Governance of Research Data

MacKenzie Smith, MIT Research Director, and
Creative Commons Science Fellow

Data Governance is…

- a *system of decision rights and responsibilities* that describe who can take what actions with what data, and when, under what circumstances, using what methods;
- includes *laws and policies* associated with data, and *strategies* for data management in the context of an organization;
- includes *processes* that insure important data assets are formally managed throughout an organization, including *business processes* and *risk management*;
- ensures that data can be *trusted* and that people can be made *accountable* for actions affecting the data

Goals of Research Data Archiving

- research reproducibility
- reusability: large-scale data interoperability
  - Includes technical, social, legal and policy aspects
  - usual focus on technical/social
  - focus here on *legal/policy aspects*
- fiscal responsibility for tax-funded research
- broadest possible impact

Trends in Research Data Archiving

- Journal publishers starting to mandate data deposit
  e.g. Journal of Evolutionary Biology
- Funders starting to mandate data management plans
  e.g. U.S. NIH, NSF
- No common practice yet, unclear expectations from funders and publishers, researchers have no clue, institutions scrambling to deal with this

Goals for Researchers

- ***Credit***
- First access and ease of reuse themselves
- Easier integration/interoperability
  (i.e. “re-purposing”)
- Help complying with grant terms
- Trusted advice
  (no interest in complex legal issues)

NSF Data Management Plans "may include"

1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;

2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);

3. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;

4. policies and provisions for re-use, re-distribution, and the production of derivatives; and

5. plans for archiving data, samples, and other research products, and for preservation of access to them.

International Collaborations

Q: If I participate in a collaborative international research project, do I need to be concerned with data management policies established by institutions outside the United States?

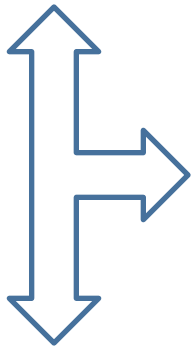
A: Yes. There may be cases where data management plans are affected by formal data protocols established by large international research consortia or set forth in formal science and technology agreements signed by the United States Government and foreign counterparts. Be sure to discuss this issue with your sponsored projects office (or equivalent) and your international research partner when first planning your collaboration.

Some Definitions

- *Credit* is what researchers want
- *Citation* is the norm in scholarly communication to provide supporting evidence, now proxy for credit
- *Attribution* is legally-imposed, remedy is lawsuit; *does not* insure credit or citation

Legal Mechanisms for Sharing Data

1. licenses
2. contracts



Require attribution

3. waivers

No attribution requirement

Copyright for Data

- Does not apply to facts, e.g. most scientific data
- Can apply to a *collection of facts*, but only to original aspects, not facts themselves
- Can extract facts from a copyrighted database without infringing

Licenses

- Licenses ≠ contracts
  - depend on *underlying rights*, e.g. copyright or sui generis rights
  - Copyright is a bundle of rights, automatic when fixed, limited in scope and duration

- US and EU differ (sui generis data rights) so different licenses cover copyright, sui generis rights, or both

Creative Commons (CC-BY) Example

- applies to data and databases to the extent they're copyrightable
- only data uses that implicate copyright trigger attribution requirement
- data uses that do *not* implicate copyright, e.g. in the public domain, do *not* trigger attribution

