Modeling Tweets in Compliance with the Portland Common Data Model

Martin Klein
University of California Los Angeles
Research Library
Los Angeles, CA USA
martinklein@library.ucla.edu

Kevin S. Clarke
University of California Los Angeles
Research Library
Los Angeles, CA USA
ksclarke@library.ucla.edu

ABSTRACT

The ingest of non-traditional digital library collections into a linked data-based institutional repository for archival and presentation purposes is challenging on many levels. We propose a model for Twitter data that is compliant with the Portland Common Data Model. Based on this model, we can derive a linked data serialization and perform the ingest into our preservation repository.

General Terms

Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Twitter, Portland Common Data Model, Fedora

1. INTRODUCTION

Digital library programs increasingly face the challenge of incorporating non-traditional collections into their preservation and presentation workflow. Examples of such collections for the UCLA library are video and text (transcripts, closed captions, on-screen text) from daily captured TV broadcast news, crawled and archived web pages relevant to particular topics, and social media content such as tweets, which are also collected on a per-topic basis.

Within UCLA's International Digitizing Ephemera Project[1], the Research Library is developing such non-traditional collections around the Egyptian Revolution in 2009 and the Iranian Green Movement in 2011. Aside from thousands of digital images, cell phone videos, and scanned flyers, the collections also contain social media content. In particular, the library has a dataset that consists of more than $400,000$ tweets from about $50,000$ distinct users on the topic of the Egyptian Revolution. These tweets are special in the sense that they all were sent from within a 200-mile radius surrounding the capitol city of Cairo and so potentially reflect the voices of activists on the ground rather than trained journalists from international media channels. It is our intention to incorporate these tweets into UCLA's library preservation and presentation framework. The underlying institutional repository is Fedora and the library is in the process of transitioning to Fedora version 4, which is based on the Linked Data Platform[2]. For collecting tweets we are using the open source software Social Feed Manager[3]. The tool obtains tweets from the Twitter API[4] in JSON format. The Portland Common Data Model (PCDM)[5] has recently gained a lot of traction in the community as a data model to describe resources. This description provides the basis for the (RDF) serialization and hence conveniently bridges the gap between an arbitrary resource (a tweet) and the ingest into an institutional repository (Fedora 4).

The contribution of this paper is a first approach of modeling tweets in compliance with the PCDM. We describe our model, the characteristics of its main components, and all their relationships (in an RDF sense). We are reporting on a work in progress and hence are actively seeking feedback from the community to help stabilize this model.

2. THE PORTLAND COMMON DATA MODEL

For a better understanding of the here presented model, we briefly summarize the three for us relevant components of the PCDM. In the PCDM, intellectual entities (works, digital objects) are modeled as objects. An object can have descriptive and access metadata associated with it, it can contain files, and even other objects. A group of resources (objects) is modeled as a collection. Collections can also have descriptive and access metadata and it has a link to all objects it aggregates. Objects and collections are per se a unordered sets but for use cases where the order matters, a proxy class can be used that establishes order via links and proper IANA relation types such as first, last, next, and previous. The bitstream (sequence of binary data) of a resource is modeled as a file. A file can be described by accompanying metadata such as size, content type, and provenance information.

