Toward Standards and Best Practices Related to the Publication, Exchange, and Usage of ~~Open~~ "Sharable" Data

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

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)
   - Standards
     - First-phase KOS for sharing of restricted data
3. Conclusions and next steps

Data sharing advantages

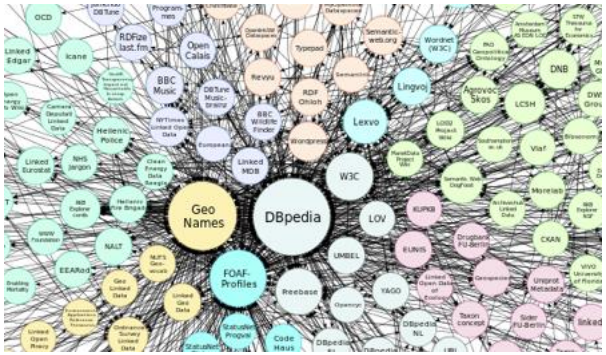
Different Reasons

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

Open data

DataONE

DFC DataNet FEDERATION CONSORTIUM



Closed, restricted 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)

https://project-open-data.cio.gov/

Project Open Data | Edit this Page | Definitions | Guidance | Discuss

PROJECT OPEN DATA

Open Data Policy — Managing Information as an Asset

1. Background

Data is a valuable national resource and a strategic asset to the U.S. Government, its partners, and the public. Managing this data as an asset and making it available, discoverable, and usable – in a word, open – not only strengthens our democracy and promotes efficiency and effectiveness in government, but also has the potential to create economic opportunity and improve citizens' quality of life.

For example, when the U.S. Government released weather and GPS data to the public, it fueled an industry that today is valued at tens of billions of dollars per year. Now, weather and mapping tools are ubiquitous and help everyday Americans navigate their lives.

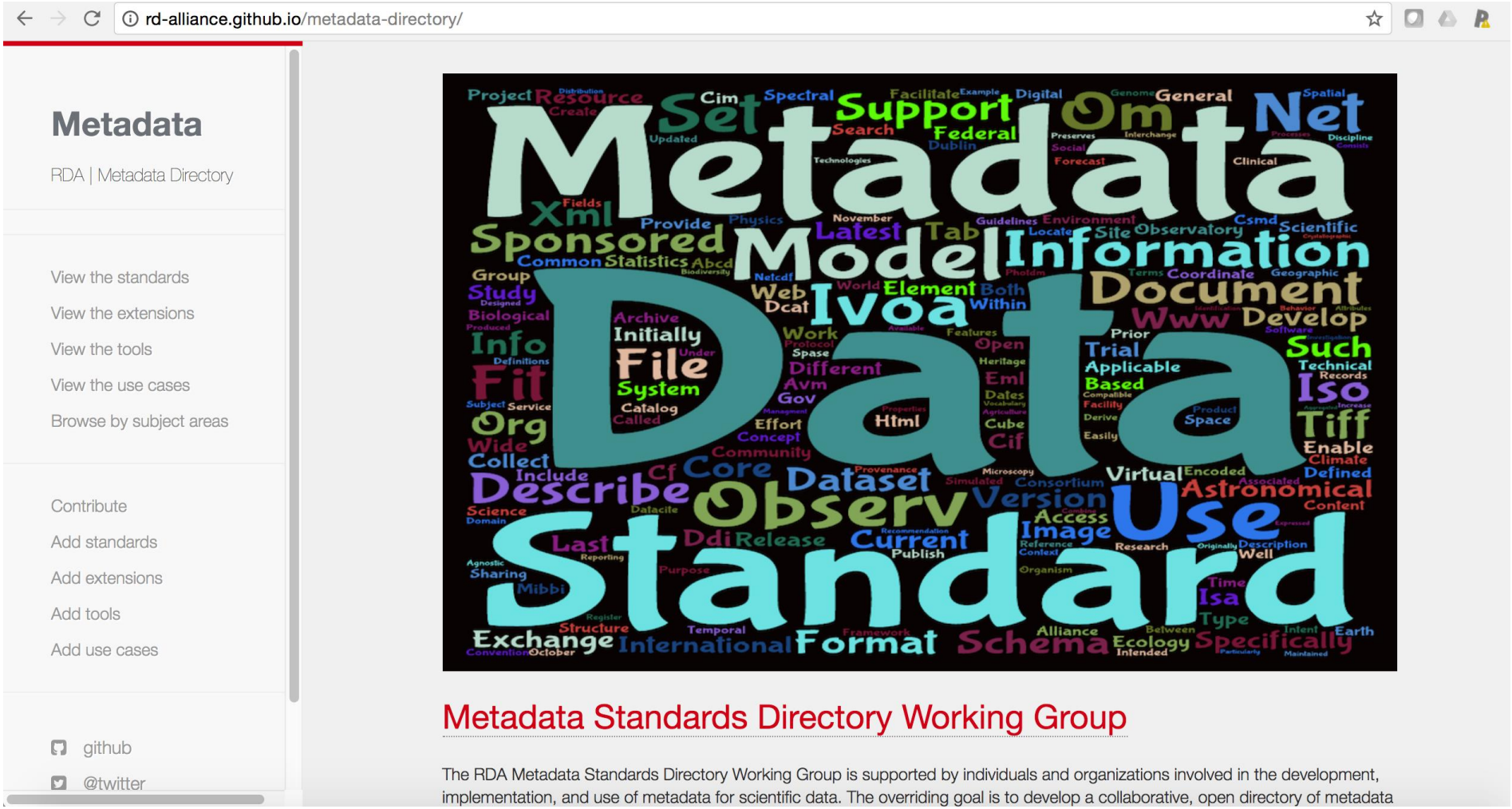
The ultimate value of data can often not be predicted. That's why the U.S. Government released a policy that instructs agencies to manage their data, and information more generally, as an asset from the start and, wherever possible, release it to the public in a way that makes it open, discoverable, and usable.

