Share BIG DATA

Sharing Restricted Data: Challenges, Protocols and Implications for Digital Libraries

8th A-LIEP Conference/19th ICADL Conference
Chulalongkorn University, Bangkok Thailand
November 14, 2017

Jane Greenberg
Alice B Kroger Professor

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

IIS/BD Spokes
#1636788

NSF

NORTHEAST BIG DATA INNOVATION HUB

Overview

1. Questions?....
2. Data sharing
    - Set the stage
    - Closed data
3. Spoke
4. Implications for DL, field of ILS

QUESTIONS?

Who is here?

- *Library, archival, information/data scientists*
- *Computer scientists*
- *Researchers*
- *Educators*
- *All of the above*
- *Other?*

Has anyone here deposited research data?

- Open
- *Restricted*
- *Don't know…*
  - *Haven't but through about it…*

Has anyone here shared research data?

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 conduct research that has a real impact
- *BUT… I cant because of…*
  - ***Legal issues…***
  - ***Privacy…***
  - ***Policies***

Data sharing

- Set the stage….

Data sharing motivations

- Data deluge
- Open science, open source
- Jim Gray (Microsoft Research) notion of a *Fourth Paradigm*
  - supporting data driven science
- Opportunity to solve grand world challenges

How open data on agriculture & nutrition can solve world hunger

07 SEPTEMBER 2016

Noi Jaitang, interviewed as part of the World Resources Institute report, waters his garden in Thailand // Laura Villadiego

How to Solve the Environmental Information Divide

TERESA MATHEW SEP 5, 2017

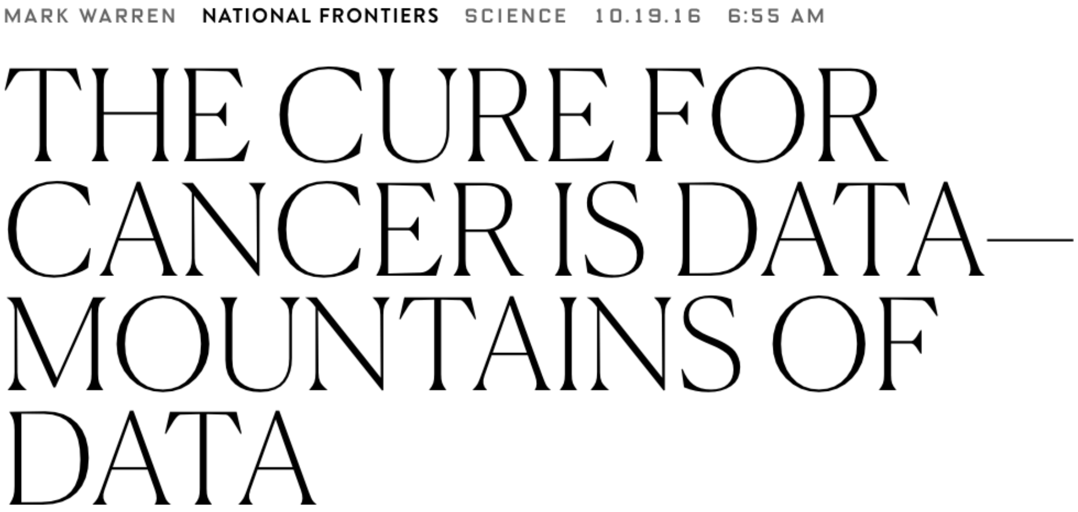
The New York Times

February 2016

SundayReview | OPINION

Give Up Your Data to Cure Disease

By DAVID B. AGUS FEB. 6, 2016

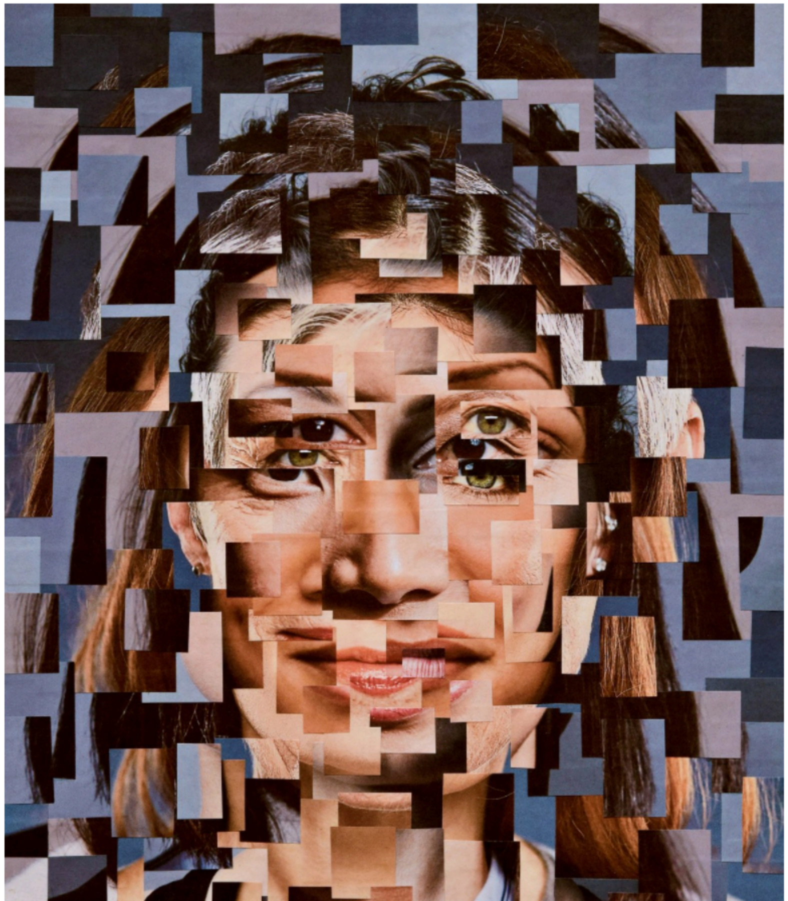


MARK WARREN NATIONAL FRONTIERS SCIENCE 10.19.16 6:55 AM

THE CURE FOR CANCER IS DATA—MOUNTAINS OF DATA

WIRED

October 2016

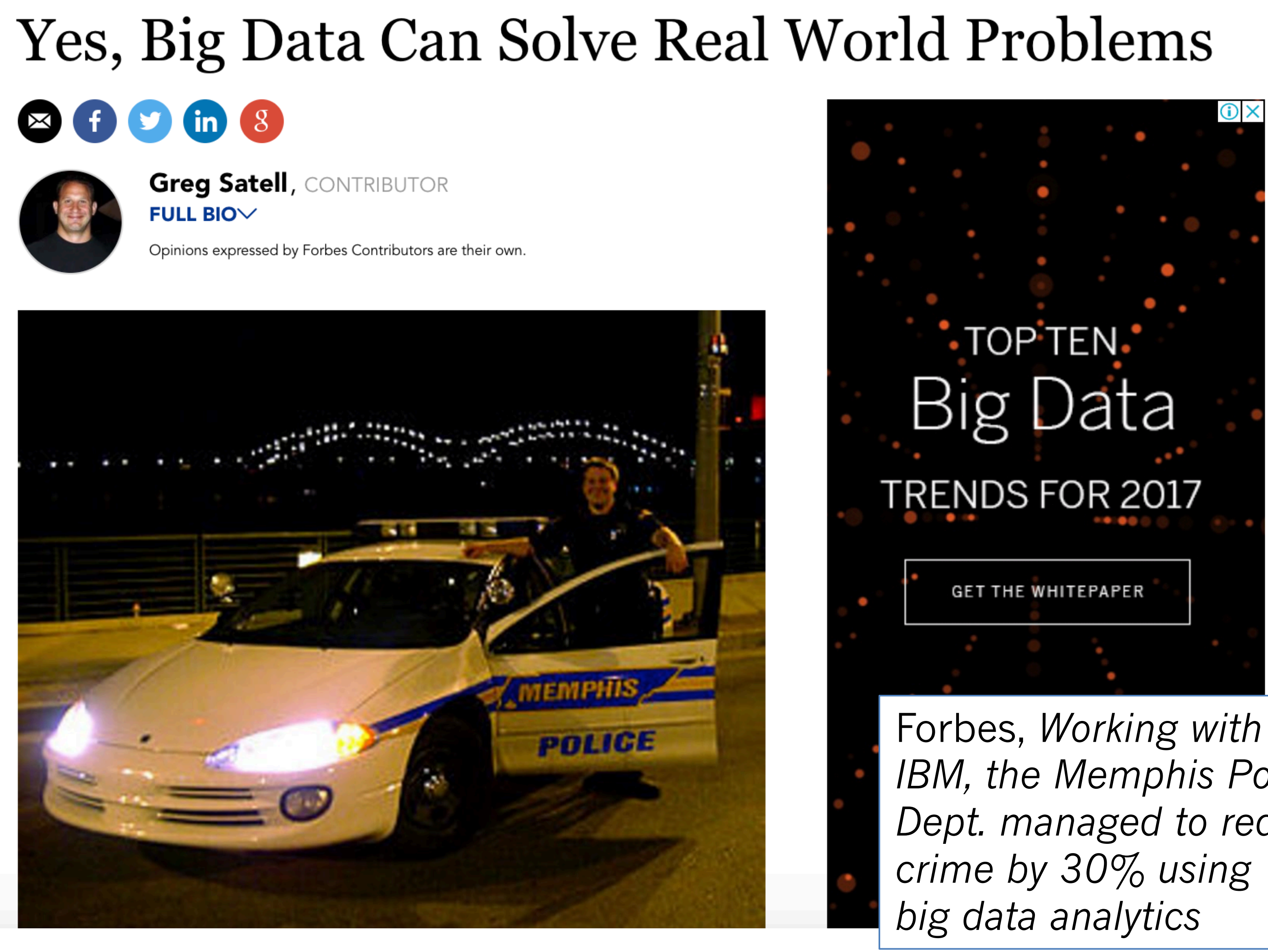


http://ipat.gatech.edu/news/students-use-data-internship-solve-real-world-problems

