Pratt Institute
School of Information

DADAlytics Project

Stakeholder Meeting - Minutes
11.6.17

In Attendance:

Jeff Rubin, Ilaria della Monica, Cecily Marcus, Rob Hudson, Farris Wahbeh, Giovanni Trambusti, Alexandra Provo, Hannah Sistrunk, Cristina Pattuelli, Matt Miller, Karen Hwang, Eric Toole, Dana Lachenmayer, Sarah Ann Adams, Rachel Egan

---

How would a service like DADAlytics be helpful?
What application would you see from your own experience?

- Tech moves quickly, but it's slow to be adopted in cultural heritage; reclaiming old metadata and how to make that useful?
- Karen: Linked data at the MET. Minting authority files automatically when non-existent. Leverage Wikidata to mint URIs.
- Farris: What about finding aids/ Using this as archival pre-processing of backlog? How would Whitney use these tools to analyze archival backlog and identify relationship with unprocessed collections, integrated into archives management workflow.
- Cecily: Or reassembling diffuse/dispersed archival materials through archivespace records. “Re-stitching”. Use for researchers as discovery layer on top of cross-institutional archival collections.
- Using extraction against a collection, to see "what is or isn’t processed"?
- Alex: Would be interesting to compare work of a human indexer with NER. Or use EPubs as source material
- Cecily: What about institutional repositories?? Enrich metadata of texts to facilitate discovery.
- Ilaria: Berenson Circles - Intellectual society. Use to extract entities from diaries and put together who is who, what venues, when? Use to identify entities across integrated archive and link between photographic collections, documents, library, etc.
- DADAlytics could assist in telling you what isn’t yet minted; minting the entities themselves would be a more detailed work flow

- How do you put collections "back together" by looking at disparately located items?
- Cecily: Identify objects to be incorporated into UMBRA by enriching metadata records with shared entity through comparison, co-referencing.
- Rob: Use NER to further linkify, create more access points to Carnegie Hall Archives Performance History Database. Because Carnegie is a venue, social interaction is important aspect. Or extract carnegie hall ticket price.
- Jeff: Tool would be integrated into digital repository ingest workflow. Also used for processed collections and Hogan Jazz Archives materials & Jazz/NOLA Heritage archive.
- Karen: Would use DADAlytics on processing workflow, ESTN.
- Cecily: Metadata enhancement would be really helpful for research.
- Matt: Used in DPLA ingest process with metadata remediation and normalization, crosswalking.
- Matt: Journalists as users.
- Metadata enrichment to index (i.e. topic) ecosystem
- Indexes vs. full-text search [MM]

Challenges and Issues:

- Karen: How to integrate this with different workflows, collection processing. Would help to be configurable with different sets of tools. How to open metadata that has been contributed to DPLA to enhancement?
- Alex: What about scale? How to automate management of entities, have unsupervised option.
- Karen: Will it be easy enough for community archives without tech team to use? What about a studio partnership with organization like METRO with setup stations and technical support?
- Alex: What about messy, garbage-y data? Topic curation toolkit for stripping formatting from data.
- Alex: Abbreviated names are a problem to be aware of.
- Jeff: How do we embed this technology into contemporary documents rather than just transforming old analog documents into digital content? Into word processor or spreadsheet?
- What about the “nobody people” (as Cristina calls them)?
- How to generate JSON LD for search engine consumption-through Bootstrapping, in order to increase discoverability of collections
- How does this data get fed back into already existing schemas or metadata structures, both cleanly/precisely and automatically? (Jeff Rubins, Cecily Martin, Karen all have a need for this)

- Disambiguation through text distance clustering [AP] for preferred names vs. abbreviations, nicknames, misspelling, etc.
- Karen: How to enable vocabulary selection?
- Karen: how easy will it be for community archives [for example] to use these tools if they don't have dedicated staff/tools?

