Benefits of LOD for Digitized Special Collections
Advisory Board Meeting

Short URL to this doc: https://goo.gl/HjcJO1

Logistics:

Day/Time: 9:00 - 3:00, March 27, 2017

Venue: Illini Room (19th floor)
Illini Center (200 S. Wacker Drive, Chicago, IL)
Check-in with government-issued photo ID at front desk.

Directions (from Club Quarters):
Walk 4 blocks West on W. Adams
Google map directions: https://goo.gl/lN02E3

Contact me: Ryan Dubnicek, rdubnic2@illinois.edu /

For Remote Attendees:

We'll have a dedicated laptop and conference phone setup for remote attendees, and will share slides from presentations virtually. In order to see slides, please use the link below to join via web browser. You may need to install a plug-in, so please test connecting via this link ahead of the meeting.

Join Online:
https://meet.illinois.edu/rdubnic2/P6VD6QHH

Join by Phone: +1 888 983 3631
Conference ID: 8544261

Attending from Advisory Board:

- In-Person:
  - Tom Teper, Co-Chair (UIUC)
  - Jerry McDonough, Co-Chair (UIUC)
  - Antoine Isaac (Europeana)
  - Francoise Leriche (Univ. of Grenoble)
  - Cecily Marcus (Univ. of Minnesota)
  - Jeff Mixter (OCLC)
  - Chew Chiat Naun (Cornell Univ.)
  - Kevin Page (Oxford e-Research Centre)
  - Doug Reside (NY Public Library)
  - Robert Sanderson (Getty)
- Remotely:
  - Mark Matienzo (Stanford Univ.)
  - Derek Miller (Harvard Univ.)

Attending from Project Team:

- PIs:
  - Tim Cole
  - Myung-Ja Han
  - Caroline Szylowicz
- Staff:
  - Ryan Dubnicek, Project Coordinator
  - Deren Kudeki, Research Programmer
  - Jacob Jett, PhD Research Asst.
  - Joo Ho Lee, PhD Research Asst.
  - Melina Zavala, Master's Research Asst.

Lodging: Club Quarters Central Loop (111 West Adams Street, Chicago, IL 60603)

Internet access:
Club Quarters has free wifi throughout the hotel, and 24-hour accounts are provided by Illini Center. Instructions for creating accounts and getting connected while at the Illini Center will be in the conference room.

Meals during the meeting: we will be providing coffee, tea, fruit and pastries at 8:15 and boxed lunch at 12:30 on 3/27. Thought the day, we will have coffee and tea available, as well.

Optional Sunday Dinner (3/26): 7:30 pm at The Village (71 W. Monroe, Chicago, IL 60603).
Meet in hotel lobby at 7:15 to walk over,Tim, MJ and Caroline will plan to meet attendees in the lobby, and walk over to the restaurant. If you'd like to walk with someone, look for them. Each attendee will be responsible for their own bill.

Agenda

8:15 - Morning refreshments

9:00 - Welcome and introductions

9:15 - Board feedback on work to date (will include a break midway):

- Demo of how LOD has been integrated into our Motley collection search and browse resource (CONTENTdm)

- Mapping Motley and Portraits of Actors Collection metadata to schema.org

- Strategies used to identify and add links to existing descriptions

- Marking up Proust social & literary network of names and mapping Kolb's notes to schema.org

- Summary and preliminary observations from baseline Google analytics data and round of user testing, and plans for Motley testing on new interface this spring

12:00 - Lunch & discussion: other projects we should be familiar/in touch with

1:00 - Work in progress

- Ideas for axes of interest for visualizing and annotating the social network of Marcel Proust
- Ideas for replacing and migrating from CONTENTdm

2:15 - Wrap-up, action items & takeaways

3:00 - Adjourn

Meeting Notes

Actions:

- MJ will add more context for Motley collection's metadata and add it to the white paper
- Should switch issue and volume in hierarchy of K-P bibliography items mapping
    - Also set granularity on date extraction to not include day--stop at month
- Deren and MJ will report out total links found for entities in collections, and then break down on manual vs. automatic finding, and send to Board
-

9:15 - Board feedback on work to date (will include a break midway):

