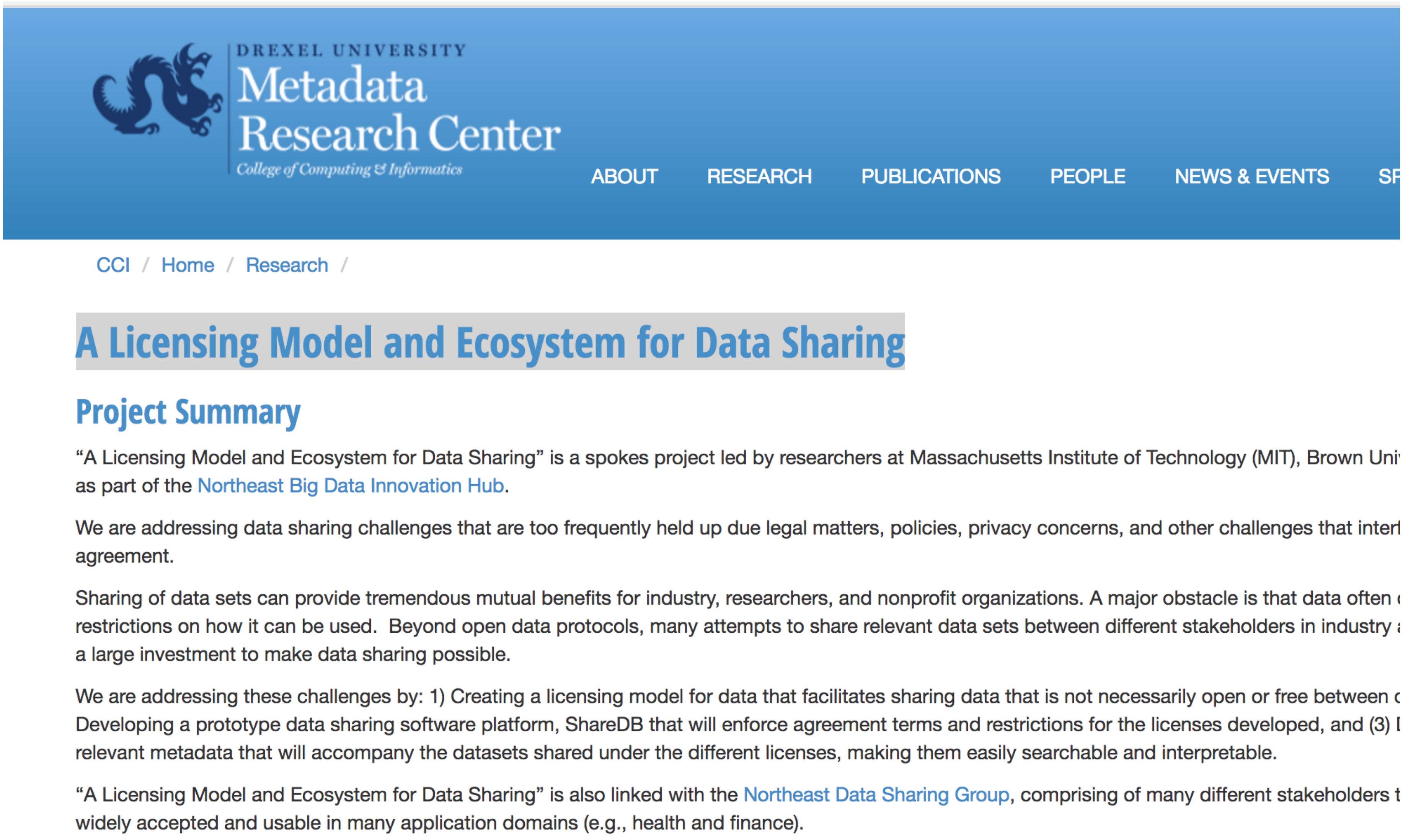
1. Licensing Framework / Generator

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

3. Metadata (Search Licenses and Data)

- Principle: Solve the 80% case!

