DOI Meeting Summary (aka ImageCite)

Raw Notes

Meeting Dates: May 4-5, 2017

Attendees:

Dan Marcus, WashU, XNAT, HCP, CNDA
Tim Olsen, Radiologics, XNAT
David Kennedy, UMass Med, NITRC, ReproNim
Christian Haselgrove, UMass Med, NITRC, ReproNim
Al Crowley, TCG, NITRC, ReproNim
Samir Das, MNI, LORIS
Ross Kelly, MRN, COINS
Fred Prior, U. Arkansas, TCIA
Ashish Sharma, Emory U, TCIA
Visakh Muraleedharan, INCF, Center TBI - Remote
Dan Hall, NIMH, NDA - Remote

Contents of this Report

Meeting Objective

To plan an implementation of an imaging a unique identification (DOI) system across a set of imaging data sharing host providers. We will review a proposed scheme of identification and attribution (Honor, *et al.*[1]) in terms of its suitability and applicability to the needs of the community and feasibility of implementation in the various image hosting systems. Specific objectives were to see if consensus about assignment of unique identifiers to 'collection'[2] and 'series'[3] level data objects could be reached. If consensus on such an identifier objective is reached, discussion about implementation details, identification of barriers, and elaboration of ways to reduce these barriers would be discussed. Outcomes will be a conceptual plan of what can be accomplished by the repositories in the short term under current, existing development cycles (Phase 0); identification of additional future objectives that can be achieved with additional, potentially coordinated, support; and a 'white paper' document that documents this identifier vision that can be circulated to additional resource providers in order to attempt to gain even wider adoption.

The Scheme

We pursue the twin objectives of enhancing credit for data sharing and providing an improved documentation system for the identification of data used in publications in order to support enhanced reproducibility and transparency of published research. We propose that making progress in this domain is dependent on identifying:

- How to uniquely identify data, and to what level of granularity;
- How to adequately cite datasets, and make those citations visible, quantifiable and reusable; and
- How to ensure the chain of proper attribution and credit is maintained, even in the case when new datasets are created from multiple existing sources.

Granularity: We posit that identifiers should be assigned to the "finest independently sharable object". In the imaging domain, we contend that the image 'series' is the basic shared element. A generic query for imaging data, in the context of identifying data suitable for reanalysis or meta analysis is typically of the form: Give me the 'images' that meet a specified set of imaging and subject characteristics (i.e. "find the resting-state images for Autism patients, aged 20-25, acquired at 3T").

Collections: We posit that basic shared element is always associated with at least one higher order grouping that identifies the originator of the basic shared element. In practice, any shared 'series' can be associated with a 'project' that represents a collection of shared data associated with a specific set of investigators, protocol, funder, etc. In addition to a collection representing the origination of the data, the basic shared element can be grouped into new collections based upon search criteria or some other method to aggregate new datasets from shared data for subsequent use. These new collections to which a basic shared element is associated should also be identified in a fashion that identifies who performed the generation of the new collection, the criteria used, and other features of the collection.

---

[1] Honor LB, Haselgrove C, Frazier JA, Kennedy DN. Data Citation in Neuroimaging: Proposed Best Practices for Data Identification and Attribution. Front Neuroinform. 2016 Aug 12;10:34. doi:10.3389/fninf.2016.00034.

[2] One of the first orders of future business will be to define "collection" in a formal sense.

[3] One of the first orders of future business will be to define "series" in a formal sense.

Identifier Format: The digital object identifier (DOI) and associated DataCite metadata scheme is posited as an appropriate identifier.

Consensus

After some deliberations, the participants in attendance agreed about the importance of 'collection'[4] and 'series' level identifiers in order to support the twin objectives of facilitating credit for data sharers and promoting a more reproducible overall data descriptor for use in scientific publications and as a way to track credit and usage of data elements, particularly as data gets aggregated across more independent and data hosting facilities.