- Demo of how LOD has been integrated into our Motley collection search and browse resource (CONTENTdm)
    - Overview of project goals, deliverables, meeting goals, and potential outcomes
    - Introduction to vision of what a new record item may look like
        - External links, dynamically generated information via RDF in left-hand panel
    - Antoine: not so bad that people use text as identifiers instead of URIs (Shema.org is supposed to be ok with this even without context)
        - Tim: Google Structured Data Tool does read it correctly, but it isn't reading the context, json-ld playground does, which can create a bit of a clash between the two
    - One unique thing about Motley--there are compound objects (example: a sketch and then a full colored drawing). These are treated as one object in CONTENTdm and Motley
    - Doug: have you looked at linking more complex theatrical sources? IBDB, Theatricalia?
        - Yes, we have done some, but we've had to rely on manual collection
        - Doug: they likely won't notice, even if their board says they'd prefer not to have us do it
            - Ryan: we reached out to them about permission for automated discovery, and they said they were interested, but needed to talk to their board. They never got back to project team. Still a possibility, but got a bit lost in the shuffle of other critical work (especially given that we had already put in so much time manually finding and adding links from IBDB and other sources
        - Doug: Playbill Vault is a good source for links, and they'll have no problem with web scraping
        - Cecily: Where was focus of heavy lifting, manual labor?
            - Tim: Translating data to schema? The process of mapping takes as long as you want; debates have occurred, we tried not to
            - Couple of ideas have been thought about. How structured is your data is the beginning line.
                - Proust data, it was difficult to automate the process because of the string formating.
            - Proust: figuring out gender of names in collection was a big hurdle,
            - especially for non-Francophones
            - MJ: this all comes back to the state of your metadata at the beginning. With relatively low detail already in Motley/PoA, it was a challenging process
            - Cecily: thinking about archival processes where there is less and less metadata over time, the worry is that so much will be lost that a project like this could be harder and harder, or possibly impossible, in the future
                - Tim: if you have properly formatted name strings, you can get a lot of the data back, automated, at least to the extent that DBpedia and Wikipedia are useful
                    - Some difficulty with getting links to work between DBpedia and Wikipedia, which was unexpected and interesting

          - We do pull some live data from DBpedia, but we’re using SPARQL instead of their live linking
        - Tim: for things that are not linked data resources, then it’s very difficult
          - Example: Tolstoy, to make sure you have the right one, you need a human eye
        - Doug: one good thing is that we tend to add more data for digital objects than physical objects, which is one thing to keep in mind
          - Tim: we’re finding the data and the impulse used to be to bring it into your metadata, but we’re resisting that to avoid duplication
  - Jeff: Have you thought about creating own local descriptions for names that don’t have VIAF etc. links,
    - Tim: Yes, but not a lot of work on this yet, it is in the grant, though
      - We do want to allow annotation from users as well, partially for this work
      - Management of the data, and who would use it?
    - Naun: Cornell has often discussed doing this as well, but not a ton of progress
    - Jerry: how does the work being done on names relate to SNAC (http://socialarchive.iath.virginia.edu/)?
      - Potential for the Proust names
      - MJ: very interesting thought, but not a ton of work/thought put into this quite yet, but would be interested in making it work
  - Kevin: want to make sure there is provenance data available for annotation and for enrichment-at least versioning of records and external info
  - Antoine: have all automatically gathered links been manually checked?
    - No, some VIAF, but mostly no
    - Also a decision to not take multiple links for someone
- Mapping Motley and Portraits of Actors Collection metadata to schema.org
  - Tim: Intro, a lot of projects working on this; discovery in non-library systems, metadata that was structured different. We looked at frameworks out there, and decided on something different.
  - MJ: with special collections, there is often more lax, or no, standard across all special collections for metadata
    - Rob: have you thought about looking at different ontologies to work around this?
      - Tim: yes, on some other, semi-synergistic projects (mostly HathiTrust Research Center work) Kevin, Jacob, Terhi N-F and others have looked at a number of different ontologies
      - Jacob: a big reason we are using schema.org is because OCLC is using it, and UIUC has developed a workflow for using schema.org
        - Tim: we are happy to consider others, of course
      - Antoine: there has been some work on performing arts at the University Library Frankfurt am Main (http://performing-arts.eu/): that might be a good project to compare work to (they extended an ontology for manuscripts that was an extension of the Europeana data model, also a bit of SPAR) [JPM: see http://pro.europeana.eu/page/edm-for-performing-arts-metadata]
      - Antoine: Schema.org is good for other research questions, would have lost time if went and tested around with aggregation model first
  - Derek: there are two things being covered here, a play and a production, but the schema.org vocabulary discusses yet a third thing, “performances”
    - We’re actually not sure if certain costumes, for instance, are used in certain productions, but Harvard Theatre Collection may have photos of the costumes being used in certain productions
      - Doug: Costumes change based on the actor, so it could quickly change.
      - Doug: Keep in mind photographs are often staged, but not actual production photos (mainly used for promotion)
    - Tim: this effort is worth it, and we are glad someone recognized it!

