Entification of Metadata for Knowledge-Graph-Guided Discovery

Honick, Brendan John Pittsburgh Supercomputing Center, USA | bhonick@psc.edu
Polley, Katherine Louise University of Massachusetts Dartmouth, USA | kpolley@umassd.edu
Qin, Jian Syracuse University, USA | jqin@syr.edu

ABSTRACT
This paper reports on the application of entification to collections of archival metadata. Entification refers to the identification of the entities described in information objects and associated metadata records. The process presents issues about whether existing records-style metadata records are suitable and how entification will affect user access and interaction with metadata and information objects. In this study, we developed an entification workflow that includes the following steps: data cleaning, entity reconciliation to create linked data through metadata enrichment, accuracy selection, and developing knowledge graphs to communicate the semantic relationships among entities. Additionally, we discuss the implications of entification for practitioners in the field of information science, especially current limitations in the process.

KEYWORDS
Metadata entification, Metadata enrichment, OpenRefine, Archival content representation, Knowledge graph

INTRODUCTION
Entities in information resources and data are considered as separate and distinct but in many ways are related. Persons and organizations as two of the major entity types share some common attributes and play various roles in other entities—events, places, objects of artistic or utilitarian values, subjects, social and political systems—hence form relations among these entities. Entities are embedded in data and information objects as well as in the metadata describing these data and information objects. A limitation of record-based entity strings is that discovery tends to be confined within cataloging records and affords only limited interactions with the search results. Such discovery is particularly limited to within-archival collections due to the convention of collection-level metadata description. As technological advances expand human capabilities in producing, managing, and utilizing large amounts of data, the string-matching style of information representation and retrieval is no longer sufficient for finer discovery tasks, such as within-archival collection searches and browsing. The maturing entity management infrastructures and techniques open doors for developing new solutions for this problem, which Lorcan Dempsey summarizes as "*entification*," a term describing the practice of transforming strings into things in representing information and data (Dempsey, 2021).

Entifying 'strings' of entities, however, is not trivial for vast metadata sets accumulated over time. It involves not only changes in data structures but also entity disambiguation and metadata transformation (Polley et al., 2021). This process is usually time consuming as human intervention is constantly needed. The highly technical nature also requires trained metadata and computer programming personnel to perform iterative data and workflow analyses and make decisions. This creates a dilemma for libraries that need to entify local archival metadata. On the one hand, local archival collections contain a large number of entities that are not covered by authority databases, which makes it difficult to link the entities to existing entity resources. On the other hand, the time and personnel costs of transforming name strings into entities are often prohibitive for many libraries and cultural organizations. To address this dilemma, we conducted an experiment of transforming name strings into entities by using multiple technologies and methods for three archival collections. The purpose is to generate a workflow that can be easily followed to avoid reinventing the wheel and to explore low-barrier tools and methods for budget-stringent cultural organizations to adopt.

In this paper, we report an experiment that builds on our earlier work to enhance the quality of entity representation and explore the tools and workflows for using the improved entity data. This experiment is an attempt to address the two questions mentioned above. In the rest of this paper, we will first review the relevant literature, then describe the methods and workflows we deployed in entity reconciliation, data integration, and graph database transformation. Finally, we will discuss the results and their implications for the entification trend.

RELEVANT LITERATURE
Entity representation has long existed in library databases as part of authority work. The Library of Congress Name Authority File (LCNAF) and Getty's Union List of Artist Names (ULAN) are two examples. Both of these entity databases have been transformed into Linked Data Services provided by these two institutions. As of the time of writing, the Online Computer Library Center (OCLC) just released more than 150 million WorldCat entities as the foundation of linked data infrastructure (https://id.oclc.org/worldcat/entity). Entification of metadata is increasingly

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

becoming an important part of metadata work and is believed to have the potential for more comprehensive applications and uses (Zeng & Qin, 2022).

Linking metadata and modeling the relations between entities in archival collections requires an understanding of the theories and tools that drive those actions. A concept central to both processes is *entification*. As Dempsey (2021) describes this idea, "Entification involves establishing a singular identity for 'things' so that they can be operationalized in applications, gathering information about those 'things,' and relating those 'things' to other entities of interest." Transforming archival metadata from "strings" to "things" reflects the gradual endeavor in information science towards Tim Berners-Lee's concept of the Semantic Web, a "linked data information space in which data is being enriched and added" (Shadbolt et al., 2006, p. 100). As demonstrated in Dempsey's definition, entification has different phases depending on the type of linked data work being performed. This literature review focuses on the stages of entification: moving from "strings" to "things," entity reconciliation, and identifying the semantic relationships among entities.

Moving from "Strings" to "Things"

Although Dempsey's (2021) terminology is novel, the idea of entification has been around in research and practice for more than a decade. In the United States, it is marked by the launching of BIBFRAME as a web of data (Library of Congress, 2012). A synonym that occurs in scholarly research is the concept of making "strings" into "things." As Fox (2016) describes, this process involves "a departure from the time worn paradigm of treating descriptions as a collection of strings, and an effort to embrace the potential for object relational and truly semantic integration" (p. 6). Even though entification is not explicitly named, Fox's description reflects Dempsey's (2021) definition. Within librarianship, entification is highly prevalent in author name disambiguation, in which unique identifiers are applied to individuals to differentiate them (Fox, 2016). Examples of this practice are numerous, ranging from the Open Researcher and Contributor ID (ORCID) and the OCLC's Virtual International Authority File (VIAF) to the Friend of a Friend (FOAF) ontology, which provides "fixed RDF name spaced URIs that contain unique identifiers for individuals" (Fox, 2016, p. 4). Having unambiguous and delineated identifiers for the entities described in information objects is the foundational level for beginning to understand the semantic relationships between those entities.