Implementation

Given the above consensus, the discussions then turned to details and practicalities of implementation. Specific implementation issues relate to how to interpret, in practice, the meaning of the DataCite metadata fields; issuance of DOIs to the identified series and collections; and support of 'permanent' landing pages that the DOIs resolve to. We reviewed an example implementation that took a first cut at addressing many of these issues in the form of the prototype Image Attribution Framework (IAF) system that was introduced as part of the Honor, *et al.* 2016 publication[5]. In these examples, issues related to standardization of the terminologies, methods for constructing relationships between DOI, and conventions that will be to be commonly adopted across all data providers were elucidated.

After these brainstorming discussions, the following implementation plan was proposed:

Phase -1: Logistics

- Talk to California Digital Library (CDL), the DOI provider for most of the current image hosting systems, about the feasibility of this proposed proliferation of DOIs
- Create a group github account – ImageCite (https://github.com/imagecite, http://www.imagecite.org/)

Phase 0: Low hanging fruit

- Create DOIs for acquired imaging data (DICOM Series equivalent) for our various platforms. Modalities to be included: MRI data and PET data.
- Create 'declared' collection (project, study) DOIs to which the Imaging data can be associated.
    - Standardization of DOI metadata and landing pages
    - Differentiate human/phantom/animal data
    - Establishing the conventions for: the relationships between DOIs (HasPart, Is DerivedFrom, etc.); versioning (DOI for releases as well); format (do different formats of the same data (i.e. DICOM, NIfTI, etc.) have different DOIs?).

It was agreed for most providers that a Phase 0 implementation could probably be accomplished under the current funding/development plans of the existing repositories within this calendar year.

---

[4] Note that a number of existing systems already generate project- or collection-level DOIs, including TCIA, NDA, INDI, etc.

[5] iaf.virtualbrain.org

Phase 1: Important tasks for success

- Create DOIs for additional data elements: processed data; provenance files; code/analysis scripts
- Develop tools to compute download stats (and ‘h’-like altmetrics)
- Create visualization tools to document data use and reuse (data use graph)
- Develop a ‘centralized facility’ to support DOI use and aggregation as a service for the various data hosts. This will help facilitate annotation of search results that arise from content contained in multiple repositories.

It was agreed that these Phase 1 activities would likely require a new funding source and a broad collaboration between the data hosting providers. It is hoped that ‘grass-roots’ accomplishment of Phase 0 would provide the necessary ‘preliminary data’ for a future collaborative proposal.

Phase 2: Additional Future Tasks

- More data types: Electrophysiology data (MEG, EEG); Atlases/templates; Behavioural/Assessment files; Genetics data; Biosamples; Other imaging modalities.

What Next?

A number of the data hosting systems present provided informal commitment to accomplishing the Phase 0 objectives in the near future.

- XNAT suggested that a plug-in architecture could be developed that supported the generation of DOIs and landing pages for series and projects that could then be generally used in other XNAT implementations, such as DPUK, NITRC-IR, HCP, etc. on the August 2017 time frame.
- Loris suggested that issuance of project and series DOIs could be also accomplished in their system on the September 2017 time frame.

A ‘whitepaper’ summary of the meeting (this document) will be circulated amongst the participants to finalise the representation of the consensus and discussions that occurred, and then the document will be shared broadly with the greater community of data hosting facilities (LONI IDA, OpenfMRI, INDI, DataMed, DataVerse, etc.) and interested organizations (INCF, OHBM, Force11, etc.) to potentially develop a broader set of parties supporting the overall consensus and implementation concepts proposed herein.

Summary

The meeting was successful at bringing a set of data host providers to discuss the issues of data identification for support of reproducibility and tracking of data credit. A better understanding of the problem, as well as approaches that the community itself can undertake to move the field forward was obtained for the participants. Specific tacks can be accomplished in the normal course of hosting systems operations in the short term which it is hoped can be leveraged to generate future financial support for a broader set of image citation and

tracking functions that will be needed to complete the vision of a data ecosystem that is completely FAIR (findable, accessible, interoperable and reusable)[6] and transparent.

[6] Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3 (March). Nature Publishing Group: 160018.