- Differentiating between the event and the process of getting the event together; stage work distinct from the play
- MJ: originally we were hoping to use theaterEvent in schema.org, but it’s very different than the idea of “performance,” so this was abandoned
  - So maybe sketches created for specific performances have a different better type?
  - Derek: production is an instance of a stage work, so it doesn’t have a cast attached to it, or a crew or design even
    - This raises a bunch of difficult philosophical questions
    - Doug: you could say it’s a business entity though--the production existed for a moment to make something happen financially
    - Derek: Opera as a counter argument - keep the costumes and sets. and bring it up again 15 yrs later with a different story
      - Derek: the production is the instance of the work, and probably the only way to accurately tie the costume or set or design to something intellectual
      - Doug: maybe “run” is more accurate--an unbroken set of performances with the costume or design
      - Kevin: there is an issue here where the more you try to describe the metadata about something, you deeper you fall into your ontology--easiest to model what you know and leave the things you don’t know aside, perhaps to area-specific experts
      - Derek: case specific, and worth being precise about the play
      - Materiality of the thing, leave production ambiguous, for discovery
      - One to one relationships not needed for everything
  - Jerry: Encode the things that you know, leave the ones that you don’t
    - Yet there are some issues with the encoding on stuff we do know--dates of stageWork, for instance, seem to be the date of the first production, not necessarily the right production
      - Antoine: it’s probably best to pilot things as they are, or as they’re modeled, and see how they work out
        - For instance, the idea that visualArtwork is part of stageWork doesn’t make sense, perhaps, until you delve into the project understanding more in detail, where it makes logical sense
        - Jeff: being able to clearly describe the description of the thing, in the reality it’s embedded in--e.g. Is the DH scholar interested in an item’s place in the collection or it’s place in the production/play?
- Kevin: Does taking the discovery perspective compromising accurate description?
- Kevin: Is there a brief description of the heritage of where the metadata for these collections came from?
    - MJ: when Motley was purchased, the descriptions came with the items (where/when it was sketched and for which performances)
      - Documentation came from the images, created by the microfilm company that created the microfilm images
        - MJ: will add more context for Motley collection’s metadata and add it to the white paper
          - Tom: Created by the artist, in operation 1930-1976
        - We don’t know where the physical collection came from.
          - No metadata librarian at UI until 2010 (when MJ was hired). Prior to this, late 90s, early 2000s, digitizing by amateurs