July 2017

Georgia Tech Data Science for Social Good & the Civic Data Science NSF REU

December 2013



Yes, Big Data Can Solve Real World Problems

Greg Satell, CONTRIBUTOR

FULL BIO

Opinions expressed by Forbes Contributors are their own.

Forbes, *Working with IBM, the Memphis Police Dept. managed to reduce crime by 30% using big data analytics*

Data sharing advantages

Different Reasons

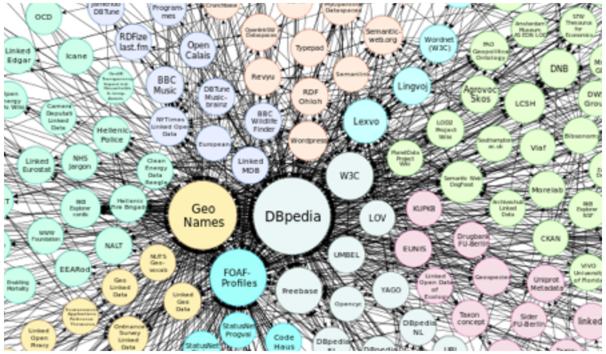
- More complete picture
- ROI
  - More data
  - More experts
  - Data reuse
- Better Insights into “Big Data”

Open data

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

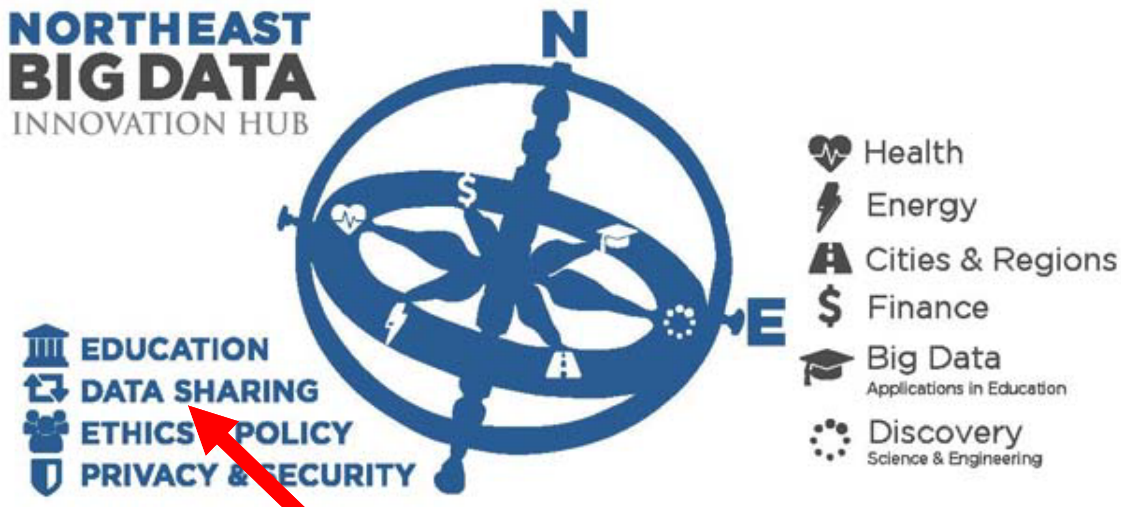
Why would someone go to all the effort to share their valuable data?

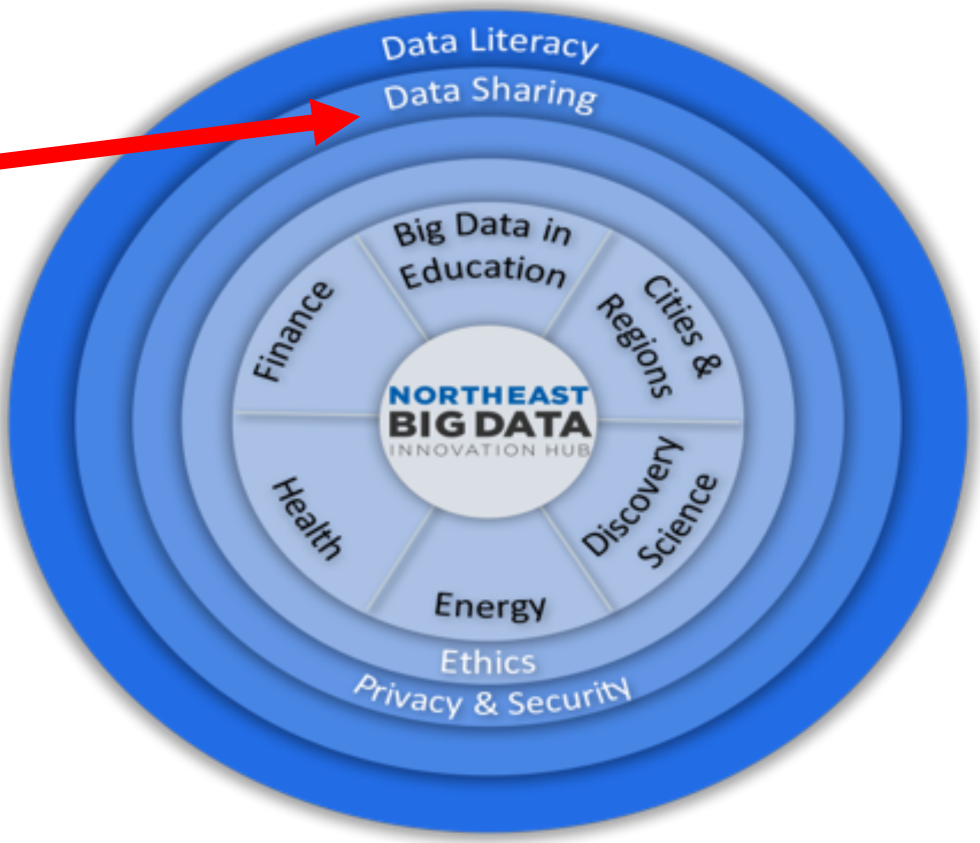
Still, merit in sharing

Sharing ‘restricted’ data today

- No sharing without a legal agreement

- Involve lawyers to create individual agreement!





