A Case Study on Entity Resolution for Distant Processing of Big Humanities Data

[1*]Weijia Xu [2*]Maria Esteva [3 +]Jessica Trelogan [4 □] Todd Swinson
[*]Texas Advanced Computing Center, [+]Institute of Classical Archaeology, [□] Department of Computer Sciences
University of Texas at Austin
[1]xwj@tacc.utexas.edu, [2]maria@tacc.utexas.edu [3]j.trelogan@austin.utexas.edu [4]swinson@gmail.com

Abstract—At the forefront of big data in the Humanities, collections management can directly impact collections access and reuse. However, curators using traditional data management methods for tasks such as identifying redundant from relevant and related records, a small increase in data volume can significantly increase their workload. In this paper, we present preliminary work aimed at assisting curators in making important data management decisions for organizing and improving the overall quality of large unstructured Humanities data collections. Using Entity Resolution as a conceptual framework, we created a similarity model that compares directories and files based on their implicit metadata, and clusters pairs of closely related directories. Useful relationships between data are identified and presented through a graphical user interface that allows qualitative evaluation of the clusters and provides a guide to decide on data management actions. To evaluate the model's performance, we experimented with a test collection and asked the curator to classify the clusters according to four model cluster configurations that consider the presence of related and duplicate information. Evaluation results suggest that the model is useful for making data management action decisions.

Keywords: Collections Management; Entity Resolution; Distant Processing; Digital Humanities

I. INTRODUCTION

One of the challenges introduced by the big data phenomenon in the context of Digital Humanities is that traditional data management methods are not suitable for increasing amounts of digital data. The concepts of "close reading and distant reading," in much use in current Digital Humanities discussions, relate to this problem. They refer respectively to the methods and cognitive demands involved in the close study of a work or passage of text, (such as that done by traditional literary critics like Harold Bloom, who finds distant reading "absurd." See [1]) versus the possibility of studying aggregates of digital texts using computational analysis methods (Moretti; Matthew Jockers prefers to call this "macroanalysis." See [2]). A similar problem applies to Humanities data management in reference to collection processing activities such as organizing, describing, accessing and preserving digital collections. Loosely borrowing on these concepts, we can say that there are close and distant data processing methods; the former understood as linear review methods traditionally used by Humanities scholars and archivists, and the latter as computational methods that narrow large amounts of information to render a meaningful representation. We contend that both; distant and close processing are needed for making sense of big Humanities data. The problem relates to the mismatch between close processing and the amount of data available for curation. Just a small increase in the amount of data makes time consuming for scholars and curators to read a text, or to look at image, or video files, one after another. On the other hand, close processing may be needed for fine grain analysis of the contents. Ideally, close processing can be preceded by distant processing methods to present the curator with a synthesis involving valid patterns found in the data for making research data management decisions.

Research data management has the goal of making data and information accessible throughout the duration of the project and to its disposition or long term archiving [3]. Managing data translates into making decisions about what actions to pursue such as what to keep, discard, merge, and organize. In projects with large teams or especially long lifecycles, increasingly typical in Humanities, researchers commonly share, copy and paste, reuse, select, duplicate, and transform data in response to research questions or as solutions to specific problems, but seldom document their actions along the way. As data are analyzed, new datasets are collected or produced as by-products, and these are added into the research loop, creating new relationships that may not be obvious or explicitly documented. Data provenance may get blurry or lost entirely in the midst of redundancy and disorder. Furthermore, for data to transition to archival collections, data provenance, inter-relationships, and transformations need to be documented or inferred by the archivists.

The goal of this project is to assist users in overcoming these challenges with large unstructured Humanities datasets. Using Entity Resolution (ER) as a conceptual framework [4], we developed a three-stage method that combines both distant and close processing. After the first stage which involves metadata extraction and pre-processing, the second computational stage as distant processing narrows and groups the metadata into clusters that the user can evaluate within a graphical interface. At the post-processing stage, the user reviews the results and decides what data management actions to pursue as close processing (e.g. discard, merge, reorganize, re-evaluate or leave data as is).