[1] http://digital.library.ucla.edu/dep/

[2] http://www.w3.org/TR/ldp/

[3] http://social-feed-manager.readthedocs.org/

[4] https://dev.twitter.com/overview/api

[5] https://github.com/duraspace/pcdm/wiki

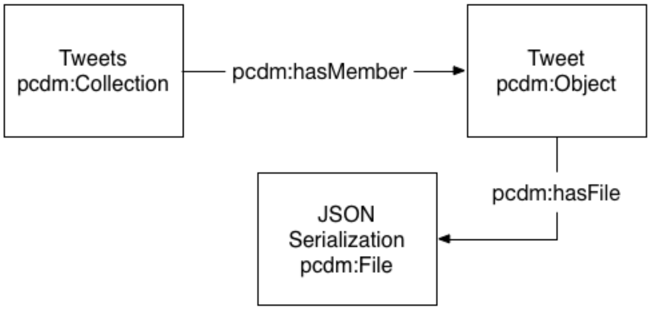

Figure 1: Collection, object, and file in the PCDM model

3. MODELING TWEETS

The diagram in Figure 1 depicts a high-level overview of our model. As we are capturing tweets by topic (a natural catastrophe, a political event) or from individual user accounts (UCLA athletics, a student organization), it makes intuitive sense that each tweet belongs to a collection, the high-level component defined in the PCDM. Each tweet is modeled as a PCDM object. Since these objects are member components of a PCDM collection, the collection links to each of them with the *hasMember* relation type. A tweet in JSON format comprises of a number of key/value pairs with notable examples being ID, text, created_at, and screen_name holding values for the tweet's unique numerical identifier, its textual content, the datetime it was sent, and the user name of its creator (the string following the @ character), respectively. One option for the PCDM would have been to deconstruct all or the for our use cases most relevant key/value pairs and model each of them individually as relationships of the tweet object. However, after consulting with the PCDM community, we decided against this approach and chose instead to model the JSON representation as a file which is linked to from the tweet object with the relation type `hasFile` as seen in 1. The main advantages of this approach is the retained simplicity and flexibility of the model. The simplicity comes from saving what otherwise would be several dozen links from the tweet object to the tweet's ID, text, creation date time, etc. and the flexibility is gained as different use cases can now individually chose which key/value pairs from the JSON serialization to process, for example in a Solr index to facilitate search.

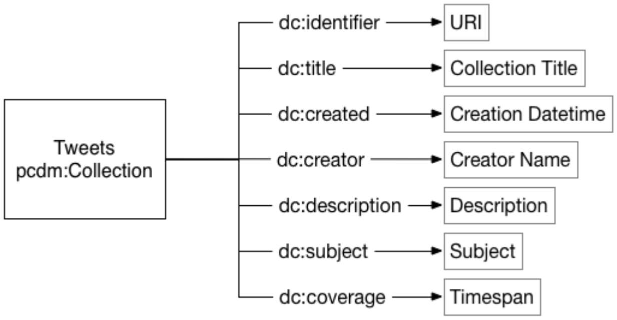

Figure 2: Link relations of the tweets collection

All components in our model have associated descriptive and/or technical metadata, modeled as links with proper relation types. Once the RDF serialization of this model is ingested into Fedora, these data points can, for example, be queried via a SPARQL endpoint. The relations of the tweets collection are shown in Figure 2. Our model contains basic metadata elements for the collection-level such as the collection's title, URI as a unique identifier, subjects, and the timespan encompassing all component tweets. Figure 3 shows that our tweet object has four such links which all reference information directly derived from the tweet itself: its URI, creation datetime, the creator's name, and containing hashtags. This introduces a certain level of redundancy as the data also exists in the JSON file and will from there be indexed in Solr. However, it also enables us to process tweets on the RDF-level, for example, extract all tweets from particular users that contains certain hashtags. The links from the JSON file in our model are depicted in Figure 4. These links point to typical technical metadata of the file itself: its size, mime type, and hash value.

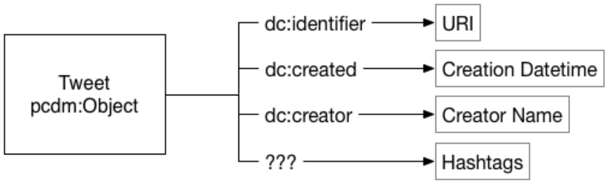

Figure 3: Link relations of the tweet object

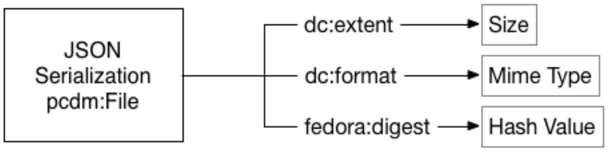

Figure 4: Link relations of the JSON file

Our model is not yet complete. For example, we have not identified a suitable relation type for the link between our tweet object and a hashtag that is contained in the tweet (as seen in Figure 3). Further, we have not yet sufficiently addressed the notion of access-level metadata but we are closely following the community discussion around the WebAccessControl system[6] and will adopt the emerging standard in due time. Also, a detailed discussion of the RDF-based linked data serialization of the model falls outside the scope of this paper.

4. SUMMARY

We introduce a model for tweets in compliance with the Portland Common Data Model in order to facilitate the ingest of such data into institutional repositories like Fedora 4. This model is still a moving target and we are actively seeking feedback on the here presented work. With the the ongoing community discussion and with feedback from the community, including the iPres audience, we are hopeful that we can derive a stable model for tweets in compliance with the PCDM

[6] `http://www.w3.org/wiki/WebAccessControl`