Spokes and rings

Co-Chairs
Jane Greenberg, Drexel
Sam Madden, MIT

A Licensing Model and Ecosystem for Data Sharing

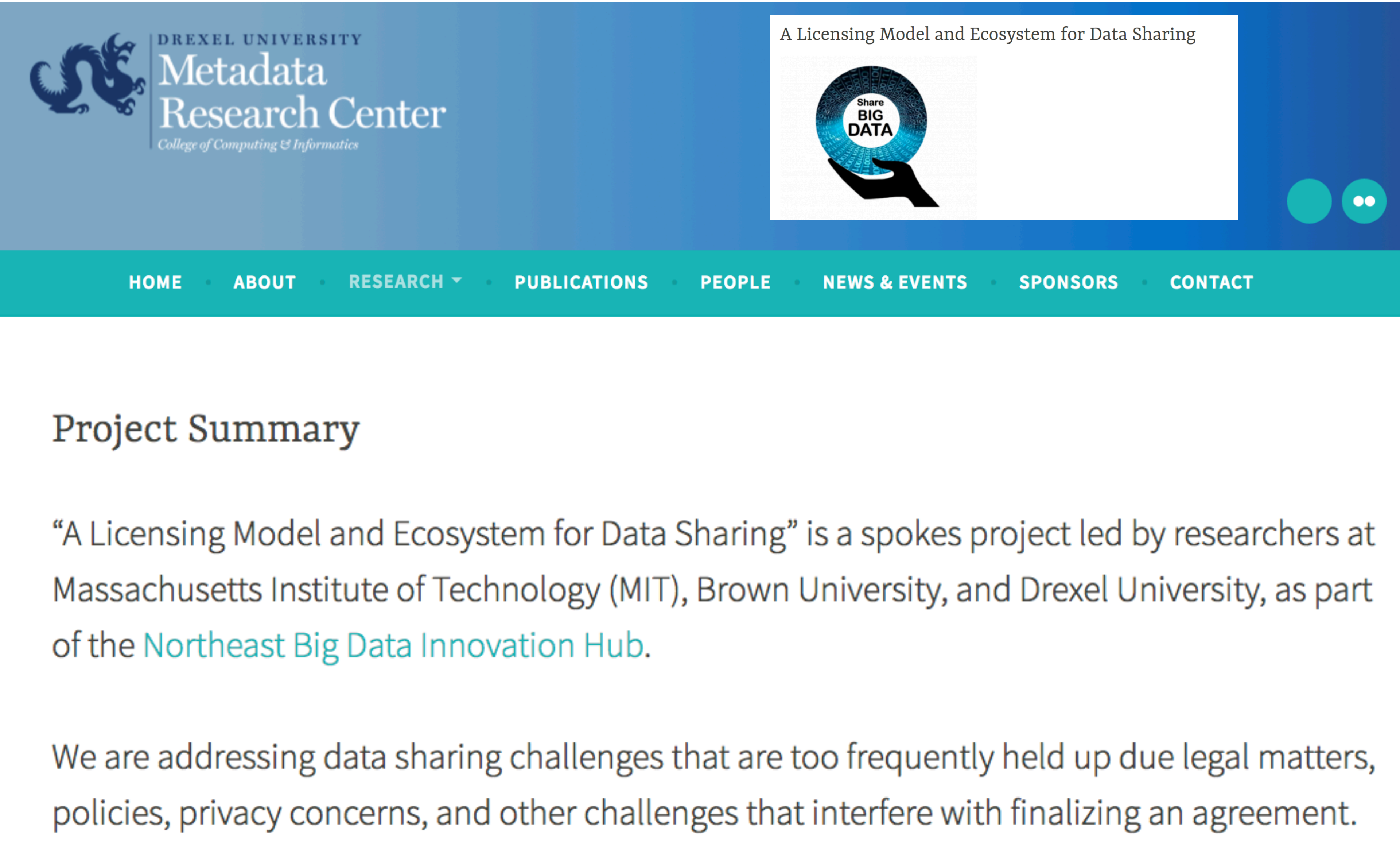
1. Licensing Framework / Generator

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

3. Metadata (Search Licenses and Data)

- Principle: Solve the 80% case!

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



Project Summary

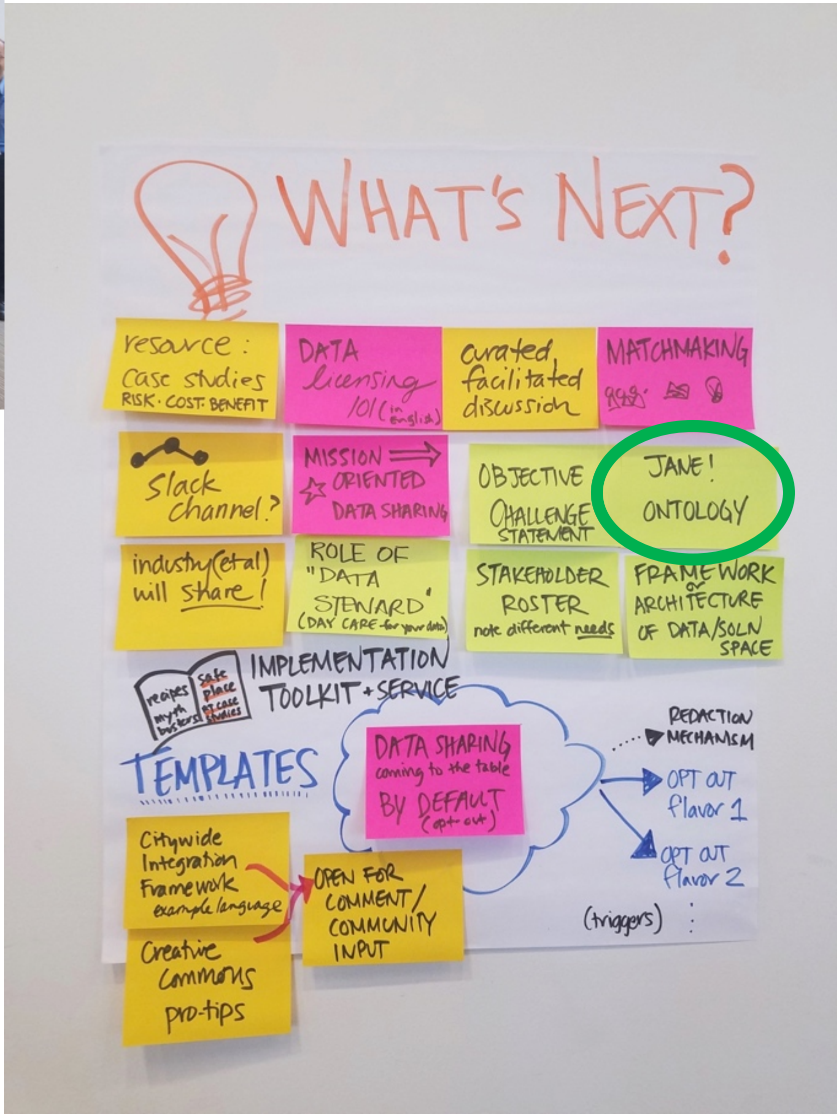
"A Licensing Model and Ecosystem for Data Sharing" is a spokes project led by researchers at Massachusetts Institute of Technology (MIT), Brown University, and Drexel University, as part of the Northeast Big Data Innovation Hub.

We are addressing data sharing challenges that are too frequently held up due legal matters, policies, privacy concerns, and other challenges that interfere with finalizing an agreement.

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

Heard from the trenches…

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


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 of 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

Early stage work

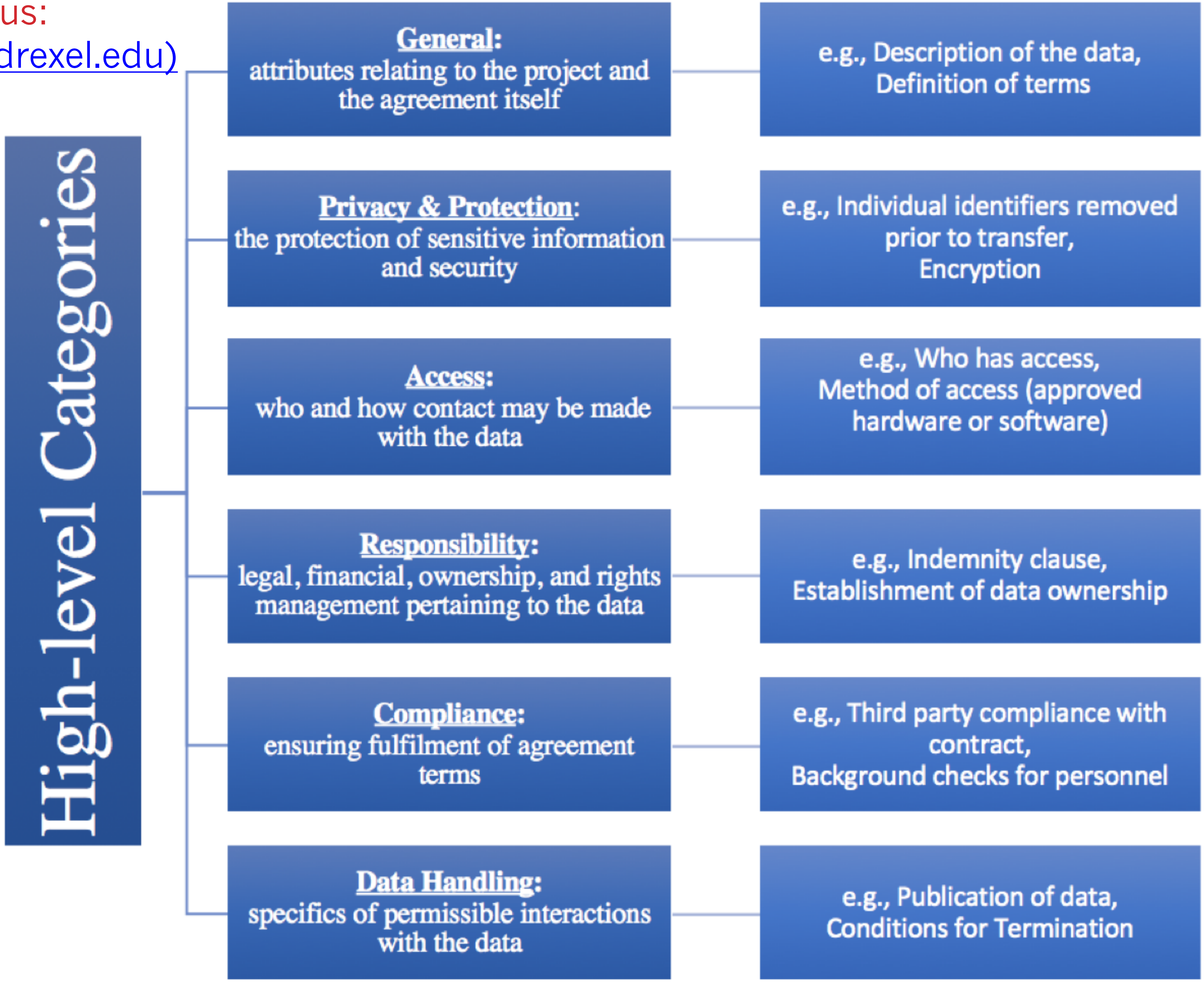
- Content analysis and clustering
- Syntactic analysis, with term proximity comparisons