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



CCI / Home / Research /

A Licensing Model and Ecosystem for Data Sharing

Project Summary

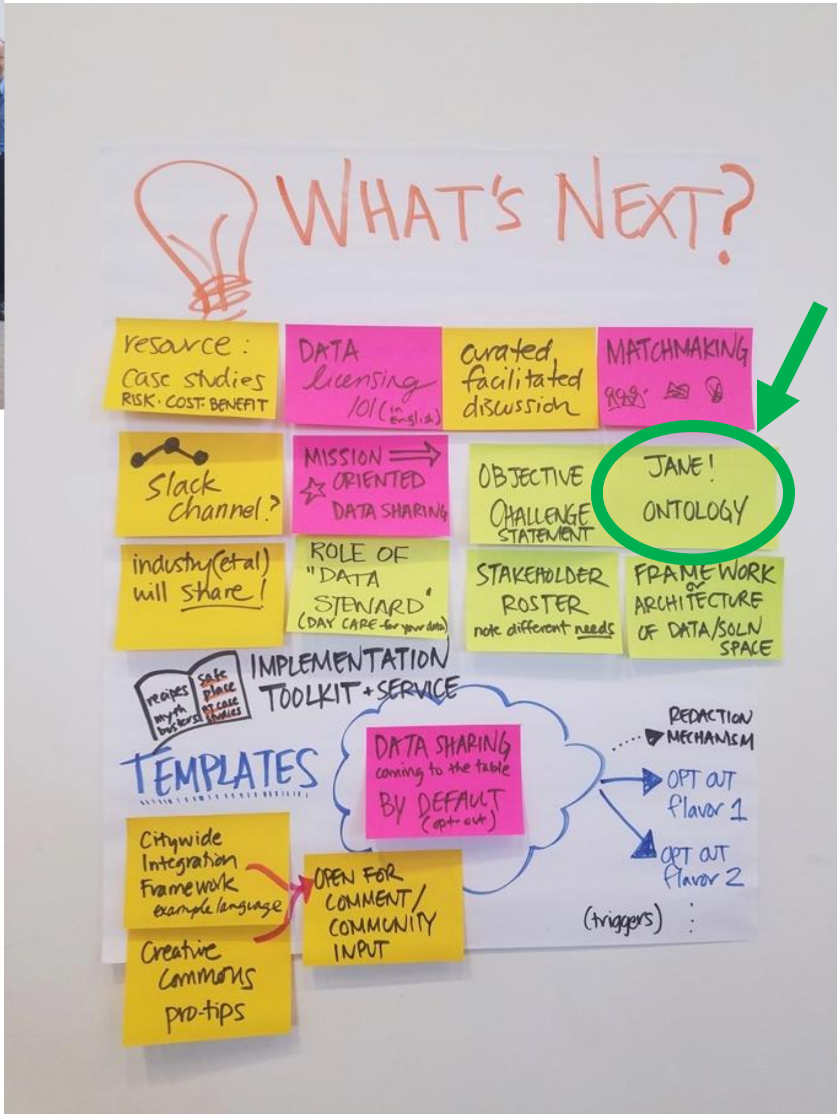
“A Licensing Model and Ecosystem for Data Sharing” is a spokes project led by researchers at Massachusetts Institute of Technology (MIT), Brown Uni
as part of the Northeast Big Data Innovation Hub.

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
restrictions on how it can be used. Beyond open data protocols, many attempts to share relevant data sets between different stakeholders in industry
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
Developing a prototype data sharing software platform, ShareDB that will enforce agreement terms and restrictions for the licenses developed, and (3)
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
widely accepted and usable in many application domains (e.g., health and finance).


WHAT'S NEXT?
resource : case studies RISK · COST · BENEFIT
DATA licensing 101 (in English)
curated, facilitated discussion
MATCHMAKING
Slack channel?
MISSION ⟹ ORIENTED DATA SHARING
OBJECTIVE CHALLENGE STATEMENT
JANE! ONTOLOGY
industry (et al) will share!
ROLE OF "DATA STEWARD" (DAY CARE for your data)
STAKEHOLDER ROSTER note different needs
FRAMEWORK or ARCHITECTURE OF DATA/SOLN SPACE
recipes myth busters safe place case studies
IMPLEMENTATION TOOLKIT + SERVICE
TEMPLATES
DATA SHARING coming to the table BY DEFAULT (opt-out)
REDACTION MECHANISM
OPT OUT flavor 1
OPT OUT flavor 2
(triggers)
Citywide Integration Framework example language
Creative Commons pro-tips
OPEN FOR COMMENT/ COMMUNITY INPUT

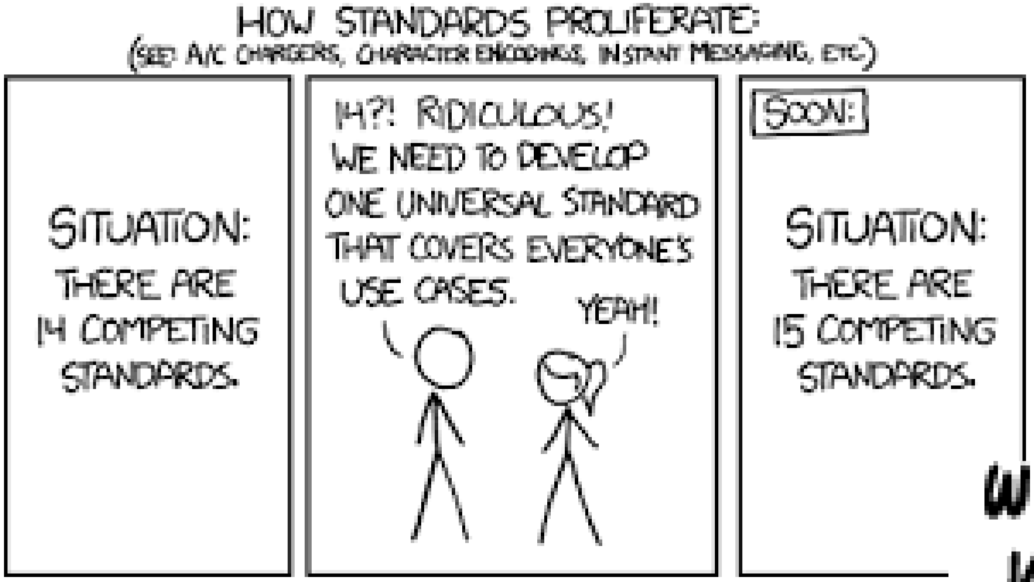
Enabling Seamless Data Sharing in Industry and Academia (Fall 2017)