    - Tim: Thought about using other vocabulary, but resisted because schema.org would not recognize it.
    - Antoine: looks like a few categories were added, are these described in a clear way for users?
        - Tim: we've been working on this, and fully implementing and analyzing these additions to schema.org proper is a bit out of scope of the project specifically, but we plan to make sure we draw out that schema.org does seem to require some additions, and we plan to discuss this a bit
        - Materiality of the object is the way to go, and forget about the event that it came from
- MJ: Intro, Motley and PoA, commonalities. Local and customized values in metadata, easy to work together at the same time.
    - Metadata is very special because of the local and customized values.
        - Describes original image, the play (name and perfomance, date)
            - So does not describe one thing, but a lot of different things.
        - Split the data in the different parts that is describing.
    - Jerry: one of the most critical things for the museum community is that the metadata record should apply to just one item (not the image and the actual drawing). There are then issues with the rights of the image and the metadata
        - Take a look at the rights statement for the collections and make sure we're accurate in rights assertion
            - Tim: luckily, UIUC owns the entire collection, the physical and the digital
        - Rob, Antoine, Kevin: CRM might handle this better, then you could build in multiple links for things rather than one
    - Kevin: Precision vs. detail; having the one that is "not wrong," are there other terms that could also work.
        - Antoine: there are some weird things in schema.org for rights and ownership. Schema.org has something called "offer" which often applies to a commercial product, but this collection is not commercial, so that makes it tough
    - Kevin: Different perspectives on the same info; can be incremented and what are the perspective?
        - Scholarly, discovery, archival--are some separate from discoverability, which pervades this project?
        - Jeff: a visualArtwork is the same as the search result page--is this a problem?
            - Tim: we're hoping to partially work around this by minting handles for the objects, which can avoid this problem
            - But we are still trying to make sure our metadata makes sense to search engines, which is a wrinkle
            - Jeff: maybe "discoveryURL" or something that isn't "sameAs"
            - Jerry: maybe a DBpedia link instead of a Wikipedia link, or something
                - 2 reasons why not used; we don't know where the link is until we get it from wikipedia and schema.org does not differentiate from our own records and other people's
    - Rob: make sure that whatever metadata structure is used in schema.org doesn't prohibit porting to CRM or other schemas later (?)
- Marking up Proust social & literary network of names and mapping Kolb's notes to schema.org
    - Intro - Caroline and MJ
    - Presentation of encoding of the name database
        - Visualization of connections between the names could be done through the data
    - Bibliographic and Chronological cards are the two types already digitized and available online
        - Presentation of the tags already in the TEI of each card, and the fields each TEI piece maps to in schema.org
            - Citations capture everything; smaller components went to sub-components
            - Google uses headlines, and doesn't understand name
            - Should switch issue and volume in hierarchy

      - Also: set granularity on date extraction to not include day--stop at month
    - Book and articles are the most common types found
- Chronology to event- mismatch of concept the meaning of event in our collection vs. schema.org's meaning.
  - Chron file is Kolb trying to reconstitute Proust's life, day-to-day, so we understand the word "event" to mean only "something that takes place"
    - Sometimes it's just Proust getting out of bed, or going to a play, and we often know the sequence of these events, but not the specific date
      - It is difficult to automate the labeling of events
  - Derek: information is unclear; not an accurate example
  - Treat chrono cards different? What's the way to bind them?
    - Class of activities in Proust live; there is a way to adhere it to a controlled vocab--"publicationEvent," etc.--scheme
    - Describing the content, reality of the card, the historical activity of the card, not the card itself
      - Antoine: although, going with the way we have, if there was better context, perhaps stats of the conversion, could be okay still. Having a mapping where 1-2% of the items don't work well, but it does for 98% is still ok--just need more documentation to explain circumstance
        - Tim: drills down to the balance between automated vs. manually--if a card can't be processed automatically, perhaps we kick to be manually reviewed
        - Antoine: some of them may need to have metadata added manually--potentially could be a follow-on project for something like this
    - What are you hoping people will discover? Info or cards? Have the cards been criticized for accuracy?
      - Tim: this collection is a really not necessarily a general Proust resource, but a resource about Kolb's perspective on Proust, which is well-regarded. But it needs to be remembered that Kolb did some mediation of data and information himself based on his own judgment
        - Going beyond schema to describe the description- might be interesting.
      - How is the collection going to be discovered?
        - Caroline: this collection is used by Proust scholars to look at his biography, which is based heavily on his correspondence, much of which is not dated, and new letters are added and discovered every day-ish
          - Kolb's work gives a panorama of Parisian life from the 1890s to 1920s, and is rich on that cultural universe.
          - We've found it's used by not just Proust scholars, but many historians of that time period (example: art scholars, people researching Dreyfus affair, etc.)
            - The event is what is to be discovered in this example
              - Says something about what will need to be done to make this happen
          - Provenance of the object; how they came up with it might be interesting to the data, to keep track of changes
  - Doug: A lot of effort seems to be into translation of TEI into schema.org; go back to the beginning, and translate the actual cards into schema.org; start over.
    - Tags inappropriate for this project?
  - Distinguish event vs. Bibliography; they're convertible