Content Analysis

1. Data collection
    - 26 data sharing agreements, industry, academia, government
2. Content analysis
    - Confirm data sharing in closed environment
    - Focused, language parsed for higher-level general categories; mid, lower-level *to* → specifications to data handling
3. Concept clustering
    - Classes, sub-classes, attributes organized on a spreadsheet in a classified, hierarchical arrangement.
4. Metadata labeling
    - Language of the categories and attributes was refined

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

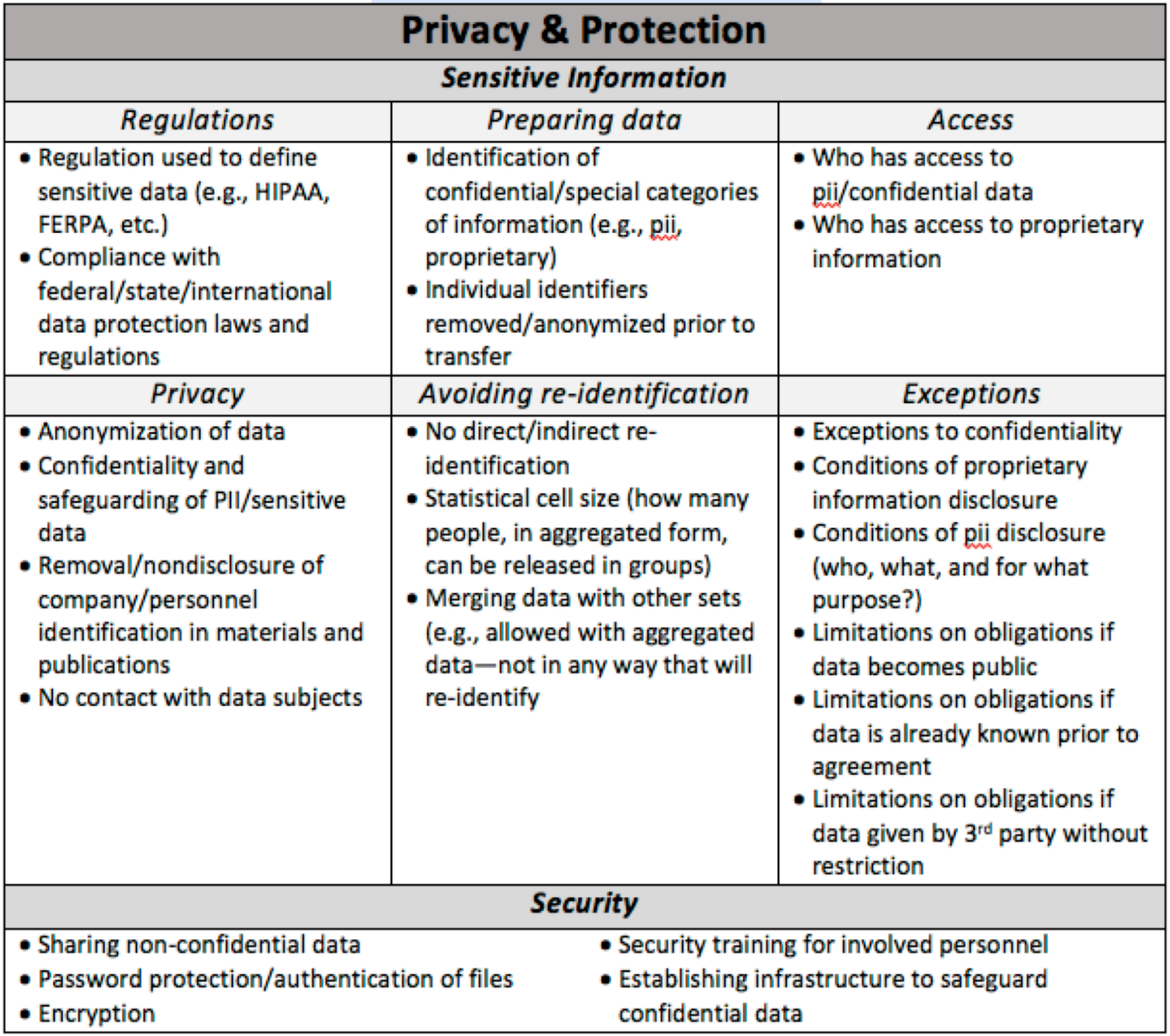
e.g., Indemnity clause, Establishment of data ownership

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



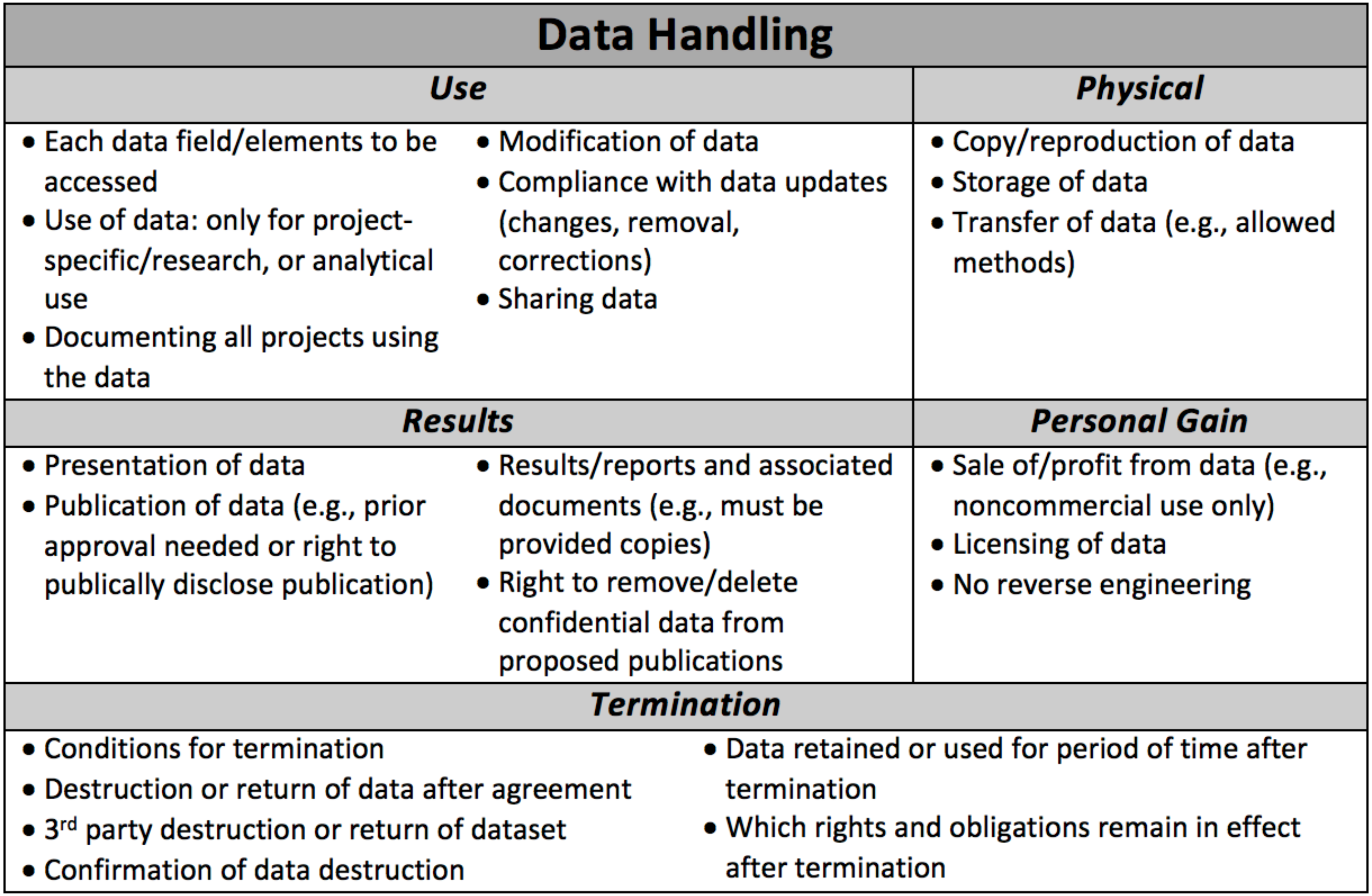
Privacy & Protection

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



Data Handling

Use		***Physical***
• Each data field/elements to be accessed • Use of data: only for project-specific/research, or analytical use • Documenting all projects using the data	• Modification of data • Compliance with data updates (changes, removal, corrections) • Sharing data	• Copy/reproduction of data • Storage of data • Transfer of data (e.g., allowed methods)
Results		***Personal Gain***
• Presentation of data • Publication of data (e.g., prior approval needed or right to publically disclose publication)	• Results/reports and associated documents (e.g., must be provided copies) • Right to remove/delete confidential data from proposed publications	• Sale of/profit from data (e.g., noncommercial use only) • Licensing of data • No reverse engineering
Termination		
• Conditions for termination • Destruction or return of data after agreement • 3rd party destruction or return of dataset • Confirmation of data destruction	• Data retained or used for period of time after termination • Which rights and obligations remain in effect after termination	

