Proposed Re-Charter, RDA-CODATA Interest Group on Legal Interoperability of Research Data (Revision Sept. 2019)

Name of Proposed Interest Group:

RDA-CODATA Interest Group on Legal Interoperability of Research Data

Proposed Co-Chairs

Christof Bruch, Bob Chen, Gail Clement, Simon Hodson

I. INTRODUCTION

The Research Data Alliance – CODATA Interest Group on Legal Interoperability of Research Data (RDA-CODATA IG) released its output, *Legal Interoperability of Research Data: Principles and Guidelines* to explain how open research data should be made widely available in order to achieve legal interoperability in the ideal case (Uhlir et al, 2016, **https://doi.org/10.5281/zenodo.162241**). The Interest Group's recommendation to release data with the most open, legally-sound mechanisms available (*ibid*., 'Principle 1' and 'Guideline 1c') derive from extensive engagement with stakeholder groups; analysis and discussion of foundational case studies; and serious consideration of the vision to achieve "open data without barriers" reflected in the legal, policy, and research literatures.

Yet since its release in Fall 2016, the *Principles and Guidelines* have elicited feedback and queries from stakeholders who feel the IG's recommendations do not accommodate their needs (please refer to the following section 'Four Focal Points for IG Consideration' for further explication of stakeholder concerns).

The Research Data Alliance – CODATA Interest Group on Legal Interoperability of Research Data therefore proposes to renew its charter. The aim of the renewal is to explore measures which can help bridge the gap between the 'open by default' premise reflected in the *Principles and Guidelines* and the real-world needs of many researchers / research institutions for a more balanced approach to data access and reuse. Under a renewed charter, we will serve as a platform to consider and resolve extant issues around the implementation of the Group's L*egal Interoperability Of Research Data: Principles And Implementation Guidelines (Uhlir et al, 2016)*. To advance the RDA Mission, we believe that the human and technical bridges necessary to improve data sharing cannot be built without a better understanding and implementation of legal interoperability practices.

Objectives

The proposed objectives of a renewed Interest Group are to:

1. To document use cases where communities of practice report barriers to implementing the *Principles and Guidelines*
2. To explore possible solutions to accommodate stakeholders needs.
3. To identify opportunities to create future working groups that will address specific issues concerning legal interoperability, and assess community priorities as to the most important activity to pursue in 2020
4. To prepare a Case Statement in support of one or more Working Groups to address the barriers to implementation and their solutions

II. BACKGROUND

A. Four Focal Points for IG Consideration

The implementation concerns surfacing in post-2016 dialog with stakeholders fall into four general categories:

1. Control of downstream use (variant degrees of openness)

   a. Diminished Authorial Control: Some stakeholder groups interested in implementing the *Principles and Guidelines* have expressed concerns with diminished authorial control in relation to rights waivers (e.g., dedicating their works to the public domain) and liberal attribution-only licenses (eg allowing all forms of modification and reuse, as long as attribution is provided).

   b. Need for more explicit Usage Terms: Specific examples of additional controls to be accommodated include such provisions as:

      i. The need for liability disclaimers for inaccuracies, lack of timeliness, or incompleteness in the released data and to limit its liability in the event a third party is somehow damaged due to the use of the information;

      ii. The licensors' desire to be notified of the reuse of his/her data;

      iii. The preference to exclude certain applications of the data for purposes not supported by the data licensor. For example, the Licensor may wish to limit the rights granted in the information to use in certain geographic region(s) or are limited for the development of products and services for certain industries/markets.

      iv. The need to address the duration or term of the agreement, what events will give one or both of the parties cause to terminate the license agreement early, and the parties' obligations upon termination.

v. The need to outline the responsibilities of both parties with respect to data protection/privacy laws. For example, a Licensee might request that the Licensor state that the data was collected in accordance with applicable law and that all necessary consents have been obtained in order. Additionally, the Licensor might require the Licensee to promise to comply with all applicable privacy/data protection laws with respect to its use of the geospatial information.

1. **Licensing practices for Multipart Objects: Need for a licensing scheme that can apply to multipart, heterogeneous objects packaged within a given data release or submission information package (SIP).**

   a. In response to calls for greater transparency and reusability in E-Science, researchers increasingly strive to produce rich representations of their findings that comprise not only the dataset itself, but also the code, protocols, and notebooks that facilitate downstream replication and reuse. Different types of subject matter (e.g. code, content or data) necessitate differences in licensing. Licenses designed for one type of subject matter — as CC licenses were designed for content, and F/OSS licenses for code — aren’t always best suited to licensing another type of subject matter. Each of these constituent parts may call for licensing regimes that differ from the other parts. The complexity may increase as released products are created using information from a multitude of sources, for which each with unique, sometimes conflicting, licensing terms.

   b. Additionally, the released object as a whole, whether shared as a zip archive, a containerized Jupyter Notebook, or compiled R markdown website, requires a metadata record for various purposes (such as registration of an identifier; indexing by search and retrieval services; etc.) that provides a clear rights statement and license that is comprehensible by humans and machines.