The White House developed Project Open Data – this collection of code, tools, and case studies – to help agencies adopt the Open Data Policy and unlock the potential of government data. Project Open Data will evolve over time as a community resource to facilitate broader adoption of open data practices in government. Anyone – government employees, contractors, developers, the general public – can view and contribute. Learn more about Project Open Data Governance and dive right in and help to build a better world through the power of open data.

2. Definitions
3. Implementation Guidance
4. Tools

5. Resources
   - 5-1 Metadata Resources - Resources to provide guidance and assistance for each aspect of creating and maintaining agency.gov/data catalog files.
6. Case Studies

http://rd-alliance.github.io/metadata-directory



RDA
RESEARCH DATA ALLIANCE

http://www.dcc.ac.uk/resources/metadata-standards

DCC because good research needs good data

Search

Home | Digital curation | About us | News | Events | Resources | Training | Projects | Community | Tailored support

Home > Resources > Metadata Standards

In this section

- Briefing Papers
- How-to Guides & Checklists
- Developing RDM Services
- Curation Lifecycle Model
- Curation Reference Manual
- Policy and legal
- Data Management Plans
- Tools
- Case studies
- Repository audit and assessment
- **Standards**
  - Disciplinary Metadata
  - DIFFUSE
- Publications and presentations
- Roles
- Curation journals

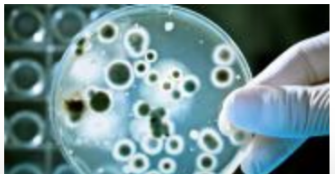
Disciplinary Metadata

While data curators, and increasingly researchers, know that good metadata is key for research data access and re-use, figuring out precisely what metadata to capture and how to capture it is a complex task. Fortunately, many academic disciplines have supported initiatives to formalise the metadata specifications the community deems to be required for data re-use. This page provides links to information about these disciplinary metadata standards, including profiles, tools to implement the standards, and use cases of data repositories currently implementing them.