6, ~ 40,
90+

Ontologizing

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

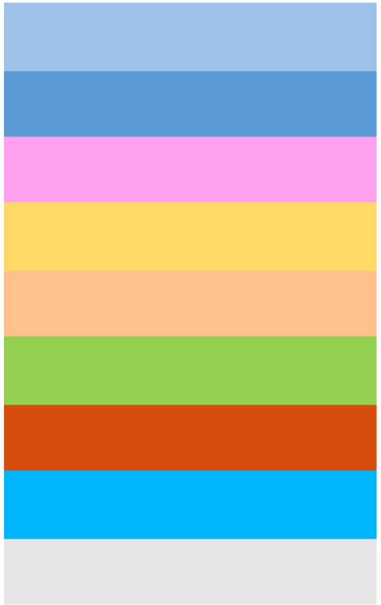
Sample 30 agreements

-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 ⋈ 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 ⋈ 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

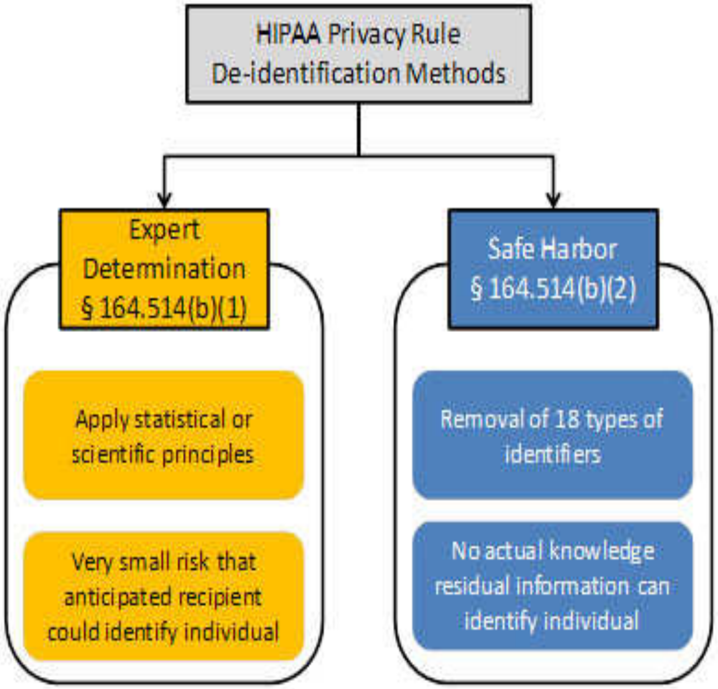
Other requirements

Platform: **First Results**

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

De-identification tool

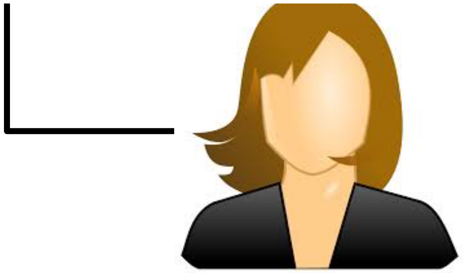
- Detect sensitive columns (rule catalog, user-defined, machine learning, …)
- Automatically de-identify


HIPAA Privacy Rule
De-identification Methods
Expert Determination § 164.514(b)(1)
Safe Harbor § 164.514(b)(2)
Apply statistical or scientific principles
Removal of 18 types of identifiers
Very small risk that anticipated recipient could identify individual
No actual knowledge residual information can identify individual

HIPAA: Interactive DE-identification

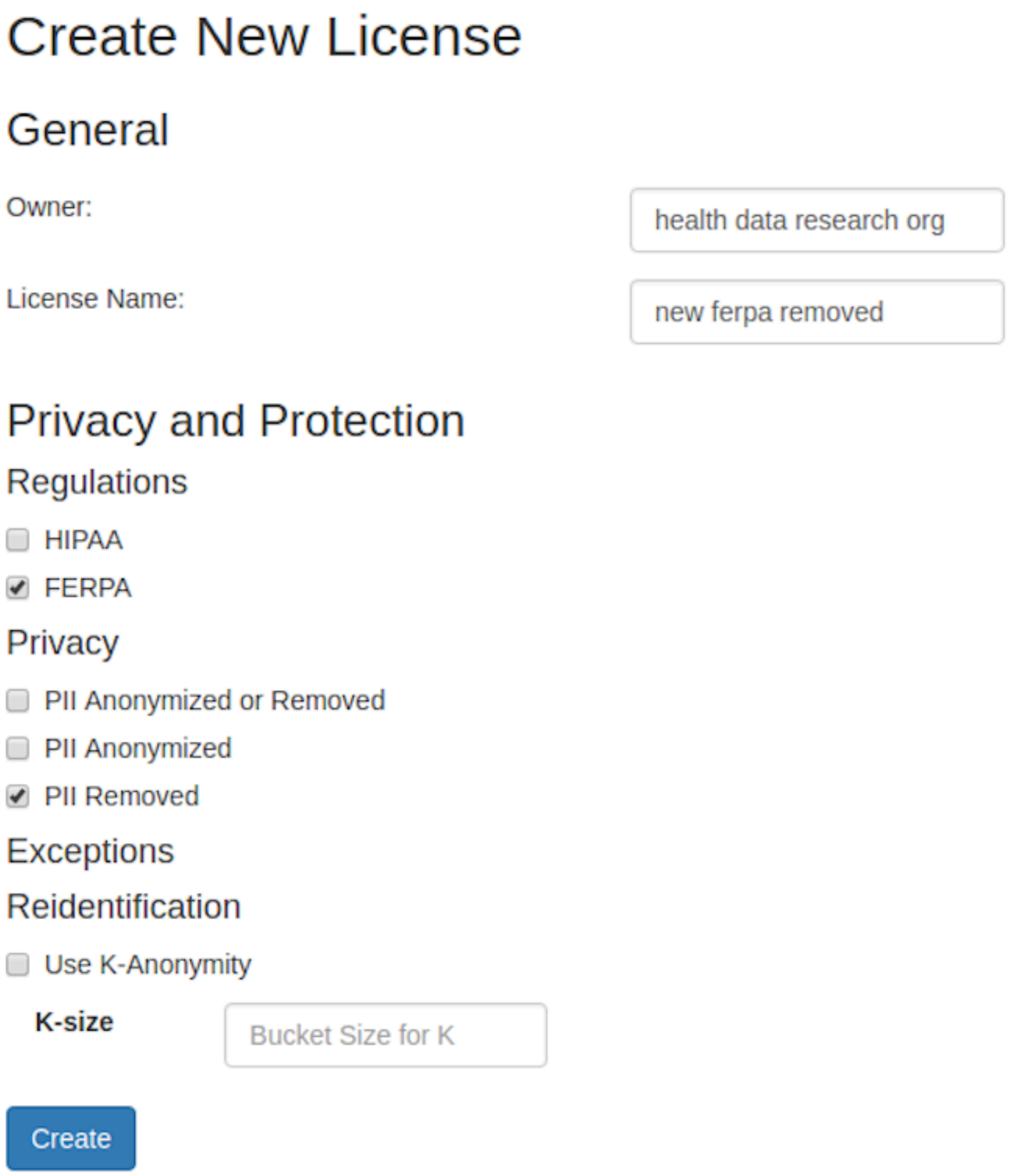
Id ✓	ame ⚠	Street ⚠	City ⚠	State ✓	P-Cod ?	Age ✓
1	J Smith	123 University Ave	Seattle	Washington	98106 ⚠	42
2	Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
3	Bob Wilson	345 Broadway	Seattle	Washington	98101	19
4	M Jones	245 Third Street	Redmond	NULL	98052	299
5	Robert Wilson	345 Broadway St	Seattle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