- Francoise: remember when the project started in the 90s, the idea was the share all of Kolb's notes to the Proust community, especially to check content and date of letters
  - This chron file is not the only one in the system--there are some in the person files, etc.
    - How do we connect the files to each other, if we think it's relevant? How do we deal with the richness of the files?
      - For instance marriage records, where many were married many times
        - For marriages, we're thinking of covering a decent amount via names (DB with names and unique codes for each person, from the person files)
        - But name data is only with respect to the chron. file, which creates a "vicious" cycle
    - Tim: the chronology is already encoded, but not the person file. So for this project, it's a bit out of scope to connect the cards, but this would be something we'd be interested in for a follow-on project.
    - Tom: 5 or 6 yrs in the future, changing ways of mapping the data; maybe better off starting over?
      - Tim: Trying to translate data n a more structured form, change of perspective might be interesting.
      - TEI-schema.org is not a great conversion, but was chosen because schema.org works well for LOD and the other collections
        - Next projects might work better with this collection with a different schema--that is part of the potential outcomes of this project
      - The discussion of "obsolete" needs to be in relation to a given something--this encoding is obsolete to what?
        - What are you trying to do here and now to provide a new intellectual work--we're not just encoding the cards, we're trying to support intellectual work for what?
        - Derek: what's the linkage on the K-P data? Is the schema.org version doing better than the existing version?
          - Derek: Not really; it's not translating into a new way of looking at the data
          - Goal: you'd find and see the card, but the entities discussed would be linked to--a gossip column mentioned links to the web version or item record for the issue, or to a person, or a historical event
          - Derek: I'd say it's hard to assess this project at all with the current example we have.
        - We're mapping the cards for now into a LOD-potential schema, and then we're trying to build the connections from the cards to other information outside, not quite linking between card sets yet
          - Tim: want to find a mapping to insert and make live to articles in people's' biographies and event documentation
        - Poetry and music- are those being mapped?
          - Might not be a good thing to do; schema.org's may not be appropriate to use for this.
          - Schema.org does not match with the Proust data; some entities lost

    - Better to not have extremes; some schema.org might work for a part of the project, some TEI might work better for other parts
    - Can this be done and should it be done? Should information be findable online?
    - Data for publication-rights, and copyright, taking a look at it is an important part. For original objects, digitized objects, metadata. Rightsstatements.org could be worth having a look at. NB: maybe the project doesn't need to solve everything before it's finished. Flagging the issue for further investigation could be acceptable.
      - How will people exploit the data that will be put out?
      - MJ: Right statement added into our goal too; all metadata is open, so anyone can use any of it
    - Antoine: Languages: it is very important to represent (with language tags) the language of metadata when it is known. It's crucial for aggregators like Europeana
    - Derek (written note): the urls for cards from the Kolb-Proust collection do not actually work
- Strategies used to identify and add links to existing descriptions
  - Motley: Manual process resulted in less results than automatic search for data
  - Run-down of slide material
  - Jeff: can you give full data of success rate in finding any match, manual or automatic?
    - Antoine: would encourage to openly report as much info as possible on the process, could potentially be part of the reason JCDL paper was not as well received as hoped
    - 240 manual hrs for Motley --likely more, and MJ will update the number
  - ISNI, from OCLC, has a lot of names not in VIAF, and a ton of already-built-in data on the names, but you can only access via a paid membership, which is prohibitive for now
    - Batch processing is an issue right now (no API to hit for the info). The manual side is possible, but time-prohibitive (and terms might be different)
  - Jeff: as much as information released as possible about the process and the challenge of reconciling against data sources would be very helpful in a publication for the community
  - Reconciliation of automatic and manual processing
    - Tim: Finding the best one that works is the end point
  - Doug: might the production company be more interesting than the venue of the performance?
    - Tim: yeah, would be great, but it's hard to find reliably. Will keep an eye out for a way to incorporate
    - MJ: we really have to have URIs, though, which can also prohibit incorporating more data
- Summary and preliminary observations from baseline Google analytics data and round of user testing, and plans for Motley testing on new interface this spring
  - Was there any scenarios where they would use linked data?
    - Tasks given could be done without them.
    - For white papers, it would be useful to include the interview script as an attachment for context and potentially for other studies to use to model their work
  - Google Analytics
    - Collecting data for use on pages (referrals, clicks, etc.) and others like devices used to access, etc.
    - Presentation of GA data from Jacob and example of interesting stats
  - What other metrics, that are relatively easy to gether, would be beneficial to do so?
    - Jeff: if Google Images inclusion is important, GA won't be able to track it, since it's based on page tagging. But Google Search Console (https://www.google.com/webmasters/tools/) will--so you just use a regex on the URL for get item to ID if people are grabbing the images from Google Images
    - Tom: are you looking at both instances of the collections?
      - No, not from Medusa yet, but we probably should--this should be fairly easy to extend to Medusa, though