ER encompasses a variety of methods developed within different disciplines to resolve cases of data ambiguity and

978-1-4799-1293-3/13/$31.00 ©2013 IEEE

repetition, to link relevant information points, and to improve the overall quality of the data [4]. ER is a well-known problem in Artificial Intelligence (AI). It specializes in finding duplicate records and locating records that might refer to the same entity such as a person, address, or any other information point. The problem that we address in this project is similar to the applications of ER in AI, but it presents considerable differences as well. In AI, ER aims to identify records of the same thing but with different partial information, the entity is clear in that case. In this project, by contrast, we do not have a fixed concept of entity, but we aim to find connections between sets of data that may or may not contain duplicate information.

For this study we used an archaeological collection generated over decades of research and publication activities by the Institute of Classical Archaeology (ICA) at the University of Texas at Austin. This collection, amassed by several generations of research staff and collaborators from a variety of disciplines, is fairly typical of a large archaeological research project. While the case study is focused on the ICA collection's particular formation history research and processing requirements, the approach is widely applicable to any large and unstructured data collection that contains redundant data alongside data that need to be assessed and re-organized for study, archiving, and dissemination.

We created a similarity model that allows directories and files to be compared based on their implicit metadata and that clusters closely related pairs of directories. We call implicit metadata the terms and filenames that users create to label their data as they go about their work, while explicit metadata refers to that which is created following descriptive metadata standards. Knowledge from the curator about the collection is incorporated in the first and third stages of the method. Useful relationships between data are found based on a scoring system assigned to numerical series and tags found in the implicit metadata. Results, presented through a graphical user interface, allow qualitative evaluation of the clusters and consequently making decisions about data management actions. To evaluate the model's performance we classify the clusters according to four model cluster configurations that consider the presence of related and duplicate information. Finally, an action score is applied to each cluster as an indicator of the curator's confidence to make decisions about data management actions.

II. BACKGOUND AND RELATED WORK

A. Formation and Processing of Unorganized Data Aggregations

Problems making sense and processing collaborative, and un-ruled aggregations of data in networked or virtual organizations emerged early on in the history of digital information management and archiving. In *Thirty years of Electronic Records,* Linda Henry describes the early efforts at the National Archives and Records Administration (NARA) to appraise unorganized materials accumulated in multiple personal computers. Her narrative points to the lack of appropriateness of traditional methods used by archivists to approach this task [5]. For cases of chaotic groups of records within unmanaged environments whose access can only be understood by the creator, the UK Public Records Office recommends leaving them out of inventories as long as the information is duplicated elsewhere [6].

Many of the reasons why these aggregations are chaotic derive from the way in which they originate and evolve. In her description of the formation process of an organizational archive dating from the mid 80's to the mid 2000's, Esteva mentions that, despite the fact that a shared network drive was implemented for file sharing and exchange, each staff member kept and discarded files idiosyncratically and nobody could find each other's information [7]. Using ethnographic methods, Boticelli studied the records derived from collaborative research projects and found that the authorship and origin of the records are difficult to pinpoint due to changes in organizational structure and functions of the projects as well as changes in the value of the data as the research proceeds [8]. In the case of archaeology datasets, Trelogan explains that taking care of data management throughout the research project and beyond, places a burden on the research team [9]. In her study of digital preservation practices of archaeologists and art historians, Beaudoin describes faculty image collecting habits as resulting in disconnected, largely invisible, "hoards" or "silos" of data that not only replicate past work, but are largely invisible as part of the scholarly record [10]. Across these papers the commonality is that faculty, research staff, and archivists are overwhelmed by the belief that they are incapable of dealing with large, unstructured collections of digital data.

Preliminary findings from the ChartEX project, that develops methods for analysis of digitized medieval manuscripts while examining researchers response to them, suggest that to conduct their work, researchers need to combine both distant and close data examination methods [11]. In this project we provide that kind of complementary approach. Our aim is to alleviate the burden on Humanities researchers and archivists trying to make sense of large varied datasets by providing a distant processing model that also allows for close processing of more manageable chunks of information.