Moving from "strings" to "things" through entification has clear benefits for library users. Gracy (2017) notes that patrons appreciate methods of searching through archival collections that are similar to how they seek information in non-academic contexts. As she notes, "Because of the tremendous power of Google, users are accustomed to having access to information from a multitude of sources automatically amassed via a single point of entry" (Gracy, 2017, p. 3). Uniform Resource Identifiers (URIs) power these "single points of entry" because URIs can be used to express entities, such as people and places, across domain or disciplinary boundaries. This un-siloing of metadata improves the findability of information objects for library users, especially in cultural heritage collections that involve multiple languages (Godby & Smith-Yoshimura, 2017; Gracy, 2017). In practice, assigning unique identifiers for entities as part of the entification process gives users greater control over searching for information. Kostakos (2020) describes how mixing subject terms with unique identifiers (i.e., "things") with strings provides a more comprehensive search experience than relying on strings alone. Likewise, Gracy (2017) points out that entities with unique identifiers can be paired with temporal events for more dynamic searching, letting users find "common threads" (p. 362). Thus, entification is part of the larger undertaking of making items in library collections more findable for users.

Entity Reconciliation

The second stage of Dempsey's (2021) tripartite definition of entification involves amassing information about entities (the "things") in a library collection. A common method for accomplishing this task prevalent in information literature is *entity reconciliation*. As Enríquez et al. (2017) define, entity reconciliation "involves identifying entities from the digital world that refer to the same real-world entity…This is a complex problem, since it is not trivial to assert that two heterogeneous data instances represent the same real object" (p. 14-15). Recent scholarship has focused on how information professionals can leverage existing technological tools to perform entity reconciliation. For instance, OpenRefine is a data-cleaning software that enables data curators to link collection metadata to a chosen database (Delpeuch, 2019). OpenRefine has built-in functionality with Wikidata, an open database for entities described in Wikipedia and other projects (Delpeuch, 2020; Vrandečić & Krötzsch, 2014).

Entity reconciliation with Wikidata through OpenRefine allows data curators to expand the semantic depth of their collections. As Zhu (2019) explains, "In Wikidata, every piece of information is described as a triple (subject, predicate, object); and each triple can itself be further described/qualified with a series of triples (such as time range, context, aspect, provenance)" (p. 228). Thus, descriptions of entities can be enriched beyond the metadata present in siloed library collections through reconciliation with Wikidata. Case studies in the literature support this concept. Cooey (2019) describes how the European Holocaust Research Infrastructure (EHRI) used Wikidata to "enhance

and expand its authority records for Holocaust-era camps and ghettos" (p. 83). Koho et al. (2021) describe the process of reconciling data from the Medieval Manuscripts in Oxford Libraries datasets to Wikidata and other external sources, such as VIAF. They distinguish between automatic and semiautomatic reconciliation (Koho et al., 2021). The former occurs when a data curator uses entirely automated processes to link data, and the latter involves human consideration of entities (based on the curator's domain knowledge) to ensure higher accuracy (Koho et al., 2021). The type of reconciliation employed during the entification process varies depending on the dynamics of a given collection.

Modeling Semantic Relationships

The final phase of Dempsey's (2021) definition of entification focuses on unveiling the semantic relationships between entities. As Khoo and Na (2007) define, "Semantic relationships are meaningful associations between two or more concepts, entities or sets of entities…The concepts/entities are an integral part of the relation as a relation cannot exist by itself but have to relate two things" (p. 2). These relationships are commonly described in the form of triples (Khoo & Na, 2007; Sadeghi et al., 2017). An entity or a concept can be in as many relationships as required. Zhang et al. (2021) note that as triples are generated to describe a collection of data or metadata, "each relation triple can provide information to other relation triples," creating an interdependent network of meaning (p. 2).

Similar to the limitations related to entity reconciliation, data curators modeling semantic relationships must strike a balance between human accuracy and technological expediency. Beyond minuscule data sets, automated tools are commonly employed to expedite semantic relation extraction. Using an expert system allows a data curator to augment their local knowledge with a rule-based artificial intelligence that "mimic[s] the reasoning procedure of a human expert when solving a knowledge intensive problem" (Roldán-García et al., 2017, p. 1). Enríquez et al. (2017) provide an expert systems perspective on how entity reconciliation is connected to semantic relationships. As they state, "[Entity reconciliation] analyzes all the information related to entities from data sources. Then, it applies probability and scoring to determine which identities can be matched and which non-obvious relationships exist among those entities" (Enríquez et al., 2017, p. 15).