2. **Metadata Integration: Develop standardized human and machine-readable rights statements as part of standard metadata.**

   a. Develop a recommendation on importance of rights statements in metadata that are human and machine readable. Advocate for rights-related metadata elements to be made mandatory in common schema implemented in data curation (such as DataCite core schema; ISO 19115-1:2014 metadata standard for Geographic Information; and analogous and equivalent schema). Surface and articulate lessons learned from the cultural heritage community and existing work under Europeana etc.

   b. Explore development of an ontology of rights statements appropriate for research data objects that are comprehensible by machines and humans.

   c. Develop recommendations on how to utilize, for legal interoperability, widely-used existing metadata standards (eg DataCite Scheme; ISO 19115-1:2014 metadata standard for Geographic Information; edm:rights field of the

Europeana Data Model; and others). These recommendations would include concrete examples on how to implement the metadata with respect to legal interoperability concerns, and will address rights statements, data licenses, software licenses that address data, license stacking, etc.

3. **Monitor, evaluate, and recommend technical means to communicate information concerning permissions/limitation concerning reuse in a machine actionable manner. What rights information do machines need to help humans determine legal fitness for use?** (ownership; rights statement; licensing terms and conditions)

    a. Are there Machine actionable, human understandable rights expression languages that could be adapted for research data?

    b. Are there emerging technologies, such as blockchain, that could be used to effectively handle legal statements made in rights expression language?

User scenario(s) or use case(s) the IG wishes to address

Discussions with stakeholders have identified approaches to improve community practices for achieving legal interoperability of research data. These approaches address needs for control of downstream use, needs for complex licensing schemes for rich data objects, needs for human and machine-readable rights statements for metadata, and needs for communicating restrictions and limitations in a machine-actionable manner.

The following illustrative case studies collected, documented, and shared by the Interest Group represent barriers to implementation of the Principles and Guidelines

- *The CaltechData repository curates and disseminates atmospheric chemistry* data from the international Total Carbon Column Observing Network (http://www.tccon.caltech.edu/), but is not able to enforce the application of standard open licenses due to researcher requirements to be notified of reuse and modification of the data (Agosti, Clement, Egloff, Morrell, 2017, *One Repository, Two Implementations, and a World of Legal Interoperability Opportunities and Challenges,* RDA 9th Plenary, Barcelona, Spain, April 7, 2017, http://doi.org/zenodo.439627; Clement and Morrell, *Data Licensing Preferences As a Barrier or Bridge to FAIR: The Case of CaltechData*, Drexel-CODATA FAIR and Responsible Research Data Management (FAIR-RRDM) Workshop 2019, Phildalphia, April 1, 2019)
- *The Reusable Data Project in the Biomedicine communit*y has devised a rubric and scorecard for measuring the level of open licensing and legal interoperability in publicly funded datasets. Findings presented at RDA plenary 10 indicate that approximately half of the examined datasets demonstrate ‘poor’ open licensing practices. (Haendel et al, 2017, *Reusable data for biomedicine: a data licensing odyssey*, RDA Plenary 10, Montreal, Canada, September 20, 2017). https://www.slideshare.net/mhaendel/reusable-data-for-biomedicine-a-data-licensing-odyssey

- Researchers in the open science community have been discussing data licensing for multipart research compendia (R Notebooks) and currently apply a diverse set of practices while admitting to legal uncertainty (https://discuss.ropensci.org/t/licensing-for-research-compendia/1581 (Boettig, posting to ROpenSci Community February 16, 2019.) The R Open Science community has proposed a model for licensing these compound objects based on the work of reproducibility expert Victoria Stodden (Stodden, Victoria, Enabling Reproducible Research: Open Licensing for Scientific Innovation (March 3, 2009). International Journal of Communications Law and Policy, Forthcoming. Available at SSRN: https://ssrn.com/abstract=1362040. The IG will review and analyze Stoden's recommended practices and will continue to outreach to the open science community to devise a formal endorsement of the Stodden model or to adapt it based on the IG's findings.
- GEOFON - A project providing access to seismic data globally collected by researchers from many research organizations and to corresponding analytical software. The key challenge to open data in this project is the risk that a competitor may copy all data plus software and start a similar service elsewhere. (Quinteros et al, 2018, *Selecting an appropriate License for Open Data. The GEOFON Experience*, RDA Plenary 11, Berlin, Germany, March 20, 2018) https://www.rd-alliance.org/sites/default/files/2018-03-23_RDA_IG-Legal-Interoperability_P11_Quinteros.pdf)
- PresQT (https://presqt.crc.nd.edu/)

