Session Organizers:
Shun-ling Chen, Tyng-Ruey Chuang, and Mike Linksvayer

Session Title:

Mass Collaboration Data Projects and Policies

Session Proposal:

Collaborations on content generation and reuse can both produce and consume various types of resources, including, among others, text, code, dataset, public domain repository, activity log and community record. These practices are increasingly multi-modal and involve many actors. How do people frame jointly created resources to allow for reuse and to encourage collaboration? Often there is an anxiety in balancing freedom (about content reuse) and fairness (no free rider). For many projects that produce copyrightable works, public licenses such as the GNU GPL and the CC BY-SA license have been instrumental in maintaining a boundary of collaboration and in regulating the use of joint output.

For data-intensive projects, we are in an early stage of framing the various data sharing issues. Rights about datasets are not homogeneous across jurisdiction boundaries, and there exist different practices. Some projects consider their datasets to be in the public domain. Some projects require attribution or citation but otherwise their datasets are free to use. Some insist on data integrity while others may impose share-alike conditions on data reuse. In cases where public licenses are used to release datasets, some projects may at the same time ask for contributor agreements and/or specify terms of service. It seems there is a large space for discussions, especially on the the practical means (legal tools or not) for governing the production and reuse of data for collaborative projects.

However, mass collaboration around creating and curating data is not a new phenomena, ranging from Project Scoresheet to MusicBrainz to Freebase to OpenStreetMap to DBpedia and more. Furthermore, "Linked Data" is making it more possible than ever for more latent forms of mass collaboration around data to occur, and for mass collaboration projects to ingest, curate, improve, etc external datasets.

By organizing this panel, we hope to elicit discussions on data sharing issues, and to help develop conceptual tools for data-intensive projects. We propose to do the following:

* a background discussion about subject matter i.e. what type of work is copyrightable. We will discuss how the language of "authors" and their "writings" in the copyright clause in the US Constitution have expanded to cover various types of work but at the same time leaving out others, what the situation is elsewhere in the world, and how various public licenses address the issue.

* an informal survey of various data communities on their data sharing practices: Observations on how they maintain the boundaries of collaborations, what tools are used to constrain/encourage data sharing, and what actions they take in face of non-conformity. Also how their objectives differ, e.g., to replace a heretofore proprietary dataset, to create a

new dataset for a particular project, field, or the universe, to exploit datasets created as a side effect of mass collaboration, and others.

* some initial thoughts on sharing content collections of a mixture nature (including e.g. datasets and copyrightable works): What restrictions ones may reasonably expect on imposing on users? Why and how?

* a review of some of the practices and tools in collaborative data collection, data extraction from content collections, and data aggregating etc. What are to be expected of such practices, and what are to be built into such tools to support data sharing and to encourage collaboration?