Heard from the trenches…

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

A Licensing Model and Ecosystem for Data Sharing” (NSF Spoke)

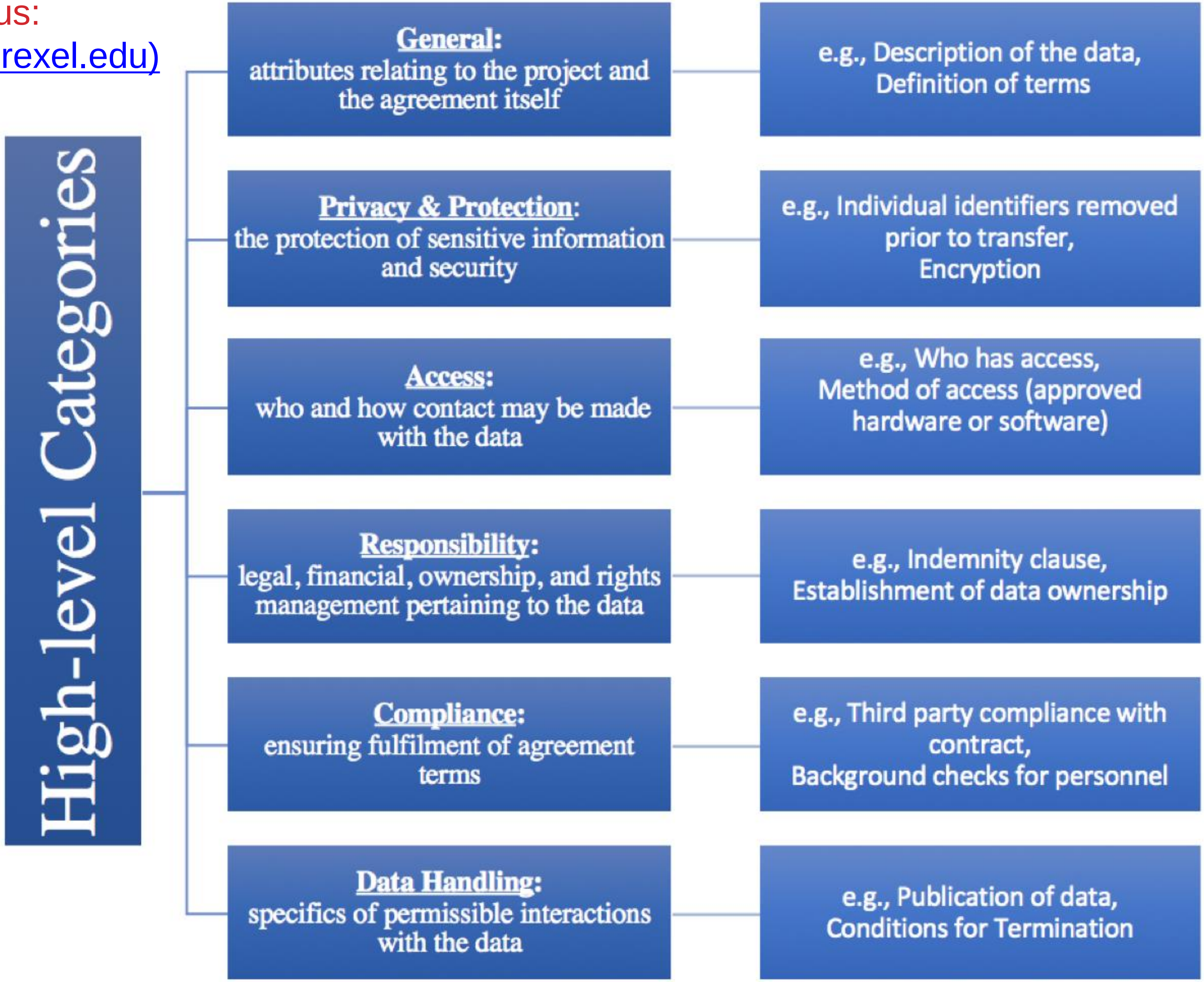
- First-phase KOS for sharing of restricted data
- Prototyping


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

WHY REINVENT THE WHEEL WHEN YOU DON'T HAVE TO?