B. Data and Collection's Organization

Xu and Esteva have experimented with various computational analysis methods for purposes of understanding data and archival collections' organization and aiding their processing. To find the stories, as records related to a same project or event, amongst the text documents of different staff members in an organization, they developed a text mining method to find similarities between text segments [12]. Using the file system metadata of that same dataset, they created a treemap visualization to

uncover individual staff recordkeeping practices across time [13]. The same team developed an interactive visual analytics framework that allows conducting collections structural and functional analyses, and includes a data mining method to infer four types of data organization criteria [14, 15]. With a focus on managing GIS archival data, Heard and Marciano created an application that identifies geographic data and allows the user, via a map-based interface, to explore metadata and records from a "top down" or "bottom up" including a geographical perspective [16]. The work presented here fills a gap in big data processing in the Humanities, in that it focuses on data organization considering the presence of duplicate and redundant information present in the collection.

C. *Entitiy Resoultion*

Entity resolution is originated from a classic problem in database research for duplicate records detection [4, 17]. In large database system in the real word, there are duplicate representations of the same object, also known as "entities" in relational databases. Those duplicate records may not be exactly the same or share a common key reflecting their connection in the system. The value of each record could be only partially matched to other records or present certain errors that make detecting those duplicates a difficult task. A common example is when multiple different formats of the same address are registered in a database. This problem is also related to record linkage, which is to identify records referring to the same entity in different databases.

A typical entity resolution process includes: data preparation, which transforms data in a uniformed model; field matching, which breaks records by field and defines similarity models between fields; and duplicate record detection, which utilizes data mining techniques such as clustering and learning classification, to identify possible duplicated records [4, 18]. Over the years, there have been a number of studies in this area [17, 19, 20]. ER application has also broadened into fields like social network analysis and web data mining [21-24].

Until now, ER has not being used in the context of collections management and curation. Our work focuses on utilizing the general ER framework to help identify related records. There are several differences between our conception and the traditional ER approach. First, most of the existing works focus on the structured data source, such as a database, or metadata from web documents. In this project, we are dealing with less structured data that also lacks uniform and consistent metadata. Secondly, the entity is generally clearly defined in ER, such as a product, or a person etc. In this project, the entity is not explicitly defined. We consider each sub-directory in the collection as a representation of some abstracted concept or theme that might be important for record keeping and should be taken into account in any curatorial decisions. Hence, the detection of duplicated representations of the same entity is translated to the detection of sub-directories that might be related to the same theme or concept. In this exploration, we focused on data modeling and similarity scores.

III. CASE STUDY COLLECTION

As a test case we used a collection generated by the Institute of Classical Archaeology (ICA), which has been conducting research in Greek agricultural territories since 1974. ICA's work has included dozens of excavation projects in southern Italy and Ukraine, in addition to intensive field surveys, campaigns dedicated to site and object conservation, and a huge variety of related studies. The resulting digital collection consists of 5TB of data (over one million files) generated by hundreds of researchers from many different countries and disciplines, each using distinct recording methods and technologies. It contains digitized and born-digital photographs, drawings, maps, plans, and notes as well as more complex datasets including GIS data, 3D models, and relational databases.

Throughout its long history, parts of the collection have been copied, transferred, and amalgamated from every available form of detached media, personal computers, and servers, and it has suffered a great deal from duplication and corruption in the process. In addition, there is a general dearth of descriptive metadata and, although some smaller projects have achieved a degree of consistency, there have been no systematically adopted principles for file naming or organization throughout the entire collection. Implicit descriptive metadata has been inconsistently applied in the form of filenames and directory labels that encode things like names or initials of places and people, abbreviations and acronyms, dates, site codes, and subjects, which provide the only information pointers.