            - Been focusing on grabbing the CONTENTdm data now, in advance of the deprecation at UI
        - Doug: might be other collections that could be better benchmarks than Harvard Theatre Collection digital images collection
            - Might be a good idea to use, but we also want to benchmark for before and after on the same set of sites
            - Will consult some of the team at Harvard to see what their referrers, etc. are to better position Motley.

12:00 - Box lunch & discussion: other projects we should be familiar/in touch with

1:00 - Work in progress (Jacob & Tim)

- Ideas for axes of interest for visualizing the social network of Marcel Proust and annotation of collection
    - Jacob presenting prototype demo of this from name database in Gephi
        - Over 7,000 names, 80,000 relationships
        - Less related to LOD as it is to adding structure to data and enriching relationships
        - Relationships between families and Proust shown
            - Does the information we have through research already done compare to it?
            - Ways to question what would be interesting to do with data visualization; what would a Proust scholar get out of visualization of the names and how they change over time.
        - What might use cases be for using this data visualized?
            - Jeff: Apache Spark (Zeppelin) has a plug-in called GraphX (http://spark.apache.org/graphx/), that can show edges between nodes and extrapolate direct relationships. Might reveal extra relationships that were explicitly missed, but implicitly exist
                - MJ: we could also connect to the cards directly from the visualization, and provide context that way for the relationship, and for the collection
            - Rob: big visualizations like this are typically hard for users of the collection, but good for informed audiences. Spark might be a way to cut out the noise a bit and focus on things that are interesting to users by faceting--slicing by time, families, number of connections, adding in
            - Francoise: remember there is bias implicitly in the cards, as Kolb asked her to find information for a marriage and add it to the card, based on her own judgment
                - Tim: it would be interesting to see how our users might enrich to collection to potentially work around this. Example: a user interested in a given family or year might be able to annotate other people to a given event, as long as the annotation is attributed
                - Francoise: the goal was to annotate and date letters, so if a letter says, “remember when we met at the wedding,” then they’d take the one from the card and say the people in the correspondence were together--always with an eye to editing the correspondence, not really for constructing a social network.
            - Doug: showed his own visualization of people involved with plays with over 1000 shows, and 4 different productions. Helps show the major figures in plays that can be researched, but also shows people who are often overlooked
                - Not entirely convinced a graph visualization is the best way to represent this data, but thinks having a db with the notion of a graph in it is important
        - What info would be valuable to add to the visualization to better support research questions?
            - Caroline/Francoise: if there was a way to make visible the context of the occurrence? Example: know if there was a big wedding or something causing this particular co-occurrence?