Methods to adjust probability and scoring exist in tools like OpenRefine (Delpeuch, 2020). Knowledge graphs are one way to display both readily apparent and "non-obvious" relationships among entities in data sets. As Sadeghi et al. (2017) define, "a knowledge graph is a fabric of concept, class, property, relationships, and entity descriptions…It aims at a holistic representation of knowledge covering multiple sources, multiple domains, and different granularity (p. 331). Knowledge graphs are powerful tools because they can integrate data from multiple third-party sources, such as Wikidata and DBpedia (Collarana et al., 2017; Sadeghi et al., 2017; Waagmeester et al., 2020). For data and metadata sets, this stage of entification is beneficial to users because the goal of the Semantic Web is to match "users' needs and content" (Zhu, 2019, p. 2). Thus, modeling the semantic relationships of reconciled entities through knowledge graphs enhances how users interact with datasets.

METHODS

The Ted Koppel Collection and the Belfer Cylinders Digital Collection at Syracuse University Library each contain metadata about thousands of individuals. Metadata for the Belfer Collection has 3,191 person names. The Koppel Collection is considerably larger, featuring 72,988 total names. In both metadata sets, some names appear more than once, with the Belfer Collection having 1,378 unique names and Koppel Collection containing 30,906. Because of the size of each of the collections, it was necessary to use automated tools to process this "big metadata" (data about big data) into linked data to facilitate the creation of a knowledge graph based on the relationships between described entities. One of the central challenges was determining the accuracy of automatically processed data. Concerns about including too many false positives or false negatives had to be balanced. In the context of this project, decisions about accuracy had to be made at the individual collection level before generating relevant knowledge graphs. This project of performing entification on person names in archival metadata consisted of the four steps described below.

Step One: Prepare the Data

Because of the size of the collections, the open-source data cleaning tool OpenRefine was selected to prepare the metadata for reconciliation. This program was selected because of the success other researchers in the literature have had with it in reconciling varying collections of library metadata to local and third-party databases (Carlson & Seely, 2017; Tillman, 2016). Each collection's metadata was obtained in the form of Microsoft Excel spreadsheets. In those files, person names were divided into given, middle, and family names columns. Delpeuch's (2019) research on entity disambiguation demonstrates that for entity reconciliation to occur as part of the entification process, those entities need to have their names in a single string. Therefore, those columns were merged in OpenRefine using the "join columns" function so that each person's name was listed in full in a singular column in the order of given name, middle name (or initial), and family name. Additionally, many individuals described in the Belfer Collection

had their birth and death dates in the same column. As OpenRefine's (n.d.) documentation on reconciliation notes, additional relevant details can be used from columns other than the list of entity names when matching with an external source. Thus, the birth and death date columns were separated through the "split into several columns" function in OpenRefine.

Step Two: Entity Reconciliation (Metadata Enrichment)

OpenRefine's (n.d.) reconciliation function was used with Wikidata, its default service, to create linked data from the collections' person names. Typically, reconciliation can be carried out against a particular "Q number," a unique identifier in Wikidata (OpenRefine, n.d.). However, when the collections' person names were reconciled with Wikidata's Q5 property, "human," OpenRefine was not able to match any of the names (Wikidata, 2022b). As a solution for this source of error, the "reconcile against no particular type" option was selected instead. Additionally, as noted above, birth and death dates were specified in the "also include relevant details from other columns" option in the reconciliation interface (OpenRefine, n.d.). Following the methodology in OpenRefine's (n.d.) documentation, these two sources of data were linked to their corresponding properties in Wikidata, "date of birth" (Q2389905) and "date of death" (Q18748141) (Wikidata, 2021, 2022a). Results from automatically linking the collections' person names to entities described in Wikidata are below.

In addition to the automatic reconciliation that Koho et al. (2021) describe, semiautomatic reconciliation was also performed on the 3181 person names in the Belfer Collection. This task was completed by using the "choose new match" option that appeared under each name post-reconciliation (OpenRefine, n.d.). Each person's name was entered, and their corresponding entry in Wikidata was selected, if applicable. This process relied on local knowledge of the collection, ensuring minimal false positives. However, it should be noted that only the results of automatic reconciliation were employed in the following steps of accuracy selection and knowledge graph creation. Instead, the data generated was used as a baseline for accuracy selection, corresponding with Tillman's (2016) method of integrating local knowledge of the collection in metadata enrichment.

Step Three: Accuracy Selection

Next, it was necessary to determine the accuracy of the reconciliation data. OpenRefine (n.d.) provides facets to support this task. For example, the "best candidate's score" numeric facet details "the range of reconciliation scores of only the best candidate of each cell." The actual scores come from Wikidata itself (OpenRefine, n.d.). For each collection reconciled to Wikidata with OpenRefine, the "best candidate's score" facet lists the scores of the names from 0 to 100, with a higher number indicating a more likely match. In preparation for creating the knowledge graphs, the scores of the reconciled names were divided into increments of 5 ($5<x \leq 100$, $10<x \leq 100$, etc.), and the corresponding number of names was recorded. The results from this step are below.

Step Four: Knowledge Graphs

The basic structure of the knowledge graph is based on the Linked Archives ontology, which included the Belfer Cylinders Collection, the Ted Koppel Collection, and the Ronald G. Becker Collection of Charles Eisenmann Photographs (Dobreski et al., 2019). The initial graph structure contained four node types (Collection, Item, Subject, and Person) and three relationship types (is part of, has subject, and is related to). Collections have subjects, items are part of a collection and have a subject, and people are related to items and subjects.