Licenses: First Results

(Sam Grabus: smg383@drexel.edu)



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

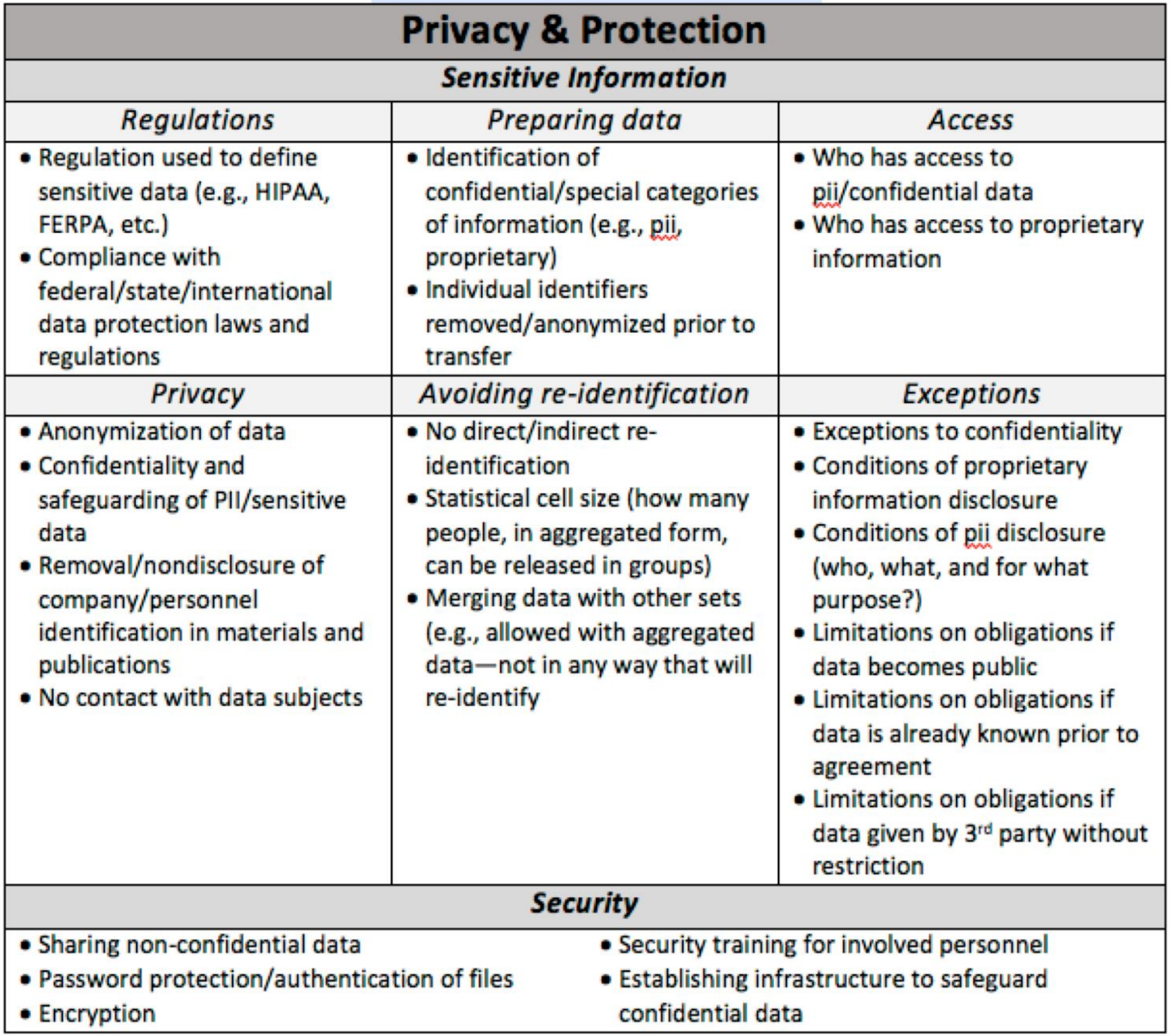
e.g., Indemnity clause, Establishment of data ownership

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



Privacy & Protection

Sensitive Information

Regulations	*Preparing data*	*Access*
• Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.) • Compliance with federal/state/international data protection laws and regulations	• Identification of confidential/special categories of information (e.g., pii, proprietary) • Individual identifiers removed/anonymized prior to transfer	• Who has access to pii/confidential data • Who has access to proprietary information
Privacy	***Avoiding re-identification***	***Exceptions***
• 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	• 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	• Exceptions to confidentiality • Conditions of proprietary information disclosure • Conditions of pii 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

Security

- Sharing non-confidential data
- Password protection/authentication of files
- Encryption
- Security training for involved personnel
- Establishing infrastructure to safeguard confidential data

Ontologizing

- **Privacy & Protection**
  - ❑ **Security**
    - ▪ Sharing non-confidential data ❼Sharing non-confidential data
    - ▪ Password protection/authentication of files ❼Password protection
    - ▪ Encryption ❼Encryption
    - ▪ Security training for involved personnel ❼Personnel Security Training
    - ▪ Establishing infrastructure to safeguard confidential data ❼Establishing Infrastructure
- **Data Handling**
  - ❑ **Use**
    - ▪ Each data field/elements to be accessed ❼Fields Accessed
    - ▪ Use of data: only for project-specific/research, or analytical use ❼ Research Use Only
    - ▪ Documenting all projects using the data ❼Projects involved
    - ▪ Modification of data ❼Modification
    - ▪ Compliance with data updates (e.g., changes, removal, corrections) ❼ Data Updates
  - ▪ Sharing data ❼Data Sharing