a



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

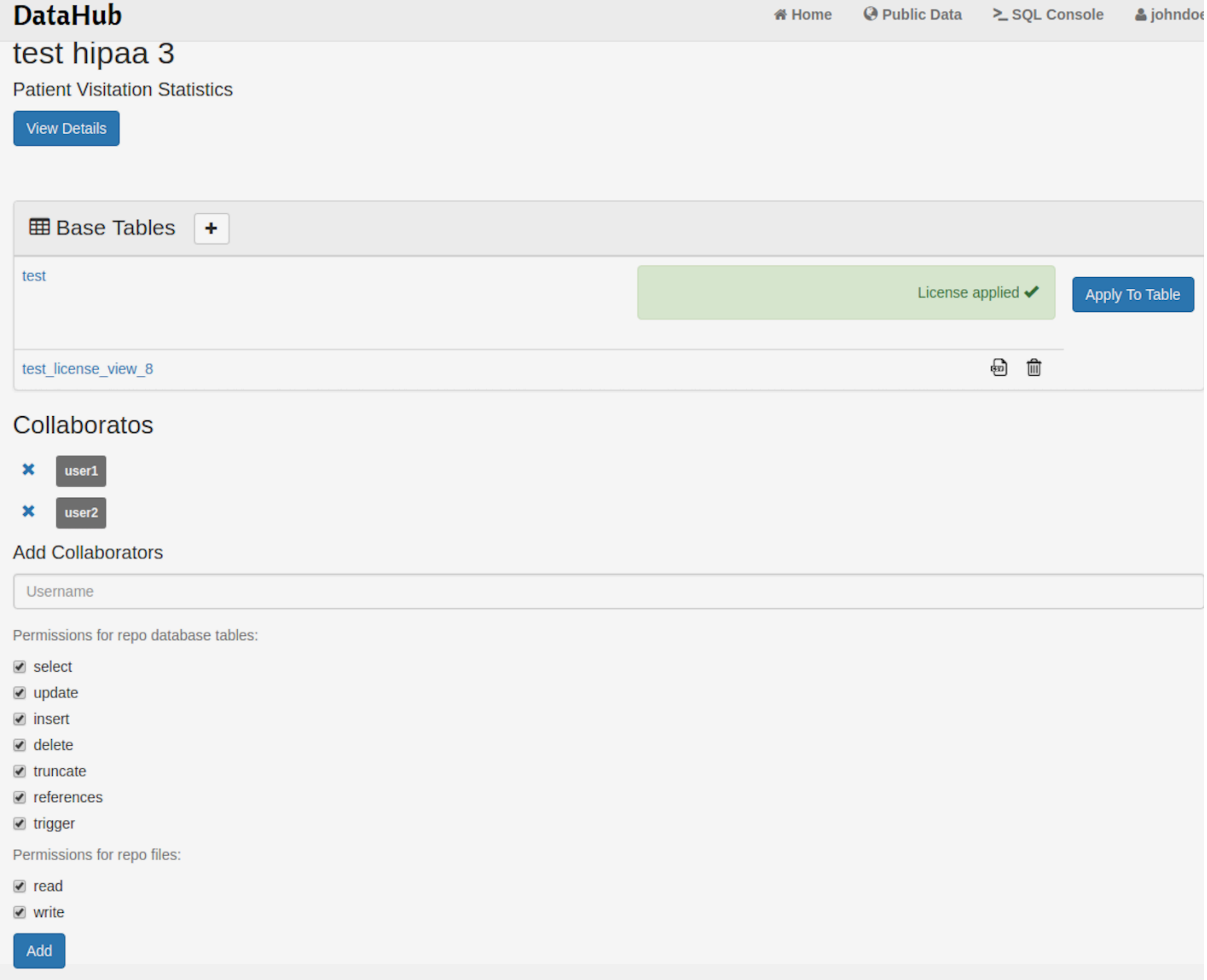
☑ PII Removed

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

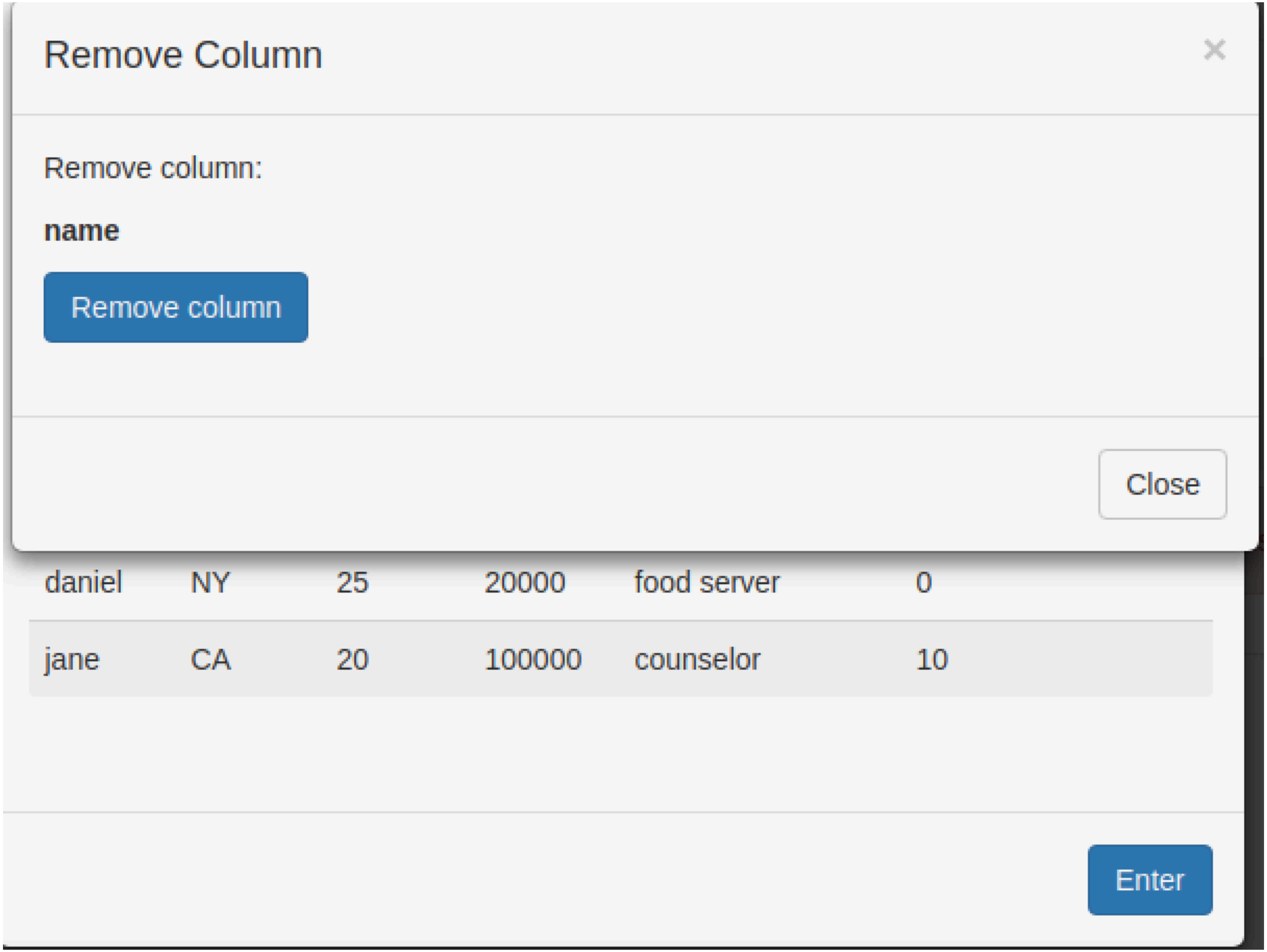
Permissions for repo database tables:

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

Permissions for repo files:

☑ read
☑ write

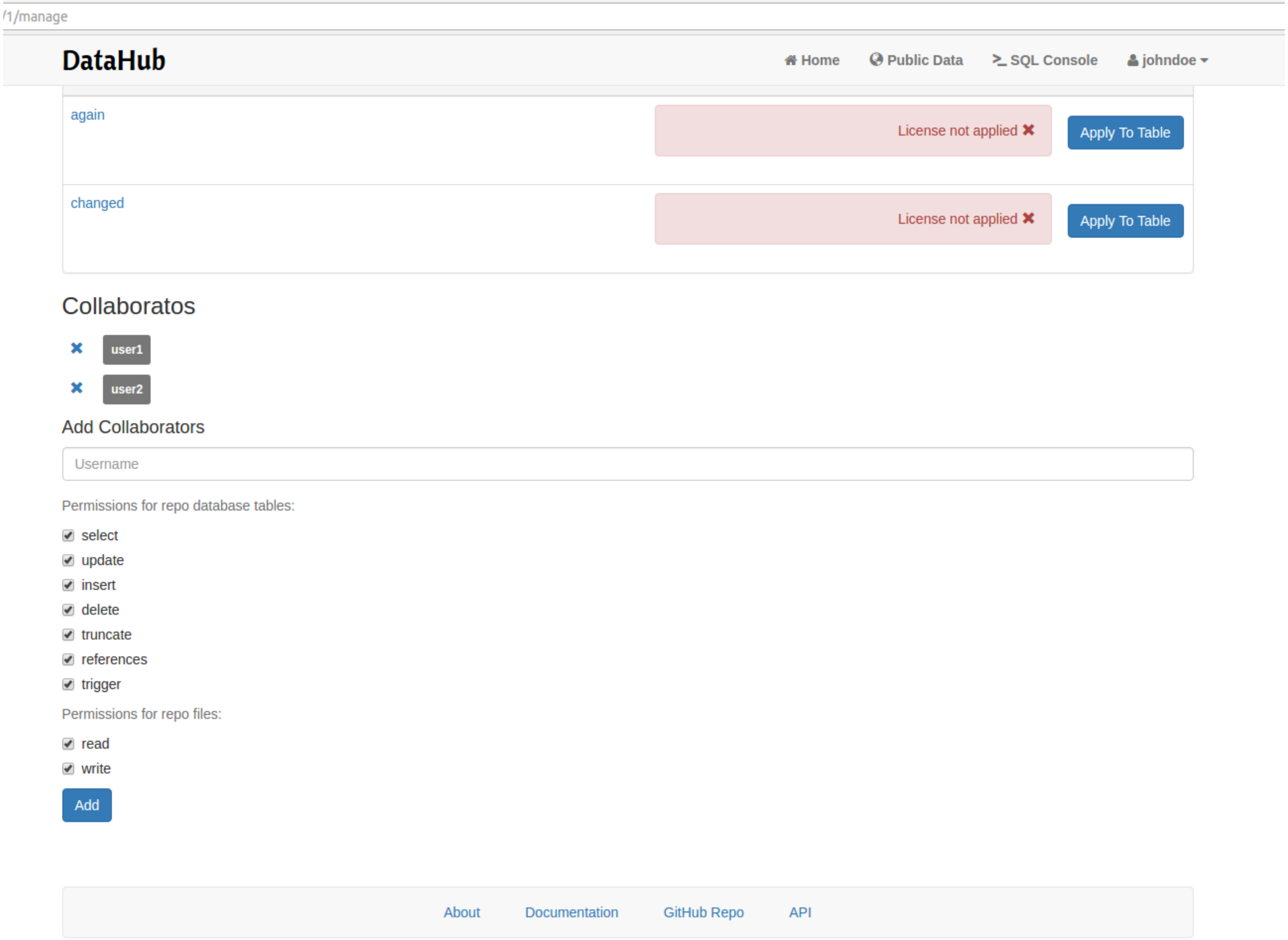
DataHub



Remove column:

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

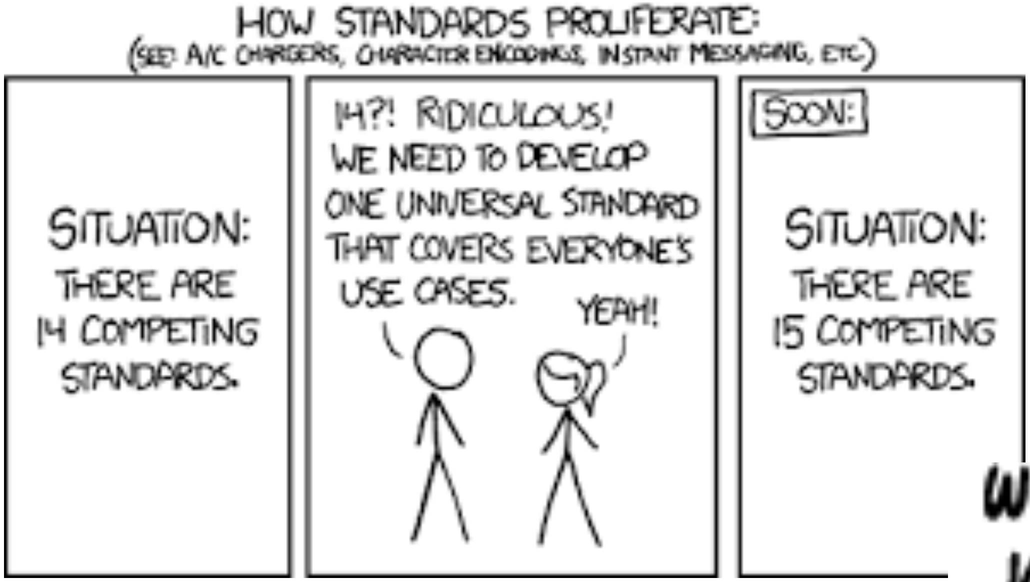
Permissions for repo files:

☑ read
☑ write

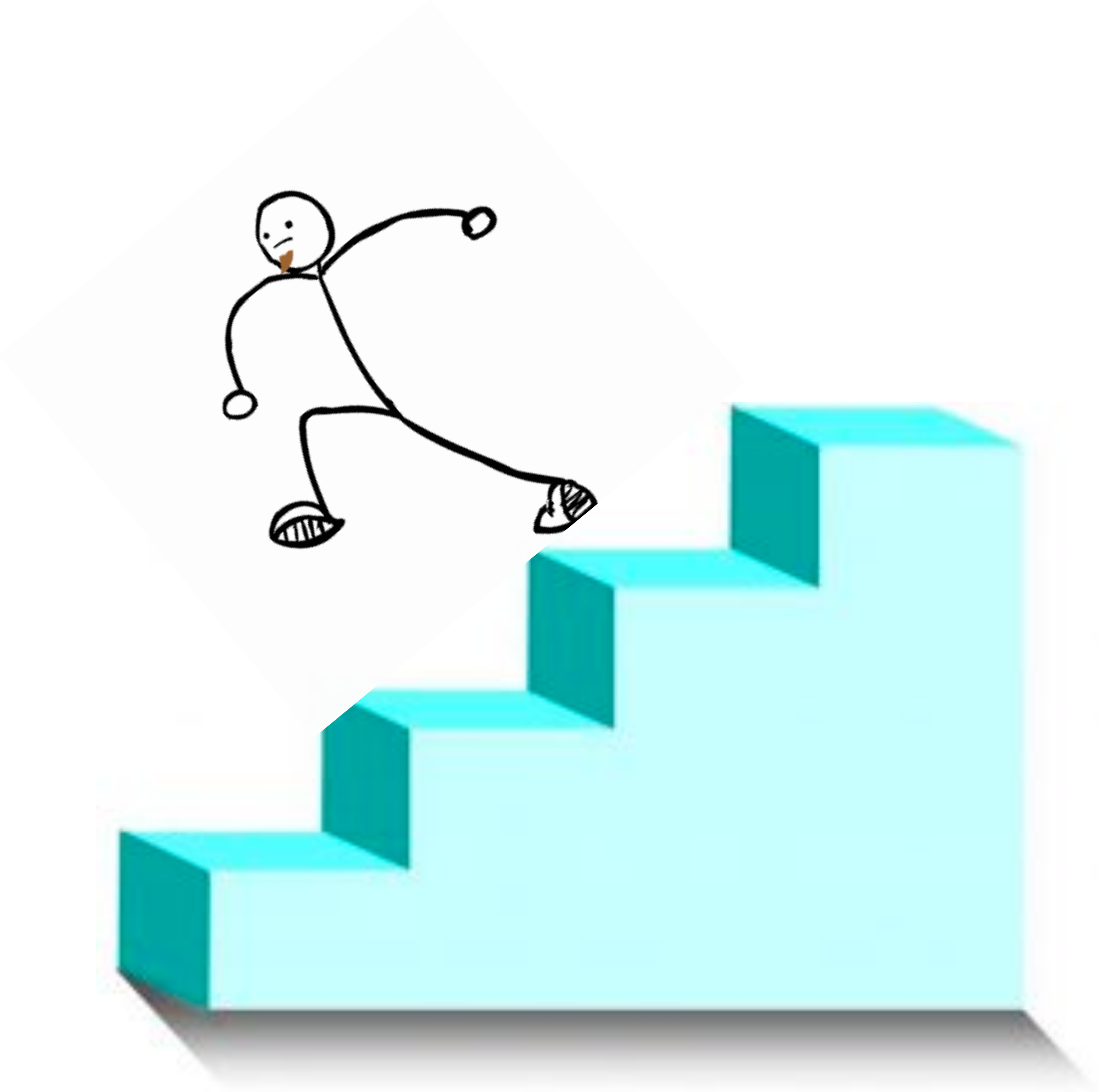
Implications for Digital Libraries?