For those disciplines that have not yet settled on a metadata standard, and for those repositories that work with data across disciplines, the General Research Data section links to information about broader metadata standards that have been adapted to suit the needs of research data.

Please note that a community-maintained version of this directory has been set up under the auspices of the Research Data Alliance.

Search by Discipline



1. *Anyone deposit data into a repository*
2. *Anyone deposit sensitive or restricted data into a repository?*

April
2017

IDEA digital repository Drexel University

IDEA Non-Exclusive Distribution License

In order for the Drexel University E-Repository and Archives (IDEA) to reproduce and distribute your work, your agreement to the following terms is necessary. Please take a moment to read the terms of this license and, if you agree, sign below.

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.

You agree that Drexel University Libraries may, without changing the content, translate the submission to any medium or format for the purposes of preservation.

You also agree that Drexel University Libraries may keep more than one copy of this submission for purposes of security, back-up and preservation.

You represent that the submission is your original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not, to the best of your knowledge, infringe upon anyone's copyright.

If the submission contains material for which you do not hold the copyright, you represent that you have obtained the unrestricted permission of the copyright owner to grant Drexel University Libraries the rights required by this license, and that such third-party owner material is clearly identified and acknowledged within the text or content of the submission.

If the submission is based upon work that has been sponsored or supported by an agency or organization other than Drexel University Libraries, you represent that you have fulfilled any right or review or other obligations required by such contract or agreement.

Drexel University Libraries will clearly identify your name(s) as the author(s) or owner(s) of the submission, and will not make any alterations, other than as allowed by this license, to your submission.

[redacted]

Signature Date

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'