The original Linked Archives ontology data was stored in an OWL file encoded in RDF. In order to connect the entities in the ontology data to the entities in the spreadsheets that went through the reconciliation process outlined previously, the OWL file was converted to a delimited text file using an 'Export to CSV' plugin for the software Protégé. An alternate delimiter to the comma had to be used due to the presence of commas in the values of some of the data properties. The resulting file was read into OpenRefine to clean the data, removing extra quotation marks and spaces from the values. In addition, any apostrophes or single quotation marks had to be replaced with alternative characters to prevent any parsing errors when inserting the data into the graph database, which uses single quotes to contain property values. Following the data cleaning, the data was then exported from OpenRefine into an Excel spreadsheet to be linked with the reconciliation data.

Using the reconciled values from Step 2, additional columns were added to the spreadsheet containing person name data in OpenRefine. For our proof-of-concept database, we decided to add columns for the following properties: date of birth, date of death, occupation, country of citizenship, and sex/gender. The data for these properties was pulled from Wikidata by the OpenRefine reconciliation service, but the process could be repeated for any additional properties present in Wikidata. For the latter three properties, as well as the person's name, whose values are entities in Wikidata rather than literal values, additional columns were added for the Wikidata IDs (in the form of the letter Q plus the unique identifying number) in addition to the entity label. For occupation and country of citizenship, which could have multiple values, the distinction between "records" and "rows" in OpenRefine was particularly important. By default, multiple values for a single property are added as multiple rows in a single record, which can be an issue when exporting the data into a spreadsheet. There is a function in OpenRefine to combine a record's data

into a single row with a delimiter, so an additional column with the combined values was created. The relevant columns (name and reconciliation data) were exported into an Excel spreadsheet to be linked with the ontology data.

Excel's VLOOKUP function was used to combine the ontology data and reconciliation data spreadsheets into a single table. The person's full name was the common column between the two source sheets. Columns were added to the ontology spreadsheet for date of birth, date of death, sex, occupation, and country of citizenship. In addition to this "master" spreadsheet, an additional spreadsheet was created as a dictionary of the relevant Wikidata entities, with columns for the label, ID, and the type of entity, based on the values obtained from reconciliation with Wikidata. The entity types became new node types in our knowledge graph structure (Occupation, Country, and Sex) as well as three new relationship types to connect Person entities to those new types (has occupation, is citizen of, and is of sex).

To transform the prepared tabular data into a graph structure, a Java program was written to read in the spreadsheets (as text files) and output a series of Cypher statements that could be run to insert the data into a Neo4j graph database. The data in the text files was parsed into Java objects that corresponded to the entities in the ontology file and the additional facets obtained from Wikidata. Each object contained a method to translate the data into a Cypher query to create a node for that object and edges for all relationships between that node and others, and these queries were saved to a text file. A Python script was written to connect to a local Neo4j database instance and run the queries in the text file to create the complete collection database. The code repository for the parsing and insertion process can be found on GitHub at https://github.com/Metadata-Lab/Linked_Archives_Neo4j.

RESULTS

Entity Reconciliation

Entity reconciliation for the Koppel and Belfer Collections was performed both semi-automatically and automatically, per Koho et al.'s (2021) research. As described in Step Two of the Methods section, the semiautomatic reconciliation results from the Belfer Collection were used to provide guidance on the selections made during the automatic process. Out of the 3,181 person names listed in the Belfer Collection, semiautomatic reconciliation achieved a match rate of 72.02%, or 2,291 entities. Although the set of semiautomatically-matched entities was employed to create the knowledge graph, the results from this phase were used to guide automatic reconciliation when decisions were made about the matches' desired accuracy.

Automatic reconciliation of the Koppel and Belfer Collections required determinations to be made about the desired accuracy of the matches. The "best candidate's score" metric from Step Three of the Methods influenced these decisions. As described in Step Three of the Methods section, scoring ranges in progressive increments of 5 were used to generate a more granular view of the collections' matches. The results from this process are shown in Figure 1 and Figure 2. The Koppel Collection had a minimum scoring range of $0<x\leq100$, while the Belfer Collection's was $10<x\leq100$ due to system limitations. From these tables, it was determined that 51,522 person names would be reconciled for the Koppel Collection, and 2,230 names would be reconciled for the Belfer Collection. These figures represented match rates of 70.95% and 70.10%, respectively. The decision to use each amount of names was made based on trends in the data in which the percentage of matches decreased as the scoring ranges approached 100. The sets of automatically reconciled entities were part of the graph database's foundation.