In previous work, the lifecycle of the case study research collection was modeled into three stages: 1) the collection of new primary data (e.g. photography and illustrations of objects), to 2) its study and synthesis (e.g. quantification and analysis of a particular class of artifact), and finally 3) to its publication and dissemination [15]. The motivation for the current project was to locate all relevant data related to one excavation, and, after having reduced as much irrelevant or duplicated data as possible, prepare it for archiving and presentation in print together with an online companion that will include the entire dataset.

The sample used here contains over 100,000 files, most of which are images of varied formats, either scanned from analog media or captured digitally, and spread across 3033 directories. The implicit metadata takes the form of names given to files and directories by the researchers, photographers and scanning technicians in the course of their work, and includes non-standardized terms and numbers that change over time.

IV. METHOD AND IMPLEMENTATION

Due to the diversity of file formats in this collection, a major problem is that there is no way to resolve redundancy with direct content-based comparison methods. Exact

duplicates can be detected by finding files with identical checksums, but checksums also indicate corrupt data, and some duplication is done deliberately, so discarding files based on checksums alone may be misleading. A minimal modification in a file (e.g. minor edits and resizing) will result in a different checksum, even though for the curator, the file may be considered a duplicate. This gets more complex in the case of image data. Using image recognition software for identification of duplicate images would not bring together different related images, nor would be useful for purposes of grouping related data in different formats (e.g. spreadsheets and text documents). Furthermore, pair-wise content comparisons among files are computationally expensive and not practical for large-scale collections. Our approach focuses on the available implicit metadata at the directory level to return clusters that may contain related files. The curator further evaluates if the resultant groupings are useful.

In a nutshell, our method includes a combination of natural language processing (NLP) and data analysis methods. We first tokenize each term in the directory structure and submit a list of terms to the curator, who determines which ones are meaningful for the collection's organizational needs. After pre-processing, each directory is represented by a set of tokens that appear in all the files within that directory. To relate and compare directories, we developed a scoring model that considers three features:: tags, series and structural information. Using this model, the comparison between any two directories yields a vector that quantifies the similarity between two directories. We proceed with a pair-wise comparison amongst all the directories in the test collection. Next, we filter out pairs of directories with low similarity as indicated by the resultant vector and conduct cluster analysis with the remaining directories. Figure 1 shows an overview of the workflow.

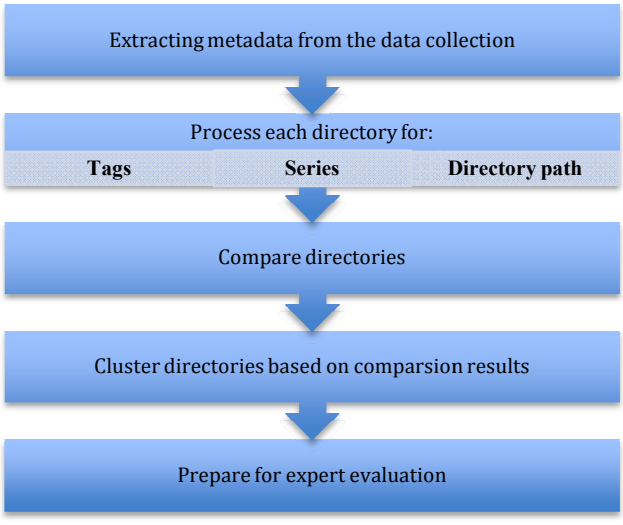

Figure 1: Overview of the processing workflow

A. *Data Pre-processing and Modeling*

The raw input data is the hierarchical structure of the entire collection including all the directories and files. We generated the structural file using a Linux operating system utility named "find." We adopted a tokenizer from the OpenNLP package (http://opennlp.apache.org/), which uses common separators to generate tokens as well as to extract numbers that appear in the filenames and directory labels.

One of the challenges presented by this collection is to identify directories that contain sets of files that, while different, may be related by provenance, function, or theme. For example, a set of image files originally created by one user or project may have been, reused, modified and versioned by other users at different time periods. Our approach is to utilize the numbers and sequences embedded in the naming conventions to identify these relations. However, since numbers or sequences in an individual file may not mean anything, we consider all numbers, sequences, and terms within a directory. Specifically, we are interested first in identifying if a *series*, exists in a given directory. A series could be numbering sequences generated by a camera, or numeric codes used in filenames. Detection of series is not trivial because numbers for example, may be present in different formats and values. We define a series in Definition 1 and introduce Algorithm 1 for series detection.