- Jeff: Linked Jazz Project (https://linkedjazz.org/network/) would be a great example to look at for ideas of visualization and relationships
- Antoine: Europeana not looking at graphs or visualization work, but relying on data re-users to create such interfaces
- Doug: Enabling query language to make it usable to generate portions of the graph.
- Cecily: if all of this data is available, scholars can use it to their own purposes. Enabling access to the data, in a good, structured form, might be a better investment in time. Then we can work with scholars on a couple of examples, but let the users do what they want with the data.
  - Tim: will want to better think about annotation methods and how to enable this in a way that is practical
  - Jerry: there's two approaches, allow someone to download the data and use it or to have access to the API. We'll want to be thoughtful on how we allow data access.
  - Jeff: will want to make sure that the access/query language isn't too advanced to inhibit use. ElasticSearch (https://www.elastic.co/products/x-pack/graph) has a way to query graphs that might be useful

- Ideas for replacing and migrating from CONTENTdm
  - Demo of new local system to replace CONTENTdm
  - MJ: we've tried to cull local fields for the collection, as much as possible, but still have a lot, as there are specific needs for certain collections
  - Tim: been trying to strategize for RDF where possible
    - MJ: have started using schema.org properties as well
    - Metadata is also in Solr, with images coming from Medusa, and all pulled together into the collection page
      - (Images are grabbed dynamically)
  - Is there a role for triple stores in the Solr system?
    - Common question: how does it enhance the search or the presentation?
      - Rob: often use the triple store to do graph queries only, but it's not the system of record searching. They'll take the information only used by the triple store fed into it, and the rest through ElasticSearch, which speeds up the search process and hopefully loops in things that might be missed by ES
      - Kevin: triple stores and graph-based search can be very helpful for dynamic exploration, where you aren't quite sure what you want back, or the scope of what you might get.
        - Dynamically means you're, on the spot, putting the search bounds on the semantic limitations of the collection
      - MJ: also a fundamental question of whether or not the system supports URIs. For instance, CONTENTdm doesn't support URIs, though schemas want them
      - Tim: you're not looking for the card, you're looking for the concepts/information in the cards
    - Jeff: search vs. query. Users are often searching, rarely querying. An index can query on behalf of users who search
      - Results are often sorted poorly or inappropriately, so building pre-computed graphs that are stored in ES index record for the thing allows for better results (triple store will often just crunch some results and put them in a random order)
      - Rob: showed an example of poor ranking in Getty's ULAN
        - Kevin: some of this is an interface issue, as information for ranking is likely implicitly in the triple store, but can be hard to collect and then serve in a way that's effective. This second piece might be an interface challenge rather than a metadata/storage/index challenge

- Jeff: query he found a problem with was not finding Rembrandt, but having already found Rembrandt, and trying to find a related entity that might be most prominent
- Kevin: sorting and contextualization of search results can be a challenge, but might always be so, as the hurdle is having the triples/relationships/data in the records, but then there is an ongoing challenge with interfaces

- All: How might we deviate from the modified CDM view?
    -

2:30 - Wrap-up, action items & takeaways (Jerry and Tom will facilitate)

- Users weren't as prominent in the first half of the discussion, which is likely an issue
    - Users as corporate users and individual users, but remember that the former is going to become more and more common and prominent as enriched information is released
- Issues with the O in LOD--if IBDB data is a problem for us to use automatically, what will the implications be for automated users of *our* data that will then be hitting IBDB data and machines as well?
    - This goes beyond just data, but metadata, and licensing, etc.
    - Remember that no matter the income stream, it likely will not cover the cost of upkeep and serving said data
- Terms of use of each collection
    - Even when we think we're pretty certain about use, we can't be totally certain
    - For K-P, we might own the papers but not the rights to the intellectual content
        - Sandberg was an example of a rights issue, as well
        - Photos; they did not know who took them, the subjects in them can be separate from author or each other (w/r/t to rights)
- Figuring out how to create harmony between RDF and the metadata
    - RDF and metadata may have different copyright holders, and it can unclear what applies where and to which part
- Value to the use and users: a lot of work needed to make clear there is a benefit
    - Something that directly exploits capabilities of LOD, in a unique way, is key
        - An uptick of the use of the links added would be good evidence
        - Could also engage commercial interests to get someone to the table to brainstorm the commercial value of the LOD
            - Related, but maybe more altruistic, is working with commercial entities to see how partners might be interested in getting UIUC images into their sites to better increase views
                - Getting images in a central site would be useful and possible
                - Seems in the scope of what Mellon likes, with the spirit of openness
                - Don't know the mechanisms to tell sites what we have, for them to evaluate what to pull, or if they want to pull stuff from us
                - Antoine: WikiMedia foundation might be interested in our data, or linking to it--**Alex Stinson** might be one to contact on this
- Policy questions, not least of all: authority work, and the cost, and if it's worth it if it isn't universalizable or released to the wider public
    - Think about internal effort as a cost for a product/information that is resultant from staff time. Would you pay the amounts of the effort for the data from an external org?
    - Is it worth it or only up to a certain point; might not be completely feasible to do and to sustain
    - What's next after RDF xml or json-ld? We need to strategize it
        - Luckily, we are able to use Mellon's money, and let the data go, since it wasn't internally funding, and Mellon is into openness, but maybe harder to say if it's as likely for internal work and data
            - Not just Mellon's money; university, indirect cost, cost share, tuition for student workers, etc.