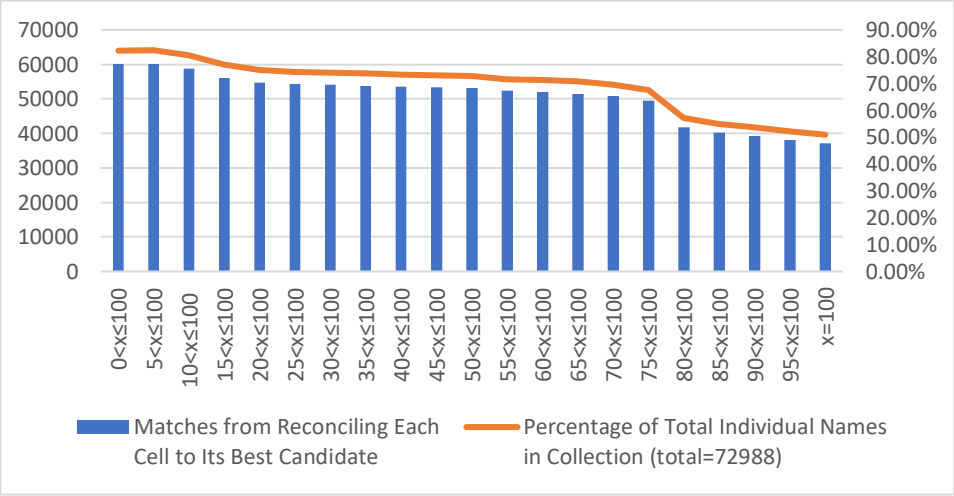

Figure 1. Koppel Collection Automatic Reconciliation by Best Candidate's Score

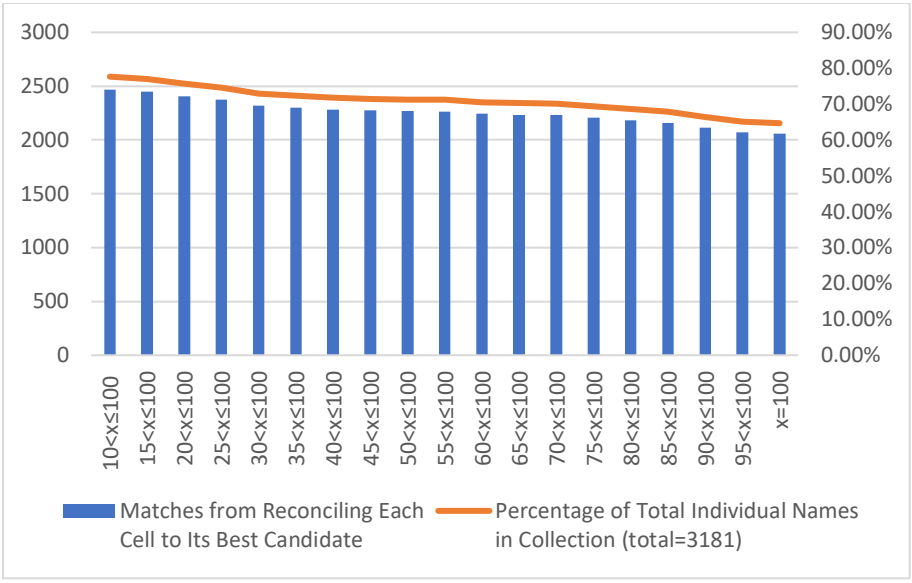

Figure 2. Belfer Collection Automatic Reconciliation by Best Candidate's Score

Graph Database

The graph database created with the metadata for all three collections and the added Wikidata reconciliation data includes seven node types and six relationship types, with a total of 44,896 nodes and 191,590 relationships. Figure 3 below shows the resulting schema of the database in Neo4j, including the types of nodes in the database and how they are related.

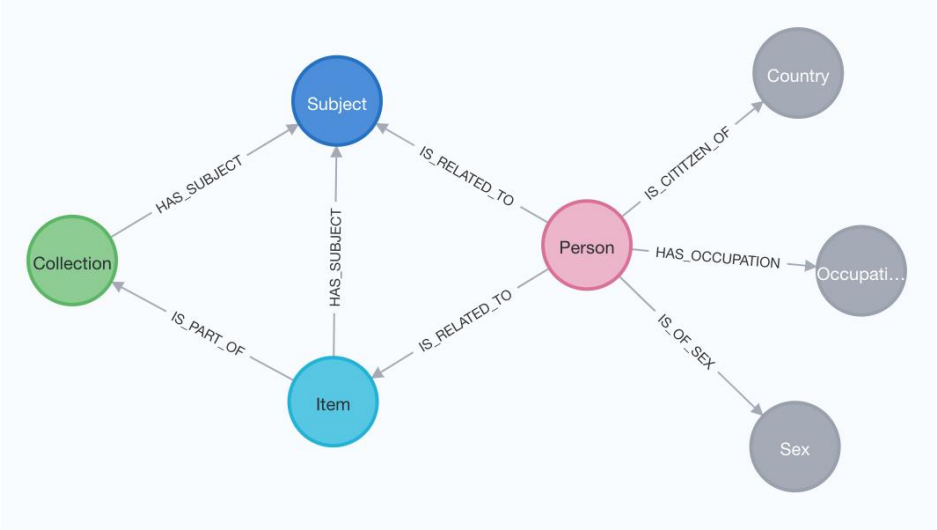

Figure 3. Database Schema Visualization

The knowledge graph can be visualized in a variety of ways. The Neo4j Browser software allows direct access to the database system as well as highly interactive visualizations of the data, including the opportunity to move and expand nodes, as well as view the node properties in a window next to the graph visualization. This can be seen in Figure 4 below. Other software can be used to visualize graph databases, and JavaScript libraries like Neovis.js and Popoto.js provide developers with the opportunity to embed a Neo4j graph visualization into a web page for access.