Design Recommendations

- What other crowdsourcing opportunities are possible?....This is also something we can look at and keep in mind.
- Not just linking to outside authorities like Wikidata and LOC. (Alex Provo)----Identity Management
- Discussion boards and other ways to engage the community as well, when it comes to fine tuning the software.
- Weighting names → TFxIDF [RE]
- Want a tool to recognize the structure of documents (document analysis)……Oral history transcripts have an interviewer and interviewee, diaries have a different structure, correspondences have another structure etc.
- Generating network from back of book indexes, not sure if all the terms are 'appropriate' for Wikidata [ie. one item is 26 words long], want to manage network as a closed system within itself
- Turn key solutions-don’t have to download anything, but can access all the tools at once
- Goal is to continually be documenting and publishing work based on this process, in order to share with the community and to gain feedback along the way
- Karen: what about having a local instance of items on the cloud
- Don't reinvent the wheel. There is a lot out there, but it isn't necessarily usable.
- "want to incorporated it back into the metadata" / "we want to create a crowdsourcing tool"
- Domain agnostic - tools not only for performing arts
- Workflow should include re-enriching collection data before publishing [JR]
- Unicode canonical and compatibility equivalence [RE]
- Leveraging Wikidata to mint URIs [KH]
- Focusing on collection level vs. document level to aid in archival processing workflow [FW]
- Semantic relationships at collection management level → SNAC?

- NER for typographies and semantic meaning (e.g. italics, bold) [AP]

- Karen: Option to create list of names at the outset.
- Cristina: Use dictionaries of names or local authority to define set of entities.
- Farris: OCR component would add a lot of value for potential users.
- Cecily: Linking to authority files is very valuable.
- Cristina: Degree of noise/error can be adjusted depending on needs.
- Rob: Should be web-hosted. IT permission is a barrier in office environment.
- Cecily: Should be behind a login with downloadable results.
- Rob: Linked vocabulary recommendation.
- Cristina: Have baseline relationship - is similar to. Or a number of base options.
- Cristina: Are we targeting people who have already bought into linked data or small communities with little awareness?
- Karen: Make it understandable to the collections manager what the end user will be able to do - creation of a new dataset and researchers are able to explore
- Managing URIs in closed system [AP]

Document Typologies

- Karen: Artist resumes and press releases from events, shows, exhibitions would be interesting, easy to OCR.
- Ilaria: Diaries.
- AP will provide ePub samples
- Document types: Press releases, artist resumes, exhibition invitations
- Document types: Handwritten ledgers, ticket stubs [RE]
- Researchers are interested in price history [RH]

TESTBED

- Edge cases
- Document types/formats
- Use cases
    - Domain agnostic or domain specific?
    - Scalability (# of documents)
- Users
    - Digital humanists
    - GLAM professionals
- Workflows
- Challenges
    - Amount of manual labor required
    - OCR conversion
- Functions

-
- Testbed
- Document types
    - XML/MODS/EAD
    - Plain text + CSV
    - JSON metadata
    - Event programs, press releases, artist resumes
    - ePubs and scholarly monographs
    - Papers from institutional repositories
    - Finding aids
    - Diaries, correspondence
    - Transcripts
    - Works of art
- Usability
    - Extensive documentation, tutorials, etc.
    - Low tech barrier to entry
    - Effective GUI
        - Attractive
        - Clear
        - Concise
        - Familiar
        - Forgiving
        - Responsive
    - Web-based (already in proposal)
        - Permission from IT unnecessary
    - Security & permissions
- Use Cases
    - Domain agnostic and domain specific
    - Scalability (# of documents)
    - Enriching metadata
        - Unknown or undefined entities/descriptions
    - Collection-level analysis
    - Linking collections
    - Unifying entities across archival collections
- Users
    - Archivists
    - Collection managers
    - DH students
    - GLAM community
    - DH community
    - eBook users
    - Researchers
    - Metadata librarians
    - Catalogers
    - Journalists (e.g. BBC, NYTimes)
- Functions
    - Merge/split entities
    - Ability to unlink/uncluster entities
    - Ability to designate document type
    - Match found entities to vocabulary of choice
    - Create local URIs

    - Attach URIs to metadata
    - Attach name authorities to metadata
    - Multilingual
    - Unicode canonical and compatibility equivalence for entity matching
    - JSON export
    -
- Workflows
    - Metadata creation
    - Embedded starting point for subjects, predicates, and objects on at the document-type level
- Challenges
    - Amount of manual labor required
    - Non-text based content (e.g. video, sound)
    - OCR conversion
    - Sustainability
    - Tool vs service → SaaP vs SaaS
    - Compatibility
    - Pre/post processing
    - This might be part of a toolkit