Definition 1: A series is an ordered set of at least four numerals (or equivalent) i.e. (a_1, a_2, a_3, …, a_n) and where n>3. Furthermore, the difference between intervals of two consecutive numerals are within a given threshold □.

Algorithm 1: Series Detection

1. sort the input sequences,
2. init: Index i ←0 ; series S ← empty
3. For a value a_i in the list;
4. if S is empty
5. S' ← (a_i, a_i+1, a_i+2)
6. If S is not empty
7. S' ← S U a_i
8. Check if S' is a series based on the computed series score
9. if S' is a series,
10. i ← i+1
11. else,
12. i ← i+1 and S ← S'
13. Continue with line 3

In this study, we used a heuristic measure to determine the series score as following:

$$L(a_0 \dots a_n) = \begin{cases} 0 & d < 1 \text{ or } d > 10 \\ 1 & 1 < d \leq 10 \\ 2 & d = 1 \end{cases}.$$

The d is calculated as $d = \frac{a_n - a_0}{n}$. A value of 1.0 of d definitely indicates a series, a score of less than 1.0 means probably a list of categorical numbers. And a set of numbers with high value of d could also be a series with a large

difference between any two consecutive means. However, based on the knowledge we had about the test collection, this could be just random sequential numbers placed by a photographic camera but with no bearing on similar content.

Tags are the tokens extracted from the filenames which are generally very short and may be idiosyncratic. To reduce noise we selected tags based on their frequency of occurrences in the directories. Next, we asked the curator, familiar with the collection, to highlight good and bad tags from the list of frequent tags. Good tags contain useful information such as locations, dates, names, and project abbreviations. Some examples of good tags in the collection are: Metaponto, CH (for Chersonesos), and IT (for Italy). Bad tags are those that, though commonly used, are less indicative of the functions and provenance of the collection. Some examples of those are: music, friends, lost, and files.

In a collection, each directory may have a different number of tags depending on how many files are included within and how varied their names are. To reduce noise and computation costs and based on studying the unique number of tags distribution, we fixed the number of tags to five per directory. The five tags were selected based on the following rules:

- analyze the tags in files in the immediate directory
- in the first position is the most frequent tag
- in the next four positions are the most frequent tags
- if there are not 4 good tags then fill in these spots with the most frequent tags
- if there are not 5 unique tags in the current directory, fill in the remaining positions with UNKNOWN

When identifying series and tags, only the names of files and directory labels immediately inside a directory are analyzed. Previous labels in the full path of the analyzed directory are not considered, and the children in directories further down the tree are not considered either

B. Overall Similarity Scoring Model

After each directory has been analyzed, pair-wise directories need to be compared. The scoring model for comparing two directories consists of the following three features:

1. lists of series identified inside a directory
2. top five tags exctracted inside a directory
3. directory path of the directory

A tag score for each directory pair is computed by calculating the number of matches between the two directories' "top 5 tag" vectors. A positive match in each position of the vector adds 1 to the tag score, for a maximum score of 5. UNKNOWN tags are considered the same, so the tag score can be high even if only one actual tag matches, for example, considering: [a, c, UNK, UNK, UNK] [b, c, UNK, UNK, UNK], score=4. Here, only "c" matches, but since we match UNK's to each other, 3 more matches are counted. A series score for each pair of directories is calculated as follows:

- 0=no series overlap at all
- 1=some series overlap (any range overlap between series of the same number of digits between the two directories)
- 2=complete series overlap (both directories contain the exact same number of series, and each of the series has the same size, min, and max)

For the structural comparison, we calculate prefix and suffix scores between the paths of each pair of directories. The prefix score is the number of consecutive individual directory labels in the path that are the same between two directories, starting from the left. The suffix score is similar but starts from the right. One can view these scores as the longest common prefix or suffix between two strings, but comparing each directory name in a path instead of characters in a string.