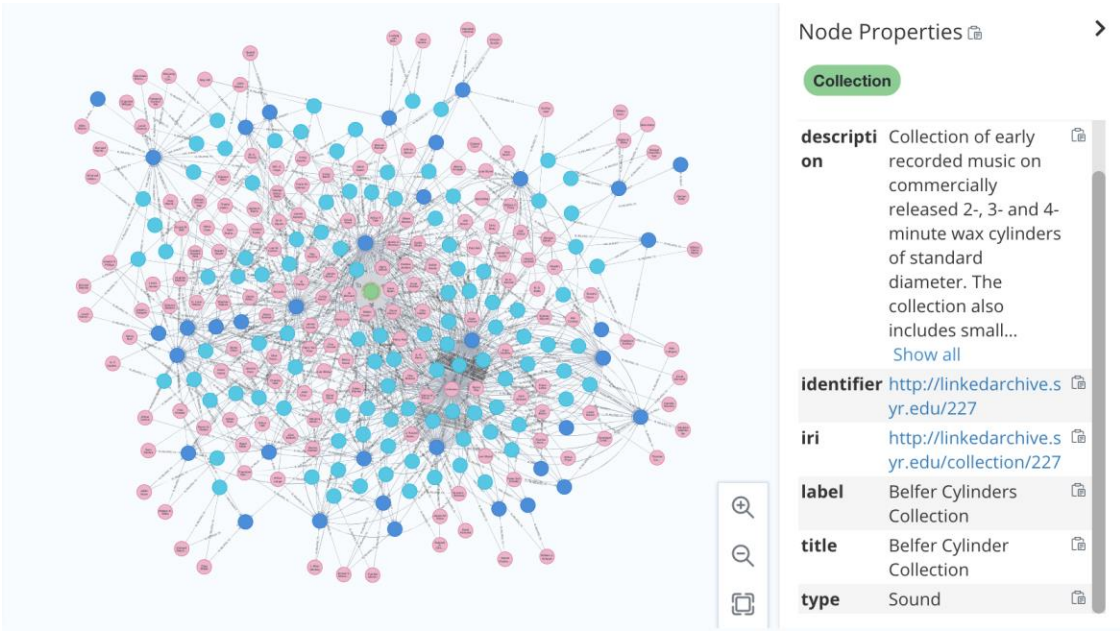

Figure 4. Example Visualization of Belfer Cylinders Collection (300 nodes)

The Neo4j graph structure provides great opportunities for exploring the collections in the database in unique ways. Browsing the database can be done to a limited extent with the Neo4j Browser visualizations, but is difficult due to the volume of nodes, and searching the database requires the use of Cypher queries. While additional software development or programming would be necessary to create an interface accessible to users without the necessary technical expertise, Cypher queries can be used to explore the data and provide insight into various research questions. For example, the query and visualization shown in Figure 5 below would allow a researcher to explore information about non-American individuals who were interviewed during a Nightline episode about the Iran Hostage Crisis. The details of the results can be examined through the visual by selecting and expanding nodes.

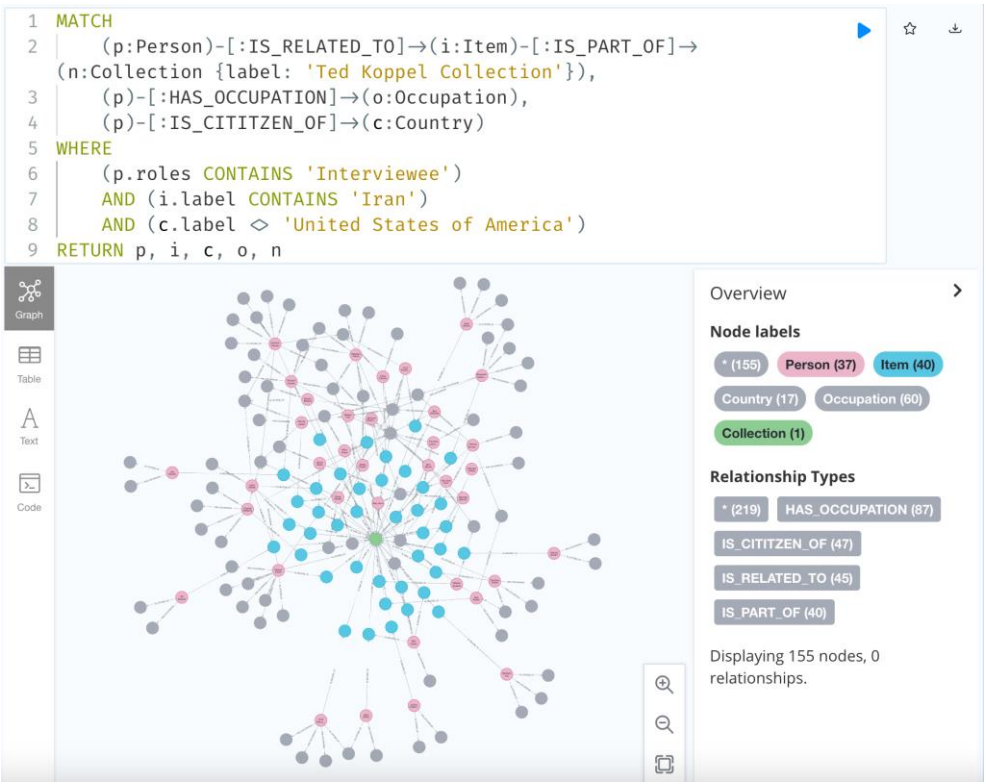

Figure 5. Example Query on Ted Koppel Collection

DISCUSSION

In the experiment reported above, we showed how entity names existing in metadata as strings were entified by semiautomatic and automatic name reconciliation and how the transformed, enriched entities may be utilized for providing in-depth discovery of information and knowledge. Although this experiment is limited in scope, it provides a proof of concept that metadata entification offers many potential benefits for knowledge discovery. There are several lessons learned worth highlighting during this process.