Standards

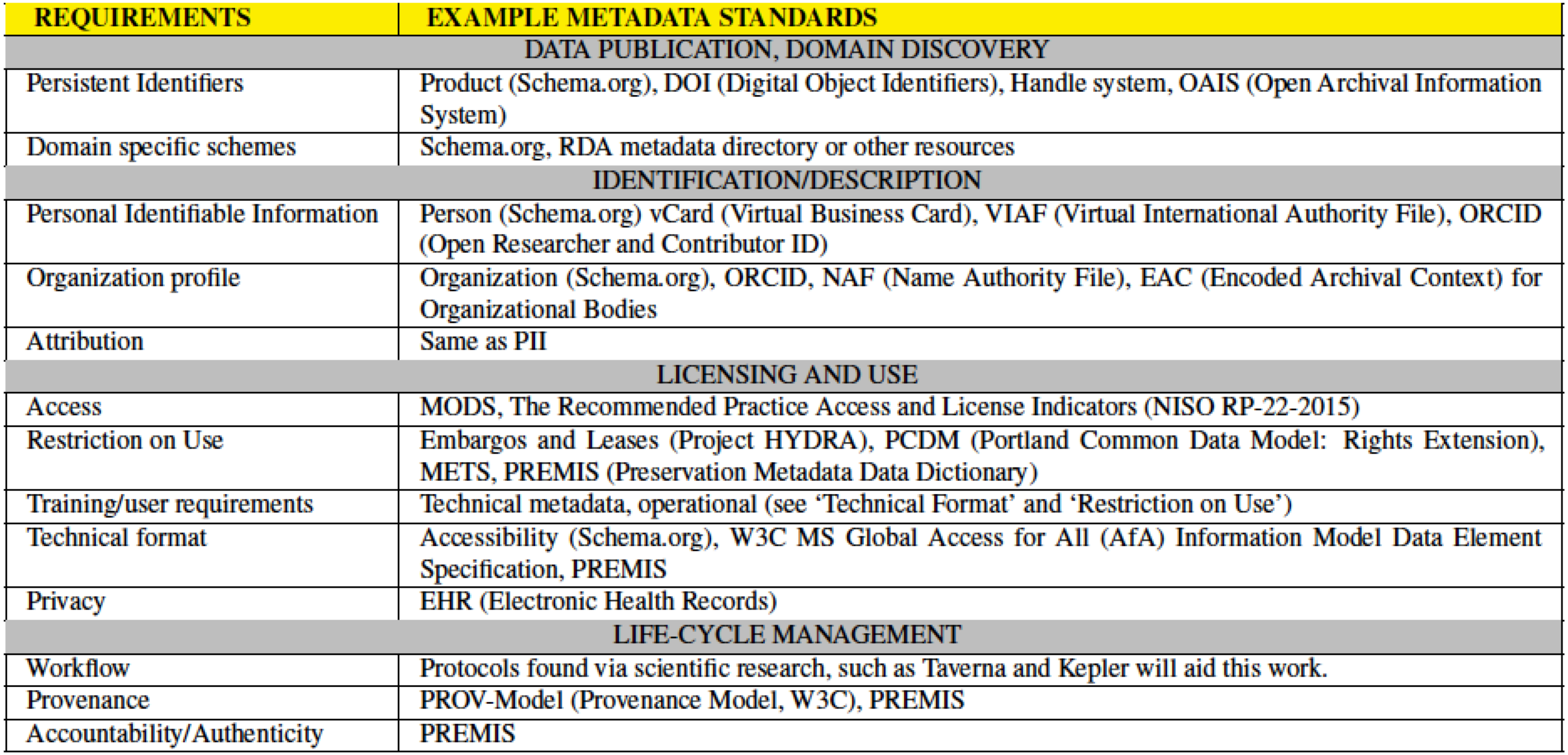
- We are good at this


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?



Lay of the land: Agent, access/rights, + workflow



REQUIREMENTS	EXAMPLE METADATA STANDARDS
DATA PUBLICATION, DOMAIN DISCOVERY	
Persistent Identifiers	Product (Schema.org), DOI (Digital Object Identifiers), Handle system, OAIS (Open Archival Information System)
Domain specific schemes	Schema.org, RDA metadata directory or other resources
IDENTIFICATION/DESCRIPTION	
Personal Identifiable Information	Person (Schema.org) vCard (Virtual Business Card), VIAF (Virtual International Authority File), ORCID (Open Researcher and Contributor ID)
Organization profile	Organization (Schema.org), ORCID, NAF (Name Authority File), EAC (Encoded Archival Context) for Organizational Bodies
Attribution	Same as PII
LICENSING AND USE	
Access	MODS, The Recommended Practice Access and License Indicators (NISO RP-22-2015)
Restriction on Use	Embargos and Leases (Project HYDRA), PCDM (Portland Common Data Model: Rights Extension), METS, PREMIS (Preservation Metadata Data Dictionary)
Training/user requirements	Technical metadata, operational (see 'Technical Format' and 'Restriction on Use')
Technical format	Accessibility (Schema.org), W3C MS Global Access for All (AfA) Information Model Data Element Specification, PREMIS
Privacy	EHR (Electronic Health Records)
LIFE-CYCLE MANAGEMENT	
Workflow	Protocols found via scientific research, such as Taverna and Kepler will aid this work.
Provenance	PROV-Model (Provenance Model, W3C), PREMIS
Accountability/Authenticity	PREMIS

Just a few…existing metadata and rights standards

- Rights statements.org: http://rightsstatements.org/en/documentation/
- Mets: 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/ns/odrl/2/
- ONIX-PL for licensing terms: http://www.editeur.org/21/ONIX-PL/

Connecting with Initiatives

- Rights Data Integration Project (RDI): http://www.rdi-project.org/about2
- UK Copyright Hub: http://www.copyrighthub.org/
- Linked Content Coalition—LCC Rights Reference Model as part of the LCC Framework: http://www.linkedcontentcoalition.org/
- Research Data Alliance
  - Legal interoperability Interest Group
  - RDA/NISO Privacy Task Group

FRAMEWORKS

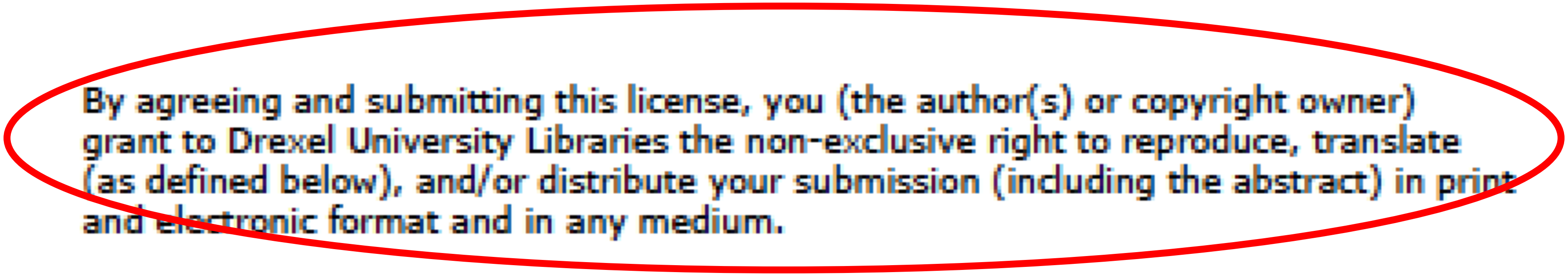
https://www.force11.org/group/fairgroup/fairprinciples

- **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.

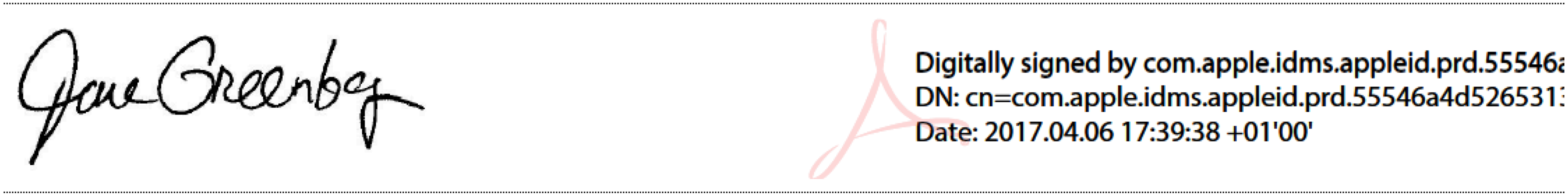
More on implications

- Never a one size fits all
- Housing data, protecting data
- Arching licenses
- Longevity of metadata describing the data
- Other implications

Alternative … repository depostion



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
- *Connecting internationally…*

https://cci.drexel.edu/ShareBigData

Share Big Data

Introduction

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

il 2017, at 19:53.

- Successful agreements
- Share your case
- Links to licenses

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