Standards and Best Practices Related to the Publication, Exchange, and Usage of Open “Sharable” Data
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

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 score (Fair Isaac Corporation)
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
Metadata

RDA | Metadata Directory

View the standards
View the extensions
View the tools
View the use cases
Browse by subject areas

Contribute
Add standards
Add extensions
Add tools
Add use cases

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

Metadata Standards Directory Working Group

The RDA Metadata Standards Directory Working Group is supported by individuals and organizations involved in the development, implementation, and use of metadata for scientific data. The overriding goal is to develop a collaborative, open directory of metadata
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?
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.
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 Greenboy

Digitally signed by com.apple.idms.appleidprd.55546a4d52653163
DN: cn=com.apple.idms.appleidprd.55546a4d52653163
Date: 2017.04.06 17:39:38 +01'00'
Open data

- DRYAD
- DataONE
- 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 score (Fair Isaac Corporation)
- FICO
## 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) does not address need</td>
<td>Concerns over sensitive information (e.g., PII)</td>
</tr>
<tr>
<td>▪ Data lifecycle – data...living thing</td>
<td><strong>Security</strong></td>
<td><strong>Incentives</strong></td>
</tr>
<tr>
<td>~ Do not want to loose control over data downstream</td>
<td>Technical and systematic aspects (policy, regulations, confidentiality/rights)</td>
<td>Why would someone go to all the effort to share their valuable data?</td>
</tr>
</tbody>
</table>
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!
Standards

...where do they fit in all of this
**How Standards Proliferate:**

(See: AC 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**

<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

• 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
• 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.
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 in data sharing agreements.

Sharing of data sets can provide tremendous mutual benefits for industry, researchers, and nonprofit organizations. A major obstacle is that data often come with restrictions on how it can be used. Beyond open data protocols, many attempts to share relevant data sets between different stakeholders in industry are hampered by 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 entities, 2) Developing a prototype data sharing software platform, ShareDB that will enforce agreement terms and restrictions for the licenses developed, and 3) Creating 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 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)
<table>
<thead>
<tr>
<th>Privacy &amp; Protection</th>
<th>Sensitive Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regulations</strong></td>
<td><strong>Preparing data</strong></td>
</tr>
<tr>
<td>• Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.)</td>
<td>• Identification of confidential/special categories of information (e.g., pii, proprietary)</td>
</tr>
<tr>
<td>• Compliance with federal/state/international data protection laws and regulations</td>
<td>• Individual identifiers removed/anonymized prior to transfer</td>
</tr>
<tr>
<td><strong>Privacy</strong></td>
<td><strong>Avoiding re-identification</strong></td>
</tr>
<tr>
<td>• Anonymization of data</td>
<td>• No direct/indirect re-identification</td>
</tr>
<tr>
<td>• Confidentiality and safeguarding of PII/sensitive data</td>
<td>• Statistical cell size (how many people, in aggregated form, can be released in groups)</td>
</tr>
<tr>
<td>• Removal/nondisclosure of company/personnel identification in materials and publications</td>
<td>• Merging data with other sets (e.g., allowed with aggregated data—not in any way that will re-identify)</td>
</tr>
<tr>
<td>• No contact with data subjects</td>
<td></td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td></td>
</tr>
<tr>
<td>• Sharing non-confidential data</td>
<td></td>
</tr>
<tr>
<td>• Password protection/authentication of files</td>
<td>• Security training for involved personnel</td>
</tr>
<tr>
<td>• Encryption</td>
<td></td>
</tr>
<tr>
<td><strong>Data Handling</strong></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>---</td>
</tr>
<tr>
<td><strong>Use</strong></td>
<td><strong>Physical</strong></td>
</tr>
<tr>
<td>• Each data field/elements to be accessed</td>
<td>• Copy/reproduction of data</td>
</tr>
<tr>
<td>• Use of data: only for project-specific/research, or analytical use</td>
<td>• Storage of data</td>
</tr>
<tr>
<td>• Documenting all projects using the data</td>
<td>• Transfer of data (e.g., allowed methods)</td>
</tr>
<tr>
<td><strong>Results</strong></td>
<td><strong>Personal Gain</strong></td>
</tr>
<tr>
<td>• Presentation of data</td>
<td>• Sale of/profit from data (e.g., noncommercial use only)</td>
</tr>
<tr>
<td>• Publication of data (e.g., prior approval needed or right to publicly disclose publication)</td>
<td>• Licensing of data</td>
</tr>
<tr>
<td>• Results/reports and associated documents (e.g., must be provided copies)</td>
<td>• No reverse engineering</td>
</tr>
<tr>
<td>• Right to remove/delete confidential data from proposed publications</td>
<td></td>
</tr>
<tr>
<td><strong>Termination</strong></td>
<td></td>
</tr>
<tr>
<td>• Conditions for termination</td>
<td>• Data retained or used for period of time after termination</td>
</tr>
<tr>
<td>• Destruction or return of data after agreement</td>
<td>• Which rights and obligations remain in effect after termination</td>
</tr>
<tr>
<td>• 3rd party destruction or return of dataset</td>
<td></td>
</tr>
<tr>
<td>• Confirmation of data destruction</td>
<td></td>
</tr>
</tbody>
</table>

6, ~ 40, 90+
• Privacy & Protection
  □ 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 (e.g., changes, removal, corrections)
  ▪ Sharing 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
Goal: Licensing Framework

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

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

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

... ... ... ... ... ... ...

**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