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

-5	-4	-3	-2	-1	0	1	2	3	4	5
			educational	right	privacy	act	health	insurance	portability	accountability
applicable	federal	law	regulation	protecting	privacy	citizen	including	family		
	license	agreement	authorized	protect	privacy	individual	subject	nd	study	
				applicable	privacy	law				
consistent	federal	family	educational	right	privacy	act	department	designates	education	alliance
subject	federal	family	educational	right	privacy	act	authorized			
education	record	covered	family	educational	privacy	act	amended			
recipient	agent	subcontractor	violation	agreement	privacy	rule	security	rule	implementing	regulation
comply	applicable	state	local	security	privacy	law	extent	protective	individual	privacy
		data	security	protection	privacy					
information	identified	family	educational	right	privacy	act				
		de	identified	applicable	privacy	law				
				applicable	privacy	law	permit	data	provider	provide
				federal	privacy	act	requirement	apply	agreement	entered
shared	state	subjected	applicable	requirement	privacy	confidentiality				
resolved	permit	covered	entity	comply	privacy	rule				
time	covered	entity	comply	requirement	privacy	rule	hipaa			
		reference	agreement	section	privacy	rule	mean	section	amended	renumbered
					privacy	rule	extent	information	created	received
					privacy	rule	standard	privacy	individually	identifiable
					privacy	rule	include	person	qualifies	personal
tern	defined	agreement	meaning	term	privacy	rule				
set	accordance	term	agreement	hipaa	privacy	security	rule			
hipaa	regulation	promulgated	thereunder	governing	privacy	security	health	information		

Sentence with highest scores:

privacy	protection	set			
applicable	privacy	law			
privacy	rule	standard	privacy	individually	identifiable
definition	set	privacy	rule		
data	security	protection	privacy		

Frequency from the most to the least:

🟦
🟦
🟪
🟨
🟧
🟩
🟥
🟦
⬜

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/Fingerprinting

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**)