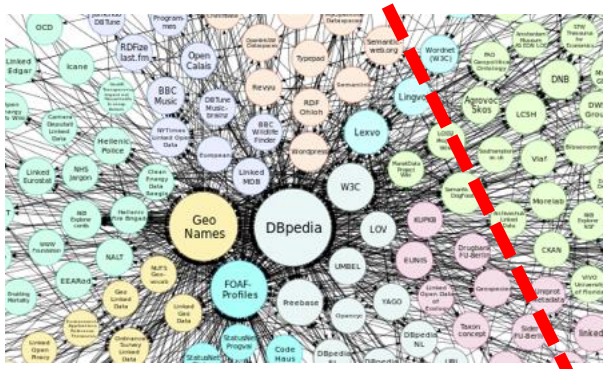
Open data

DRYAD

RDA
RESEARCH DATA ALLIANCE

DataONE

DFC DataNet FEDERATION CONSORTIUM



Closed, restricted data

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

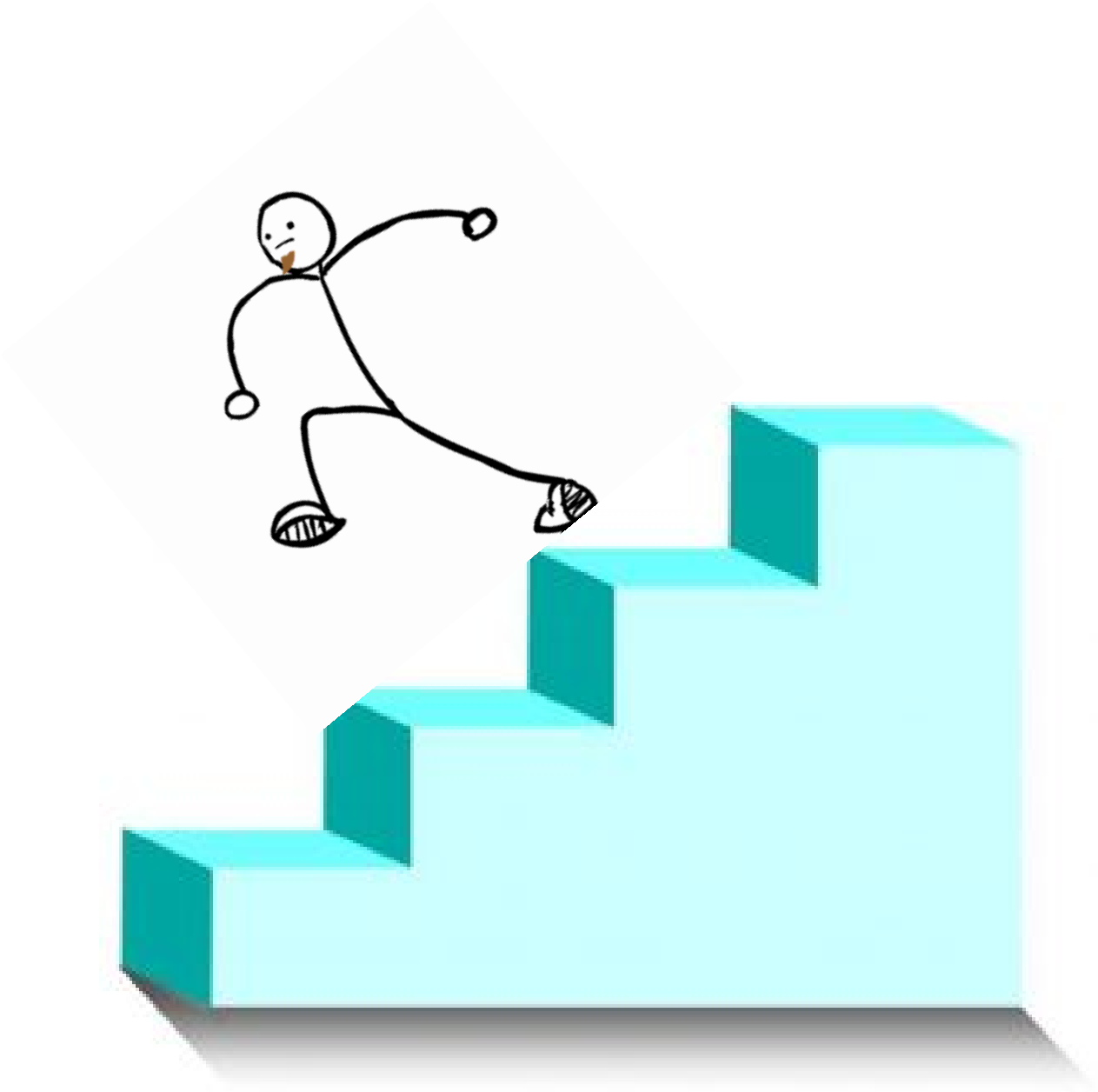
Collaborative Genomics Cloud (CGC)colocalizing massive genomics datasets)

FICO®

FICO score (Fair Isaac Corporation)

Data sharing barriers

Policy	Licensing, agreements	Rights, privacy
▪ 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?*	"Creative commons" (CC) does not address need	Concerns over sensitive information (e.g., PII)
	Security	
	Technical and systematic aspects (policy, regulations, confidentiality/ rights)	**Incentives**
		Why would someone go to all the effort to share their valuable data?



Still, merit in sharing

No sharing without a legal agreement

Involve 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!