The workflow of entifying name strings in metadata records did not follow a straight linear progression, rather, it went iteratively for some steps, along which trials and errors contantly occurred to adjust for an optimal result. A key component in successful entification is data curators' familiarity with collections. The semiautomatic reconciliation performed for the Belfer Collection is an example of this concept. Each name in that collection was individually checked for a corresponding entity in Wikidata, yielding a match rate of 72.02%. This metric represents a "gold standard" for reconciliation quality because this type of reconciliation relies heavily on researchers' local knowledge of the collection. The semiautomatic match rate provided direction about the ideal percentage of automatic matches when the latter was being carried out. However, semiautomatic reconciliation takes considerable time for substantial collections, meaning that automatic reconciliation tends to be more efficient. Local knowledge is still valuable in that context, although it must be applied at a macroscopic level. Without considerable staff time, automatically reconciled collections will have inaccuracies. For instance, selecting a "best candidate's score" range that is too broad will introduce false positives. Conversely, a range that is too narrow will generate false negatives. Therefore, while entification is a generalizable process for data curators working with archival collections, its accuracy is dependent on local needs.

Traditional information retrieval in archival collections is predicated upon records-based discovery in which users enter query terms into a search system. Entification, as shown in our research, allows for new ways for users to engage with information resources. The linked structure of entities and relationships that defined the existing archival metadata for the collections used in this project translated well to the graph database format, as did the linked data structure of the Wikidata properties used to enrich the metadata. The graph database, built using Neo4j, revealed new opportunities for interacting with the collection and searching for items based not only on their individual properties, but also their relationships with other entities. However, while this is a useful proof of concept, there are some issues with the current setup that pose problems for usability and access.

One limitation is that an additional interface would be necessary to allow users to access and search the collection metadata without needing the technical knowledge to query a graph database with the Cypher query language. There are two JavaScript graph visualization libraries that are built to use a direct connection to a Neo4j instance, but each have their own limitations. Neovis.js is useful for visualizing complex graphs in a web page, but still requires a Cypher query as input, though one could hypothetically be constructed from user input with further development. Popoto.js includes built-in features that are more amenable to browsing behavior by allowing users to explore the metadata using the relationships that exist between different entity types. However, this library, unlike a native tool like Neo4j browser, only allows users to search in the direction of the original relationship. For example, the relationship between an Item (*I*) and a Subject (*S*) is directional, such that *I* has subject *S*, but *S* is not subject of *I*. Though this is not an issue for querying with Cypher, the default behavior in popoto.js would not allow users to move backward from the Subject to the Item. This would need to be remedied by implementing the reciprocal relationship or by using an alternative tool. Of course, the graph database tools selected would depend largely on the institutional needs. This project was done using Neo4j as it is a well-known graph database system that is relatively easy to learn, but there are alternative databases available, as well as other options for interfacing with the database, though a larger development team might be required to implement such a project.

Using the graph database to create a metadata knowledge graph displayed the ability of a linked or graph-based data structure to enhance the user's ability to explore not only an entity's internal properties but also its relationships. However, at present, relations can only be identified at the highest levels. This issue is due to both the amount of information available for each entity in Wikidata as well as technological limitations. For the latter, there exists a bottleneck in artificial intelligence with regards to establishing reliable and valid relationships between nodes. Further software developments will improve the efficiency of creating semantic links between nodes. The entification of metadata will provide the foundation necessary to make those future tools.

CONCLUSION

The entification of metadata records for archival collections allows for new ways of discovering the semantic relationships between entities in those records. Successful implementation of this process relies on a combination of local knowledge of the collections and the use of software to create linked data. The results of entification can include graph databases that empower users to access and interact with semantically richer metadata than what records-based collections provide. However, our research raises important questions about the technological

developments necessary to overcome present limitations in the process. Further study is necessary to examine whether other tools may aid in entification better than the ones described here. Despite current restraints, our results demonstrate that entification is valuable not only for archival research, but also for any discipline that employs metadata records to describe information objects.

REFERENCES

Carlson, S., & Seely, A. (2017). Using OpenRefine's reconciliation to validate local authority headings. *Cataloging & Classification Quarterly*, *55*(1). https://doi.org/10.1080/01639374.2016.1245693

Collarana, D., Galkin, M., Traverso-Ribón, I., Lange, C., Vidal, M.-E., & Auer, S. (2017). Semantic data integration for knowledge graph construction at query time. *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 109–116. https://doi.org/10.1109/ICSC.2017.85

Cooey, N. (2019). Leveraging Wikidata to enhance authority records in the EHRI Portal. *Journal of Library Metadata*, *19*(1–2), 83–98. https://doi.org/10.1080/19386389.2019.1589700

Delpeuch, A. (2019). A survey of OpenRefine reconciliation services. *ArXiv:1906.08092 [Cs]*. http://arxiv.org/abs/1906.08092

Delpeuch, A. (2020). Running a Reconciliation Service for Wikidata. *Wikidata@ISWC*. http://ceur-ws.org/Vol-2773/paper-17.pdf