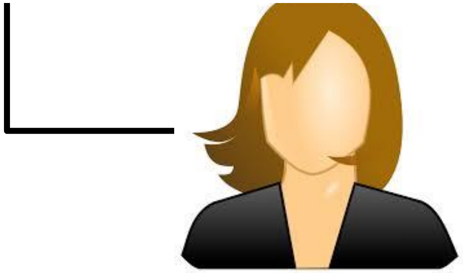
Duration of use

Warrantees of correctness/completeness/availability

Other requirements

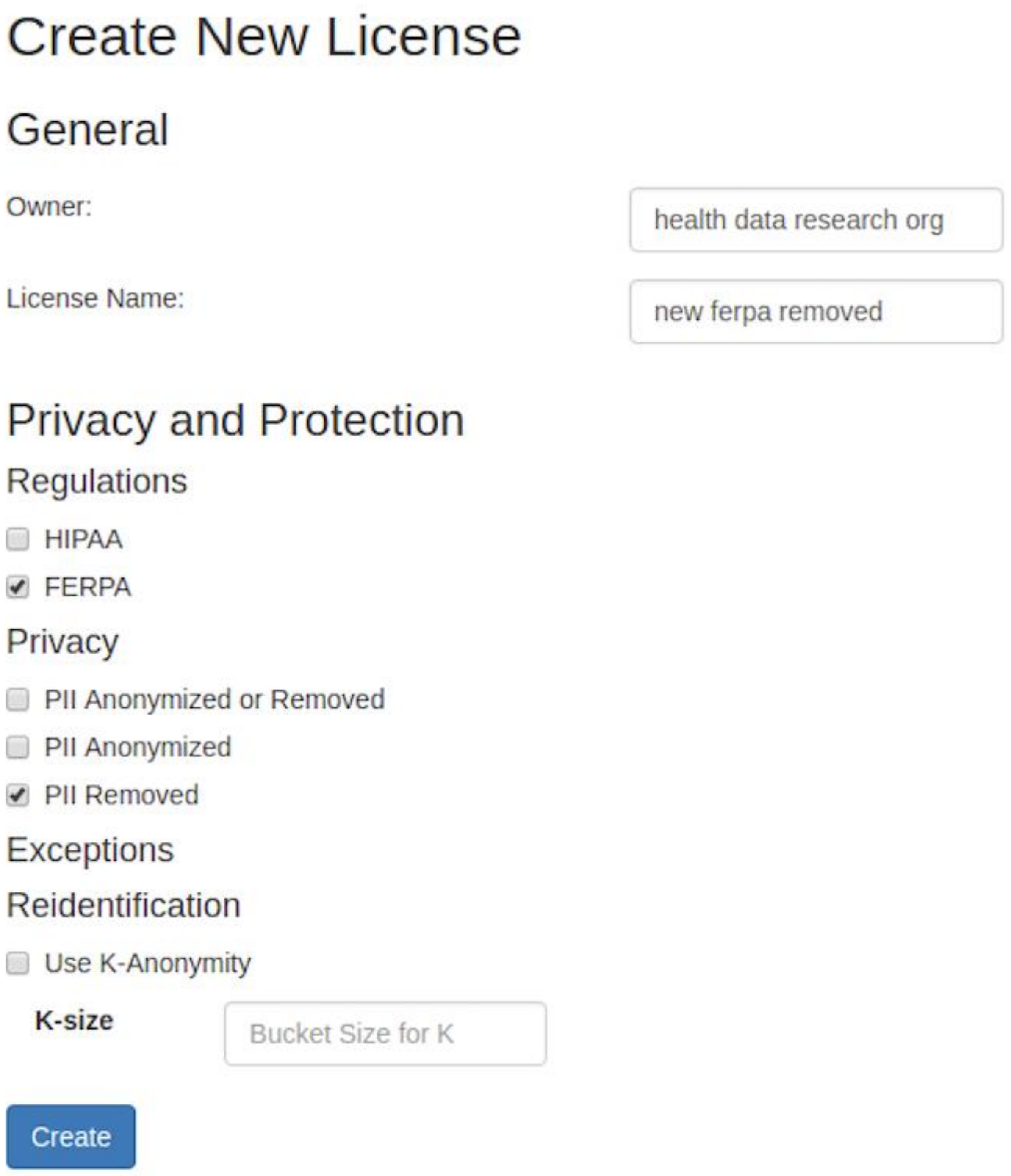
HIPAA: Interactive DE-identification

Id ✓	ame ⚠	Street ⚠	City ⚠	State ✓	P-Cod ?	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	J Widom	123 University Ave	Palo Alto	CA	94305 ⚠	NULL
…	…	…	…	…	…	…