    - Cecily: not just on sustainability, but what is the most valuable investment of time for collections? You can't do it all, and our audience is potentially very wide
        - We can't discount one set of users in favor of another, and in favor of those with specialized expertise and use cases
        - Antoine: on this front, also want to make sure knowledge sharing is high--what is our experience? What are our processes? What are our results? What are our recommendations? What are the wider implications of this work for libraries, librarians, researchers, students, content creators, etc.?
            - Might we engage the linked data group from Belgium to knowledge share? Europeana was working on mapping, and put together a paper in Code4Lib--not necessarily prestigious, but reaches a practical audience where the knowledge can be shared widely and have good impact on the ground
        - Doug: keep in mind, the better metadata and description will lead to more work on the collection, but also potentially more licensing advantages--is there a licensing group that might be interested in looking into the repercussions of better described items
        - Our collections are uniquely placed at the intersection of libraries, archives and museums, try to be public about this to build more interest in the collection in its central location.
- The "Faces of Abraham Lincoln" collection might be very advantageous to highlight, as it may have been used in Ken Burns's documentary Civil War, with links added to the items and showing how it can be crunched on back end to show these connections, which are new knowledge.
    - Especially for K-P: you could build a new knowledge graph for Marcel. If modeled and made searchable and queryable, it could be a powerful demonstration of how LOD could be leveraged and used
- Antoine: likely can't be done in project, but the annotation piece, and using annotation tools to support this, will allow other researchers to enhance the data
    - For instance, the idea that they found footage of Proust at a wedding of late (likely not, according to Caroline and Francoise), could this data be added to ours, or linked to ours?
        - Create discussion and or interest
        - Though likely not Proust, there is a ton of great data in there: who attended, who gave which gifts, etc.
            - Comparison of data possible
            - Great example of how key the human component is to research with data--the context between the linkable data
            - Caroline trying to figure out a name, and a good visualization and connection of the people in the files could help
- What are your driving positions; greater use vs. specific users important for final papers.
    - Balancing out completeness, usability of it
- Antoine: Measure the data quality before and after your work.
    - Harvesting schema.org's data is a possibility for Europeana
- Jerry: What does shareable data mean and open data? And how does it impact your schema decisions/modifications?
    - Kevin: even within users vs. other entities, there is nuance. Vast diversity amongst the pool of users. Best idea is to document what our choices are--no pressure to necessarily get the choices all right.
- MJ: balancing how linked we can get without hurting our own traffic/exposure/fit with other collections?
    - What's best for our metadata and users may not at all be so for DPLA or Europeana, etc.
    - But we might be able to serve as pilots for larger orgs to make the jump to more complicated structures of metadata
    - And how might we fit into their broader visions?
- Cecily: Worth having a convo with those funding digitization to glean:
    - Best practices for metadata generation for items that you can share with the community to make sure future efforts could fit our model

- Jerry: Longevity- TEI claimed to be outdated yet it showed why it was such a good idea of late with some efforts, and this project shows that, once the hard intellectual work of mapping is done, the dumping from TEI to a new schema is simple
    - Jerry: What does longevity mean for the data?
    - Jerry: For instance, a lot of what we used schema.org to do to the metadata could've been done in TEI as well--what are the benefits of the former over the latter?
        - So many versions of TEI out there, it's hard to pin down to one down
            - Curve of creating new vocab is short, just adding on to what's there
- Tim: Might try some ideas and circulate to the rest of the group