Dempsey, L. (2021, July 15). *Two metadata directions in libraries*. LorcanDempsey.Net. https://www.lorcandempsey.net/metadata-directions/

Dobreski, B., Qin, J., & Resnick, M. (2019). Side by side: The use of multiple subject languages in capturing shifting contexts around historical collections. *Proceedings from the North American Symposium on Knowledge Organization*, *7*. https://doi.org/10.7152/nasko.v7i1.15615

Enríquez, J. G., Domínguez-Mayo, F. J., Escalona, M. J., Ross, M., & Staples, G. (2017). Entity reconciliation in big data sources: A systematic mapping study. *Expert Systems with Applications*, *80*, 14–27. https://doi.org/10.1016/j.eswa.2017.03.010

Fox, R. (2016). From strings to things. *Digital Library Perspectives*, *32*(1), 2–6. https://doi.org/10.1108/DLP-10-2015-0020

Godby, C. J., & Smith-Yoshimura, K. (2017). From records to things: Managing the transition from legacy library metadata to linked data. *Bulletin of the Association for Information Science and Technology*, *43*(2), 18–23. https://doi.org/10.1002/bul2.2017.1720430209

Gracy, K. F. (2017). Enriching and enhancing moving images with Linked Data: An exploration in the alignment of metadata models. *Journal of Documentation*, *74*(2), 354–371. https://doi.org/10.1108/JD-07-2017-0106

Khoo, C. S. G., & Na, J.-C. (2007). Semantic relations in information science. *Annual Review of Information Science and Technology*, *40*(1), 157–228. https://doi.org/10.1002/aris.1440400112

Koho, M., Burrows, T., Hyvönen, E., Ikkala, E., Page, K., Ransom, L., Tuominen, J., Emery, D., Fraas, M., Heller, B., Lewis, D., Morrison, A., Porte, G., Thomson, E., Velios, A., & Wijsman, H. (2021). Harmonizing and publishing heterogeneous premodern manuscript metadata as Linked Open Data. *Journal of the Association for Information Science and Technology*, *73*(2), 240–257. https://doi.org/10.1002/asi.24499

Kostakos, P. (2020). Strings and things: A semantic search engine for news quotes using Named Entity Recognition. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 835–839. https://doi.org/10.1109/ASONAM49781.2020.9381383.

Library of Congress. (2012). *Bibliographic framework as a web of data: Linked data model and supporting services*. https://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

OpenRefine. (n.d.). *Reconciling*. OpenRefine Documentation. Retrieved March 26, 2022, from https://docs.openrefine.org//manual/reconciling

Polley, K., Tompkins, V., Honick, B., & Qin, J. (2021). Named entity disambiguation for archival collections: Metadata, Wikidata, and Linked data. In: Proceedings of 84th ASIST Annual Meeting, October 30-November 2, 2021, Salt Lake City, UT. https://doi.org/10.1002/pra2.490

Roldán-García, M. del M., García-Nieto, J., & Aldana-Montes, J. F. (2017). Enhancing semantic consistency in anti-fraud rule-based expert systems. Expert Systems with Applications, 90, 332–343. https://doi.org/10.1016/j.eswa.2017.08.036

Sadeghi, A., Lange, C., Vidal, M.-E., & Auer, S. (2017). Integration of scholarly communication metadata using knowledge graphs. In J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, & I. Karydis (Eds.), *Research and advanced technology for digital libraries* (pp. 328–341). Springer International Publishing. https://doi.org/10.1007/978-3-319-67008-9_26

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, *21*(3), 96–101. https://doi.org/10.1109/MIS.2006.62

Tillman, R. (2016). Extracting, augmenting, and updating metadata in Fedora 3 and 4 using a local OpenRefine reconciliation service. *Code4Lib*, 31. https://journal.code4lib.org/articles/11179

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM ACM*. https://doi.org/10.1145/2629489

Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., Good, B. M., Griffith, M., Griffith, O. L., Hanspers, K., Hermjakob, H., Hudson, T. S., Hybiske, K., Keating, S. M., Manske, M., Mayers, M., Mietchen, D., Mitraka, E., Pico, A. R., Putman, T.,

Riutta, A., Queralt-Rosinach, N., … Su, A. I. (2020). Wikidata as a knowledge graph for the life sciences. *ELife*, 9, e52614. https://doi.org/10.7554/eLife.52614

Wikidata. (2021). *Date of death*. https://www.wikidata.org/wiki/Q18748141

Wikidata. (2022a). *Date of birth*. https://www.wikidata.org/wiki/Q2389905

Wikidata. (2022b). *Human*. https://www.wikidata.org/wiki/Q5

Zeng, M.L. & Qin, J. (2022). *Metadata* (3rd ed). ALA-Neal-Schuman.

Zhu, L. (2019). The future of authority control: Issues and trends in the linked data environment. *Journal of Library Metadata*, *19*(3–4), 215–238. https://doi.org/10.1080/19386389.2019.1688368

Zhang, N., Chen, X., Xie, X., Deng, S., Tan, C., Chen, M., Huang, F., Si, L., & Chen, H. (2021). Document-level relation extraction as semantic segmentation. *ArXiv:2106.03618 [Cs]*. http://arxiv.org/abs/2106.03618