MIT CSAIL

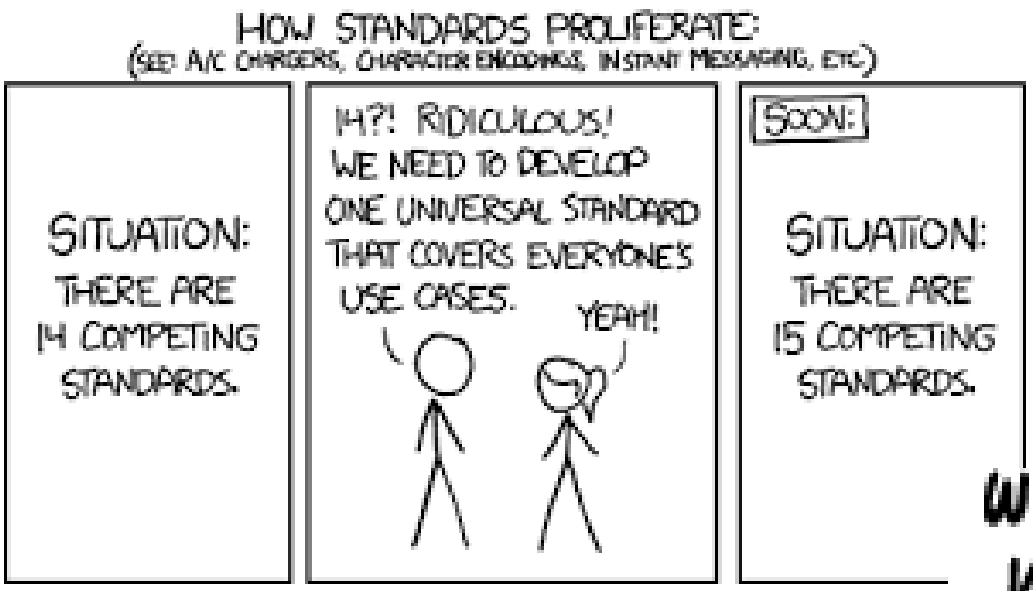
DREXEL UNIVERSITY Metadata Research Center College of Computing & Informatics

BROWN Computer Science

NORTHEAST BIG DATA INNOVATION HUB

Standards

…where do they fit in all of 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

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.

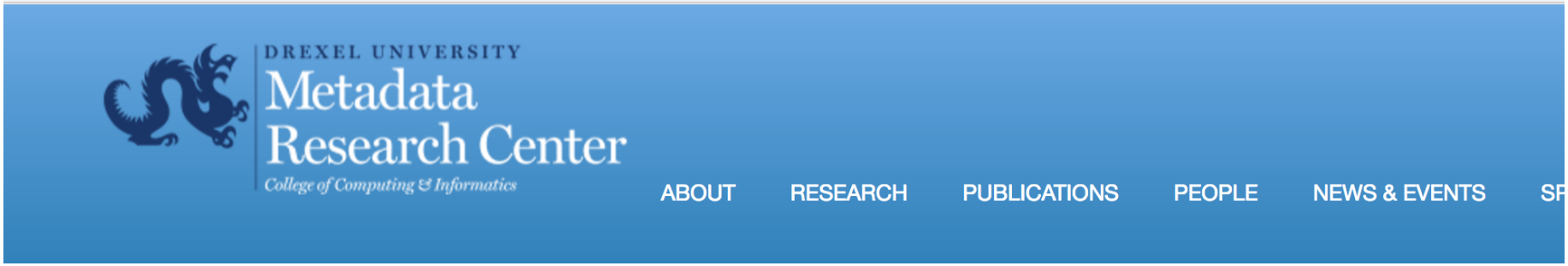
- **INTEROPERABLE:**

- 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.

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



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 interf
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.