III. PROPOSED WORK PLAN

A. Participation

The Interest Group has been actively sharing the *Principles and Guidelines* with interested researchers, policy makers, librarians, curators, data managers and data stewards through numerous venues. Examples of engagements include:

1. Teaching data licensing at the CODATA-RDA Summer School and the Force11 Scholarly Communications Institute in 2018 and 2019;
2. Answering implementation questions with interested groups of researchers and research data managers such as the NIH-Biomed community and the Belmont Forum
3. Sharing the Principles and Guidelines with interested data sharing and data curation communities, such as:
   a. the OceanBestPractices repository (https://www.oceanbestpractices.net/handle/11329/295);
   b. OpenAire (https://www.slideshare.net/OpenAIRE_eu/legal-interoperability-of-research-data-principles-and-implementation-guidelines-christoph-bruch-helmholtz-open-science-coordination-office-rdacodata-legal-interoperability-interest-group);

c. Force11 (https://www.force11.org/article/legal-interoperability-research-data-principles-and-implementation-guidelines);

d. the International Association of University Libraries (https://www.iatul.org/about/news/rda-codata-legal-interoperability-research-data-principles-and-implementation-guidelines);

e. The Food and Agriculture Organization (FA) http://aims.fao.org/activity/blog/rda-codata-legal-interoperability-research-data-principles-and-implementation

f. The Australian National Data Service https://www.ands.org.au/news-and-events/latest-news/news/legal-interoperability-of-research-data-guidelines-published

g. Researcher communities who are actively creating compound data objects, such as the ROpenSci community; and the model organism community particularly the editorial board of their new *Micropublication: Biology* data journal

h. Creative Commons. There will be a session led by a co-chair as part of the CC Summit 2019 where he will report on the work of the IG, the feedback it got in its P13 session. Based on this he will seek input from the CC community.

Stakeholder groups outside RDA

Supporting the development of standards and practices relating to rights metadata and rights statements necessitates engagement with a broad range of stakeholder groups outside of the Research Data Alliance. Groups identified for outreach and collaboration include:

- Cultural Heritage Community which has successfully developed a taxonomy of rights statement for use with objects in museums, libraries, and archives (https://rightsstatements.org/en/about.html)
- DOI Registration Agencies (DataCite, CrossRef) and their member organizations who register data objects as first class citizens of the research record
- Creative Commons

Coordination with RDA-CODATA groups

Data licensing and legal status of research data are core concerns for this IG, but also may be relevant to other data policy and data ethics groups within RDA. Particular groups we intend to reach out to for possible input and collaboration include:

- FAIR Data Maturity Model WG
- IG for Surveying Open Data Practices
- Research Funders and Stakeholders on Open Research and Data Management Policies and Practices IG
- Education and Training on handling of research data IG
- WDS/RDA Assessment of Data Fitness for Use WG
- Blockchain Applications in Health WG (Their knowledge concerning blockchain may be of interest.)

B. Mechanism

The Interest Group will pursue the proposed scope of work via weekly/fortnightly virtual meetings and sessions at RDA Plenary. We also propose to organize a face-to-face meeting in 2019-2020 with a number of options being explored (colocated/pre-RDA events, Force2019, CODATA meetings, or a free standing workshop at a research institution with membership in RDA or CODATA). Finally we will continue to collect and share case studies, published literature, and open educational resources via our shared Zotero database online: https://www.zotero.org/groups/1757514/legalinteropdata

C. Outcomes

- Establishing and promoting Data Licensing guidelines for multi-part data objects. This output would complement the *Principles and Guidelines* with an informational resource and model for how to license the heterogeneous parts of a compound digital object, as well as the object as a whole
- Case Statement for a Working Group
    - Examples of topics under discussion by the IG that may contribute to the Case Statement for a new Working Group:
        - Beyond Creative Commons: An analysis of which licences accommodate the needs of concerned data producers/owners in order to find out if more/new licenses are needed. This may connect to the efforts of Jane Greenberg et al at Drexel and their NSF funded project “A Licensing Model and Ecosystem for Data Sharing” (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1636788&HistoricalAwards=false)
- Recommended practices for sharing and using data. The objective will be for the recommended practices to be matched with the FAIR data principles and adopted by research communities