data owner

DataHub



Create New License

General

Owner: health data research org

License Name: new ferpa removed

Privacy and Protection

Regulations

☐ HIPAA

☑ FERPA

Privacy

☐ PII Anonymized or Removed

☐ PII Anonymized

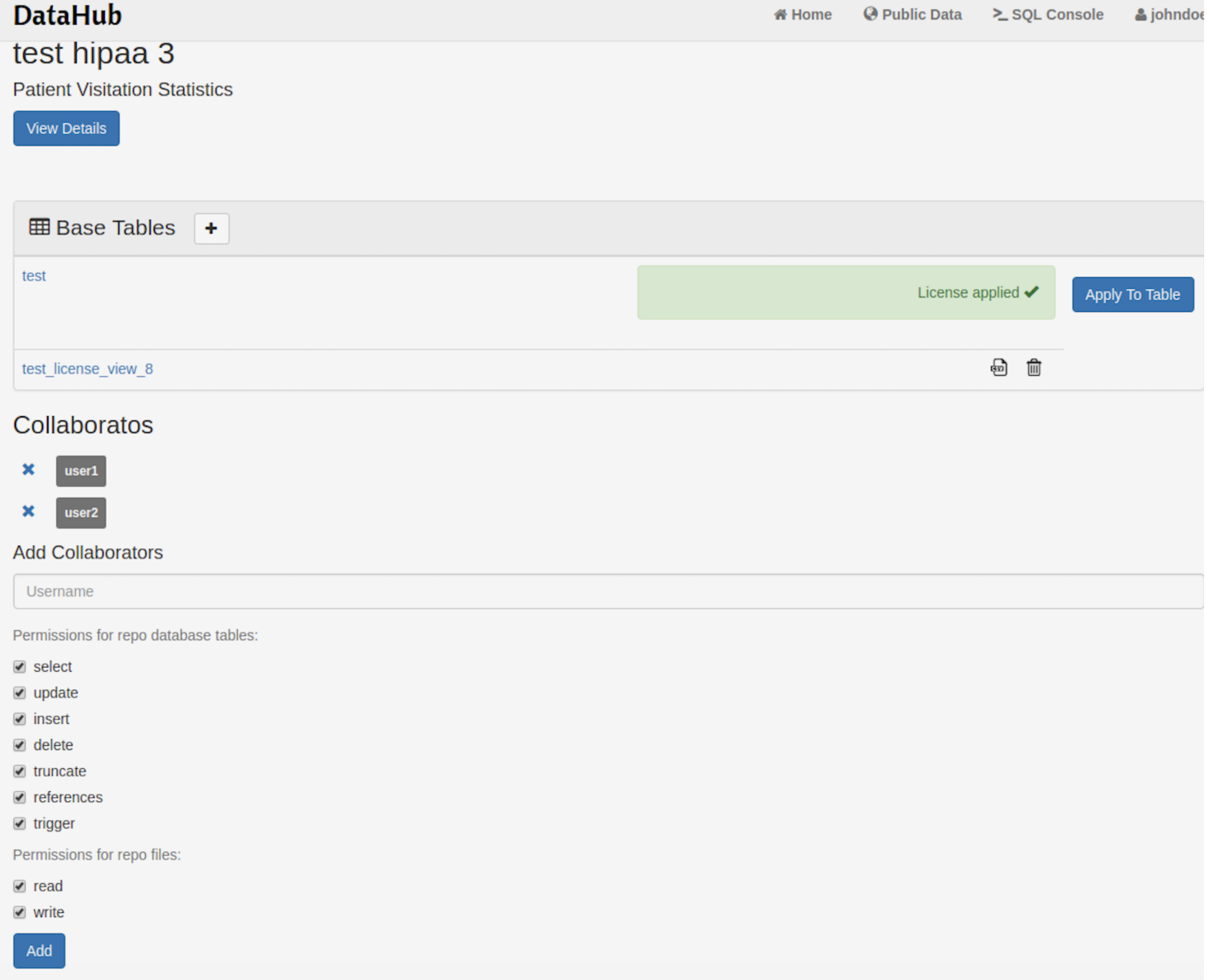
☑ PII Removed

Exceptions

Reidentification

☐ Use K-Anonymity

K-size Bucket Size for K



test hipaa 3

Patient Visitation Statistics

Base Tables

test — License applied ✓

test_license_view_8

Collaboratos

✖ user1

✖ user2

Add Collaborators

Username

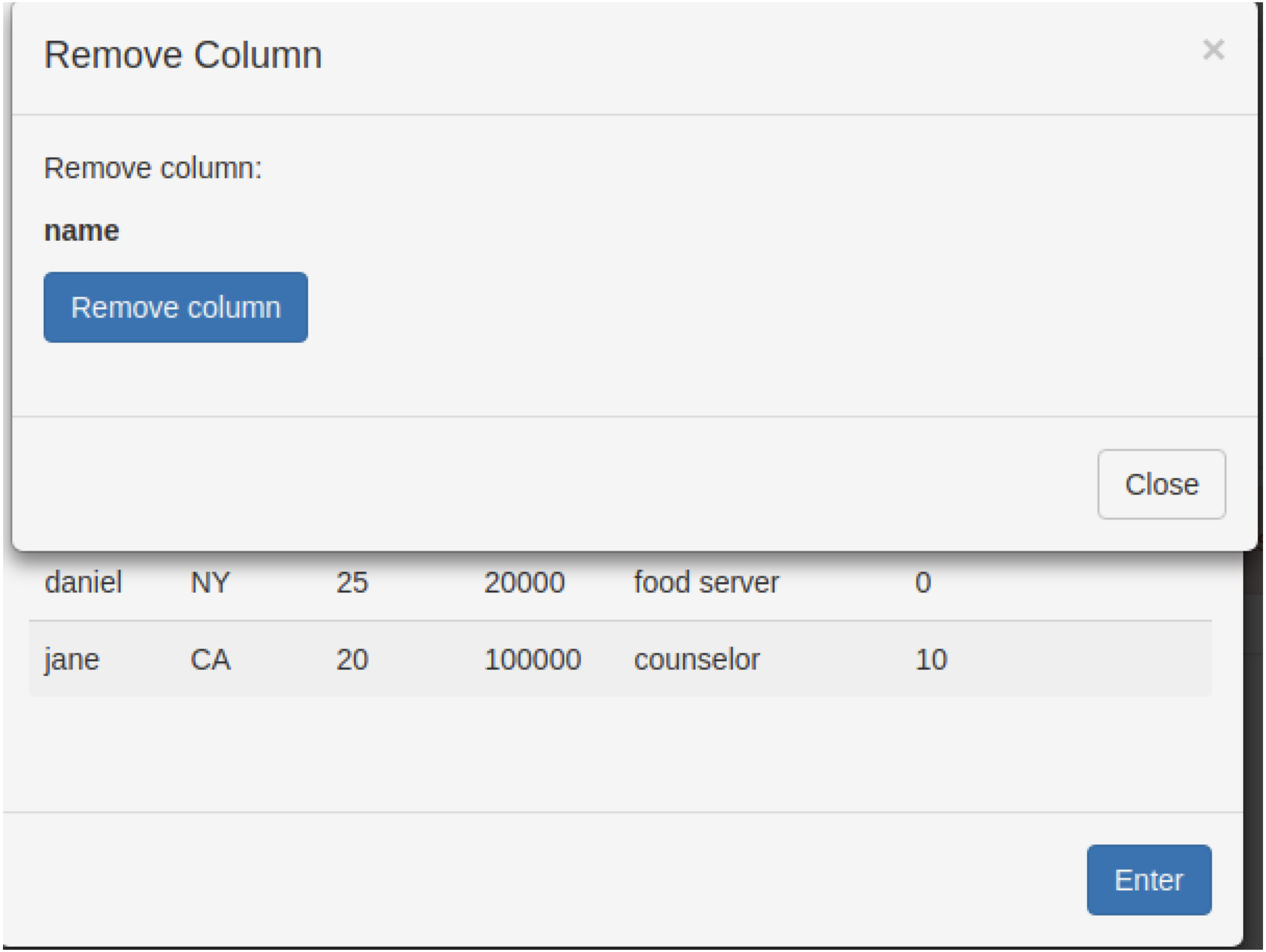
Permissions for repo database tables:

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

Permissions for repo files:

- ☑ read
- ☑ write

DataHub

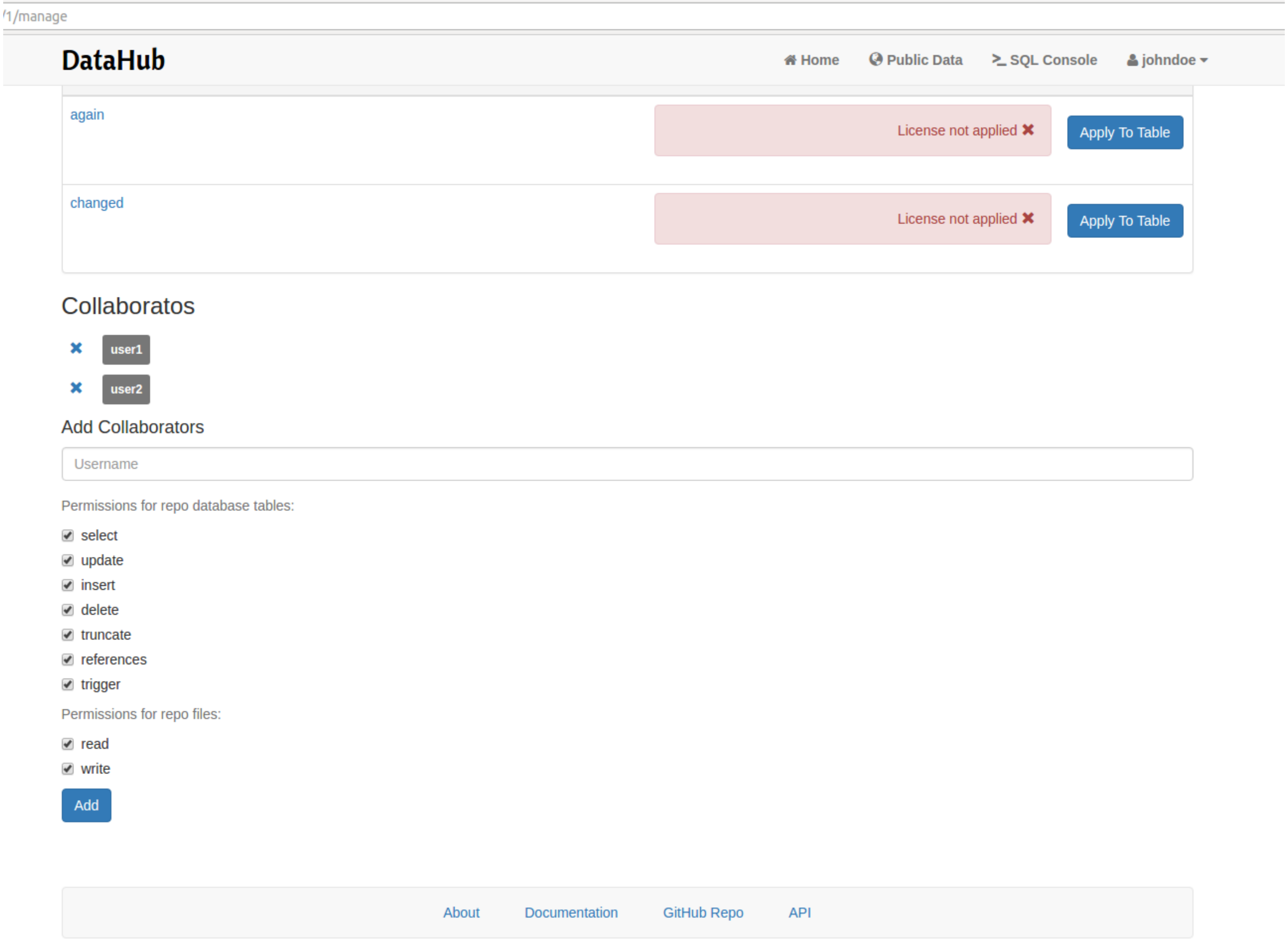


Remove Column

Remove column:

name

daniel	NY	25	20000	food server	0
jane	CA	20	100000	counselor	10



again — License not applied ✖

changed — License not applied ✖

Collaboratos

✖ user1

✖ user2

Add Collaborators

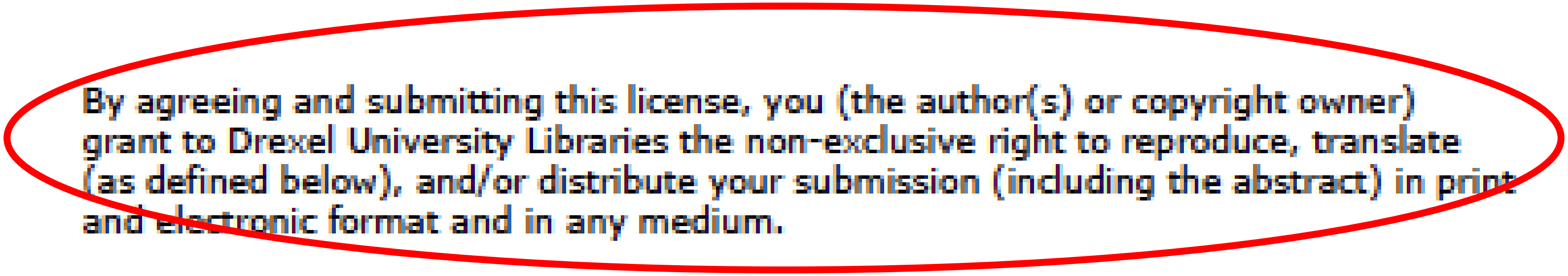
Username

Permissions for repo database tables:

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

Permissions for repo files:

- ☑ read
- ☑ write



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.

Jane Greenberg

Digitally signed by com.apple.idms.appleid.prd.55546a
DN: cn=com.apple.idms.appleid.prd.55546a4d526531
Date: 2017.04.06 17:39:38 +01'00'

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, RDA – Research Data Alliance