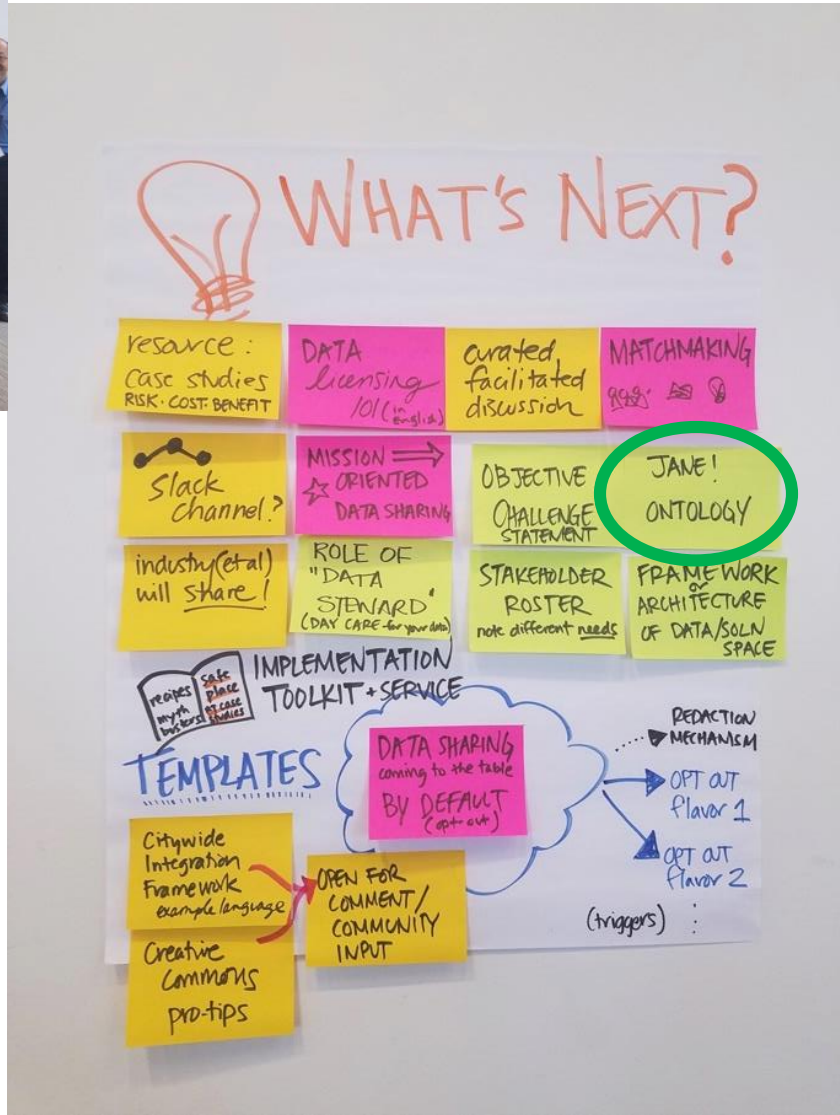
We are addressing these challenges by: 1) Creating a licensing model for data that facilitates sharing data that is not necessarily open or free between
Developing a prototype data sharing software platform, ShareDB that will enforce agreement terms and restrictions for the licenses developed, and (3)
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 t
widely accepted and usable in many application domains (e.g., health and finance).

