Title: Navigating through the Panoply of Provenance Metadata Standards

Organizers: Rhiannon Bettivia[1], Jessica Yi-Yun Cheng[2], Michael Robert Gryk[2]
[1]Simmons University, Boston, United States of America
[2]University of Illinois, Urbana-Champaign, United States of America

Abstract: The following provenance models and metadata standards will be discussed and evaluated using real-world research data provided by the organizers. Emphasis will be placed on highlighting the strengths and capabilities of each model, as well as shortcomings of any individual model which are handled by one or more of the others. The morning session will cover the models while the afternoon session will provide hands-on cross-walking exercises to explore the modeling differences in greater depth.

- PREMIS: This is an international metadata standard developed by Library of Congress to support digital curation and preservation. https://www.loc.gov/standards/premis/
- PROV: PROV is a family of models and standards proposed by the W3C. It is used to document provenance information of data and digital objects. https://www.w3.org/TR/prov-overview/
- ProvONE: This is an extension model of PROV which includes concepts and attributes for specifying workflows and data products produced by their execution. https://purl.dataone.org/provone-v1-dev

Description

Purpose and Intended Audience: Provenance is ubiquitous in information science. We anticipate this workshop will appeal to any interested in workflows, provenance, and digital preservation. In addition, the exercises in metadata modeling and crosswalking will be of interest to library and information scientists in general.

Proposed Format: The organizers of this workshop appreciate the candor of iConference in designating the online format of the proceedings from the onset. As such, this workshop has been designed specifically to work in a distance, synchronous learning environment, combining the iConference meeting platform with a virtual collaborative workspace and GoogleDrive for key anchor links and materials. The instructional flow begins with participant introductions where background and motivations are shared. Next are a series of 3 short lectures covering the 3 provenance metadata models (PROV, PROV-ONE, and PREMIS). Each lecture includes 'hands-on' exercises for the participants to practice working with each model. These exercises will take place in GoogleDocs and Oxygen, with individual worksheets and group folders offering participants the opportunity to experiment with concepts and view and learn from others in a gallery format. After the lunch break, there is an extensive group session in which the participants work on a provided dataset - recording the provenance of a Twitter thread. We plan to utilize the World Café method in which the participants will be broken in to small, virtual groups using breakout sessions. The participants will have access to a virtual shared whiteboard in which they will tackle micro-tasks. In

10-15 minute intervals, the participants will migrate to different groups, always leaving one member behind to explain the progress to the incoming participants. First, the participants will model the data with both PROV and with PREMIS. Numerous small groups will all work on different aspects of the models. When completed, all participants will rejoin the main workshop space to discuss the process of metadata making. During this discussion, one of the moderators will organize the products of the participant work in preparation for the next activity. Following the large discussion, the participants will return to their small groups and the virtual workspaces to crosswalk the metadata records to illustrate the similarities and differences between the models. Finally, all participants will come together and each micro-task group will share their experiences and findings. The day concludes by asking participants to share outstanding questions for future workshops and to share any changes to future metadata plans they may have garnered as a result of interacting with colleagues during the workshop.

Engagement: Prior to the workshop, participants will be asked to share their past experiences and future plans for provenance metadata. A website will be set up for the workshop, both to advertise the content and process and to share materials relevant to the day's activities. This latter includes the special software used during the workshop, including links and tutorials for the shared workspace platform and for the metadata editing software. Participants will be encouraged to spend 15-30 minutes prior to the workshop familiarizing themselves with the specified platforms and software.

Goals or Outcomes: As a result of attending this session, participants will:

- Be able to identify, classify and distinguish the top level Provenance entities: "Objects", "Agents" and "Activities".
- Be able to properly use provenance relations to connect the entities and create a provenance record.
- Be able to make basic records using a variety of provenance metadata models.
- Gain an understanding of the differences between PROV, PROV-ONE, and PREMIS in order to make informed decisions on their use.
- Gain an understanding of the uses of provenance metadata in different domains and sectors and within existing workflows at their institutions.

Relevance to the iConference: The focus of this workshop is both on the important topic of provenance and established standards for documenting provenance, as well on the common dilemma faced by information professionals: which standard should I choose when there are multiple standards available to me? The approach of this workshop is to cover the various standards with short lectures and simple exercises followed by group activities in which the standards are enabled to engage with real-world data, inspection and group discussion. The organizers previously gave a successful workshop at the 15th International Digital Curation Conference (February 2020, Dublin, Ireland); this proposal builds on findings and feedback from this previous workshop in a format designed specifically for a virtual conference environment.

Duration: full-day event

Draft Schedule:

TIME	ACTIVITY
8:30 – 9:00	Registration and pre-workshop survey
9:00 – 9:30	Opening remarks & introductions
9:30 – 10:30	Provenance models/standards overview
10:30 – 11:00	Coffee break
11:00 – 12:30	Provenance models/standards continued
12:30-13:30	Lunch break
13:30-14:00	Small group discussions on how these models/ metadata standards can be applied to data curation
14:00-15:30	Theory & Practice: Virtual Whiteboard Activities Modeling Twitter dataset using PREMIS & PROV
15:30-16:00	Coffee/tea break
16:00-17:00	Theory & Practice: Virtual Whiteboard Activities continued Crosswalking between the models to illustrate differences between the models. Closing remarks & post-workshop survey

Attendance: Preferred Maximum: 40 participants

Special Requirements: None