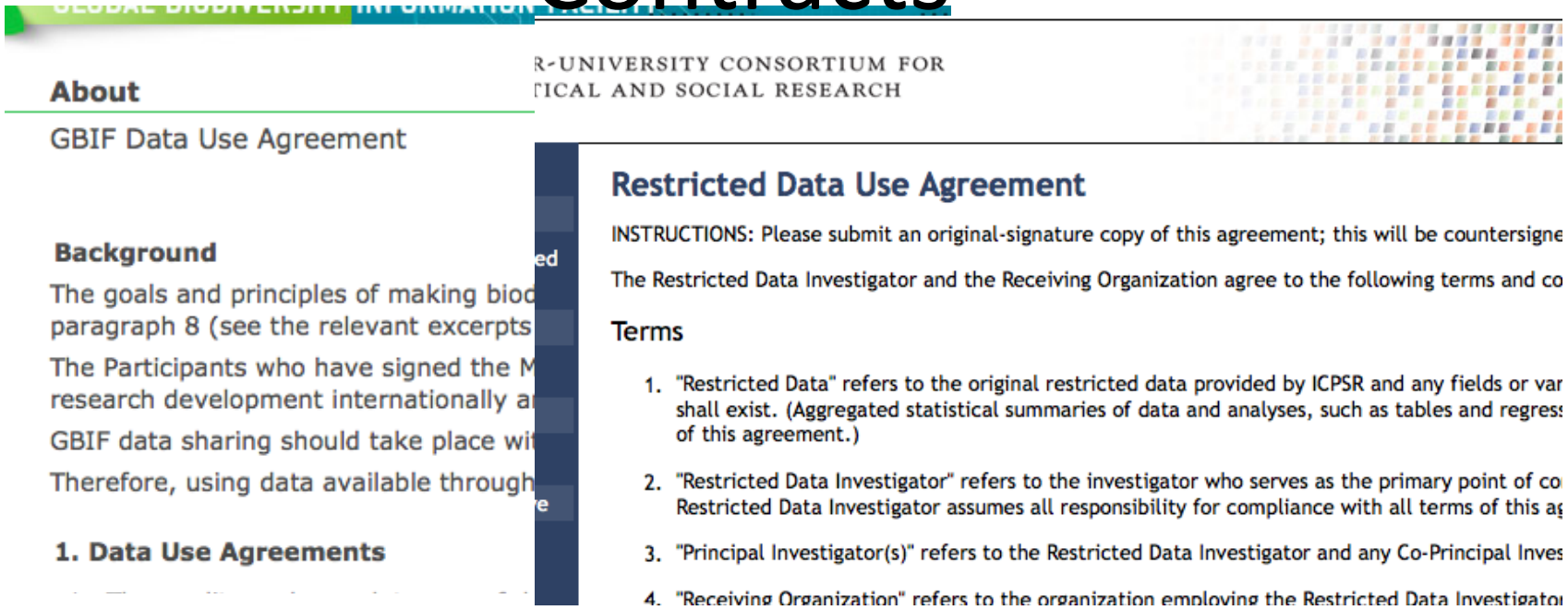
Licenses

- Hard to assess copyright for particular data and databases
- Hard to know when license applies, creates risks:
  - data provider might be misled about protection
  - data user will under- or over-comply

Licenses

- Attribution requirements are inflexible, causing absurd situations
  - e.g. providing attribution to 1,000 providers in 1,000 different ways
  - known as ‘attribution stacking’
- Could provide good attribution and still not satisfy norms or expectations

Contracts



About

GBIF Data Use Agreement

Background

The goals and principles of making biod
paragraph 8 (see the relevant excerpts
The Participants who have signed the M
research development internationally a
GBIF data sharing should take place wi
Therefore, using data available through

1. Data Use Agreements

R-UNIVERSITY CONSORTIUM FOR
TICAL AND SOCIAL RESEARCH

Restricted Data Use Agreement

INSTRUCTIONS: Please submit an original-signature copy of this agreement; this will be countersigne

The Restricted Data Investigator and the Receiving Organization agree to the following terms and co

Terms

1. "Restricted Data" refers to the original restricted data provided by ICPSR and any fields or var shall exist. (Aggregated statistical summaries of data and analyses, such as tables and regress of this agreement.)
2. "Restricted Data Investigator" refers to the investigator who serves as the primary point of co Restricted Data Investigator assumes all responsibility for compliance with all terms of this ag
3. "Principal Investigator(s)" refers to the Restricted Data Investigator and any Co-Principal Inves
4. "Receiving Organization" refers to the organization employing the Restricted Data Investigato

LTER Network Data Access Policy, Data Access Requirements, and General Data Use Agreement

approved by the LTER Coordinating Committee April 6, 2005

Long Term Ecological Research Network Data Access Policy

The LTER data policy includes three specific sections designed to express shared network policies regarding the release of LTER data products, user registration for accessing data, and the licensing agreements specifying the conditions for data use.

Contracts

- Don't require underlying legal right
  - rely on offer/acceptance, click through, terms of use
  - require formalities, e.g. attribution
- Have big downsides
  - confusing obligations, no standardization, each user agreement can have different requirements
- Researchers may avoid data if they can't understand the terms of use

Contracts

Unlike licenses, contracts only binds *parties*

- If someone obtains *licensed* data and shares it, anyone who obtains data from that user is still bound by the license
- If data had been shared by *contract*, anyone obtaining data from the second party is *not* bound by the contract since they aren't a party to the contract
- In this respect, contracts are more *limited* than licenses

Contracts

- Can have a *broader reach* than licenses
  - not tied to a legal right
  - can take away rights of public