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

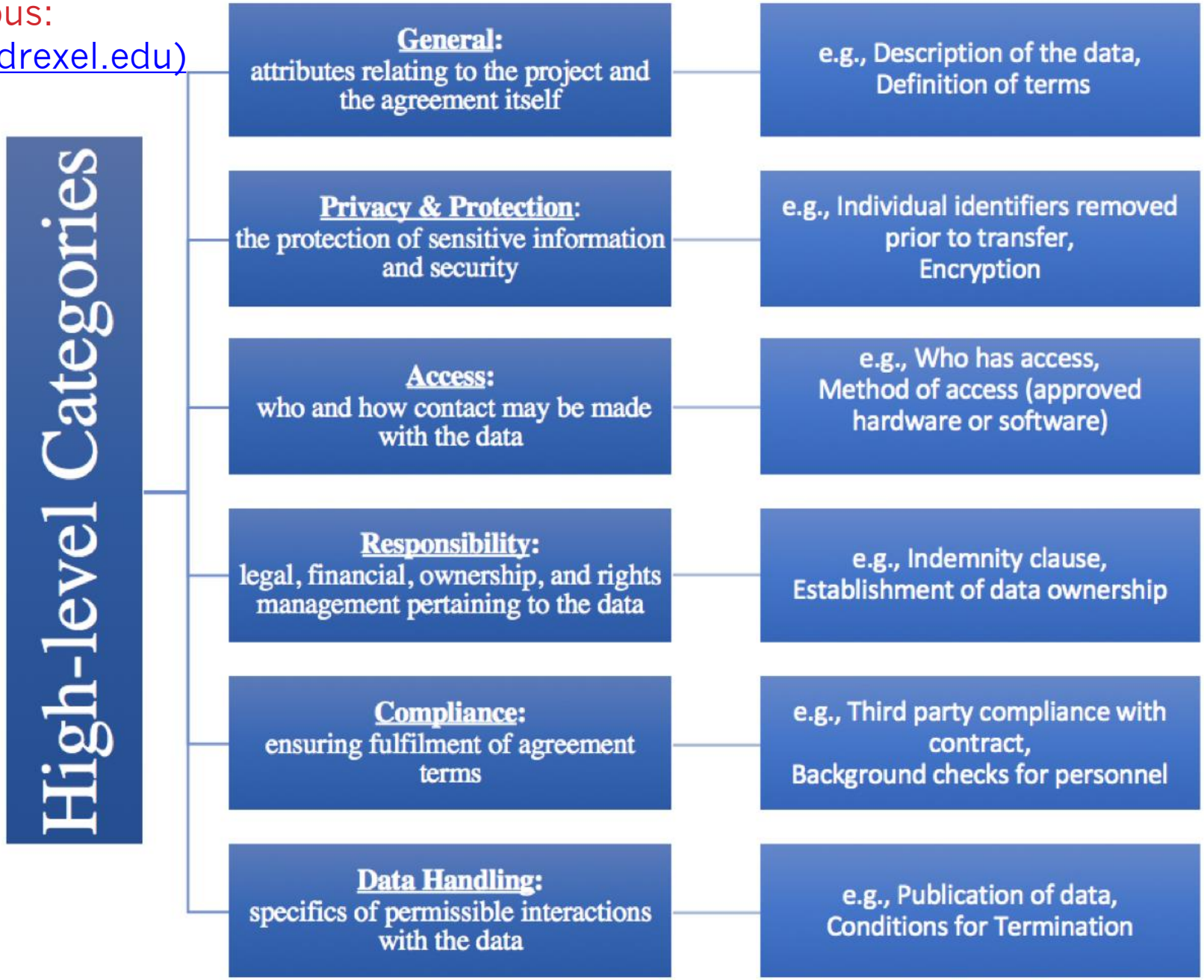
Heard from the trenches…

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



Licenses

(Sam Grabus: smg383@drexel.edu)



Privacy & Protection

Sensitive Information

Regulations

- Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.)
- Compliance with federal/state/international data protection laws and regulations

Preparing data

- Identification of confidential/special categories of information (e.g., pii, proprietary)
- Individual identifiers removed/anonymized prior to transfer

Access

- Who has access to pii/confidential data
- Who has access to proprietary information

Privacy

- 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

Avoiding re-identification

- 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

- 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

- 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

Physical

- Copy/reproduction of data
- Storage of data
- Transfer of data (e.g., allowed methods)

Results

- 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

Personal Gain

- 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 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

Goal: Licensing Framework

Standard terms that researchers, lawyers, and compliance teams conform with

- [x] Controlled access
- [ ] Tracking of access
- [x] Usage rights (e.g., publication, copying)
- [ ] Duration of use
- [x] 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

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**)

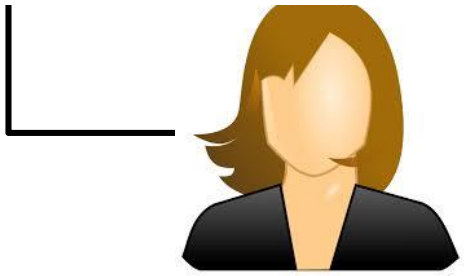
Duration of use

Warrantees of correctness/completeness /availability

Other requirements

HIPAA: Interactive DE-identification

Id	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	J Widom	123 University Ave	Palo Alto	CA	94305	NULL
…	…	…	…	…	…	…



data owner

Conclusions and next steps

- A lot of different efforts in rights area that needs to be brought together
- FAIR principles,
- Data sharing
- Specific to our Spoke, work underway, heavy lifting
  - Mining licenses shows great diversity, but similarities
  - Metadata expertise
- Community building through the NEBDIH and connecting, RDA – Research Data Alliance

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