Example


w.data.gc.ca/default.asp?lang=En&n=46D15882-1
global congress american
Login - Creative...
Creative Commo...
Government of Canada
Gouvernement du Canada
Canada
Open Data
www.data.gc.ca
Français
Home
Contact Us
Help
Search
canada.gc.ca
Home > Licence Agreement
GC Open Data
Backgrounder
Advanced Search
Browse
Licence Agreement
FAQ
Tools
Links
Research Data Canada
The Canadian Astronomy Data Centre
Government of Canada Open Data Licence Agreement for Unrestricted Use of Canada's Data
This licence agreement is between Her Majesty the Queen in Right of Canada, as represented by the Ministers of Participating Departments (Agriculture and Agri-food Canada, Canadian International Development Agency, Citizenship and Immigration (Canada), Environment Canada, Fisheries and Oceans Canada, Library and Archives Canada, Natural Resources Canada, National Research Canada, Statistics Canada, Transport Canada, and Treasury Board Secretariat collectively referred to herein as "Canada") and you.
The following are terms and conditions governing your use of this data. By browsing through, downloading, accessing or otherwise using this data, you acknowledge that you have read, understood and agree to be legally bound by the terms and conditions set out below. If you do not agree to these terms and conditions, you may not browse through, download, access or otherwise use this data.
Canada may modify this agreement at any time, and such modifications shall be effective immediately upon posting of the modified agreement on the Open Data site. Your continued access or use of this data shall be deemed your conclusive acceptance of the modified agreement.

Waivers

- Provide legal certainty
  - No need to decipher copyright protection or sift through confusing legalese
  - Better than silence, to avoid forcing people to guess what their risks are
- Mean loss of control
  - Can't *require* attribution or any other terms
- Avoid problems and rely on scholarly norms
  - no attribution stacking or inappropriate obligations

CC0 1.0 Universal (CC0 1.0)
Public Domain Dedication

3 levels: Waiver, Fall-back license, Non-assertion pledge

This is a human-readable summary of the **Legal Code (read the full text)**.

Disclaimer

No Copyright

The person who associated a work with this deed has **dedicated** the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law.

You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. See **Other Information** below.

Other Information

- In no way are the patent or trademark rights of any person affected by CC0, nor are the rights that other persons may have in the work or in how the work is used, such as **publicity or privacy** rights.
- Unless expressly stated otherwise, the person who associated a work with this deed makes no warranties about the work, and disclaims liability for all uses of the work, to the fullest extent permitted by applicable law.

Summary of Legal Tools

- Law is messy, each approach has consequences
- Licenses – (1) legal uncertainty about scope, (2) can be inconsistent with scholarly norms
- Contracts – (1) burdensome requirements with custom terms, (2) can include requirements that take away normal rights
- Waivers – (1) avoid problems, but (2) lose control and rely solely on norms

Summary of Legal Tools

- Each approach requires loss of control
- No mechanism imposes legally-binding obligations in way that perfectly maps to scholarly credit, e.g. citation
- Ideal is least friction to science while giving credit where due, i.e. waivers and norms

Privacy and Confidentiality

- Governed by law (e.g. HIPAA) and norms (e.g. IRB reviews)
- Not addressed by licenses or waivers, only contracts
- Growing tension between privacy and sharing interests

  e.g. portable consents for disease research using human genomic data

Data Licensing Practices

- Wide variation in specialized subject archives, usually custom contract (DULA)
- IR platforms (e.g. DSpace) support optional CC licenses and CC0 waivers
- US Gov data is in the Public Domain, but usually lack explicit rights statement

Data Reuse

Scientific progress requires international interoperability and frictionless data integration at very large-scale (e.g. the Web)

Data interoperability includes

1. technical issues (data integration, protocols)
2. social issues (scientific norms, credit mechanisms or lack thereof)
3. legal issues (incompatible laws and policies for data and databases)

Technical Issues

- License interoperability
- Attribution stacking
- Persistent IDs
- Provenance
- Metadata
- Data structure
- Data accessories

Persistent IDs

Need “citable” URIs for

- People (e.g. ORCIDs)
- Institutions (e.g. NISO I2 identifiers)
- Documentation (e.g. CrossRef DOIs for articles)
- Data (e.g. DataCite DOIs for datasets)
- Data extractions? Individual data points?

Provenance

- Delineating the source of the data to help with
  - determining quality
  - integration strategies
  - preservation planning

- Includes citation-type metadata, but also research methodology, instrumentation, protocols, etc.
  - same as the research article?
  - how to build into data archiving systems?

Metadata

- Includes “who/what/when/where”
  (“how” is provenance)
- Includes data structure and semantics
- *Also must be legally and technically interoperable*
- Uncertain copyrightability, waivers are best

Data Structure

- Data integration at large scale is difficult, labor-intensive
- Many disciplines have proprietary or custom data structures (e.g. FITS in astronomy)
- LOD (Web) standards work better, e.g. RDF and XML
- Massive effort to define standard ontologies, convert legacy data and tools

Data Accessories

- Data is useless without documentation and metadata
- Also useless without software to create/process/analyze/visualize
- Governance extends beyond data to the tools needed to use it, should be
  - open source
  - discoverable and documented
  - archived and preserved

Conclusion

- Data Governance is an institutional issue

- Libraries are well-positioned to help
  - Experience with licensing content (OA and closed)
  - Experience with privacy/confidentiality (archival collections, patron confidentiality)
  - Experience with standards and large-scale interoperability (e.g. MARC)
  - User advocates

Questions

and Poll

Thank you!