We calculate the overall similarity score between each pair of directories as follows. For each pair we construct a vector: [tag score, series score, prefix score, suffix score]. The scores are normalized and transformed so that 0 is the best score, and the highest score is the worst. The similarity score for the directory pair is calculated by multiplying by a weight vector to obtain the weighted sum of elements in the score vector. Lower scores mean more similar. Zero is most similar. To group directories and present this we used k-mean clustering to group the directories into a k value of groups. K values of 21, 50 and 125 were used in our experiment. For each k value, we ran the algorithm five times and chose the one with the lowest sum of squared errors as the output to present to the evaluators.

C. Clustering Analysis and Evaluation

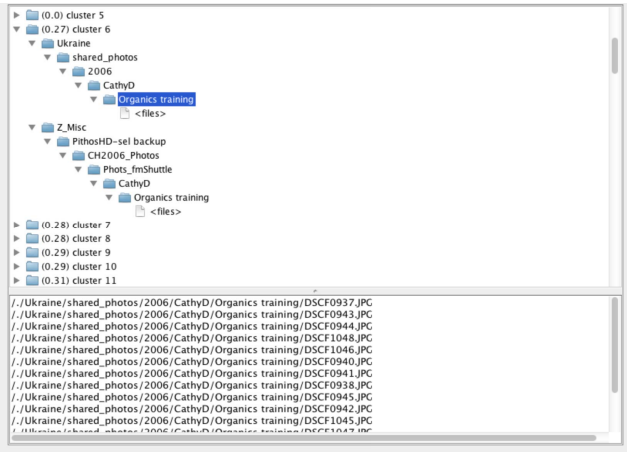

Figure 2. Screenshot of the GUI showing how the contents of each cluster can be closely evaluated by the curator.

Using the similarity scoring model we conducted cluster analysis with a standard k-mean clustering algorithm. For evaluation, for each cluster we also calculated its compactness and sum of square errors. To help users access the clustering results we implemented a graphical user interface shown in Figure 2.

V. RESULTS EVALUATION AND ANALYSIS

The test collection consists of 3,033 directories each with various nested sub-directories. Computationally, this requires running 4.6 million pair-wise comparisons. We computed value distributions based on tags and series features in logarithmic scale. The results are shown in Figure 3.

The tag feature includes the top 5 tags selected from each directory. The tag feature comparison is based on the number of common items found between two directories. Therefore, the result ranges from 0 (no matches found) to 5 (exactly the same set of tags). For the series feature, the 0 value means that there is no overlap (or no series feature) between the two directories. Value 1 means that there is an overlap between the two series, and value 2 means that one directory might have a series that is a superset of the series from the other directory. Figure 4 indicates that in both cases, the majority of the directories have no overlap in the top 5 tag selection (88.8%) and no relationship in the series (93.8%). This means that we selected the subset of directories which have highest scores in both tag and series features. In this set, there are 415 directories, each of which has the exact set of top 5 tags and strong series connection with at least one other directory. Note that the 415 directories account for about 13.7% of total number of directories in the test collection.

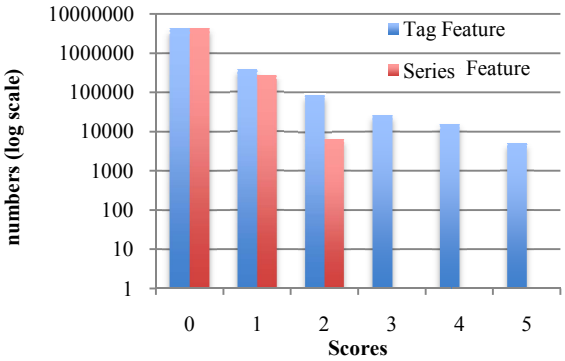

Figure 3. Tag and series features value distribution.

A. *Evaluation Criteria*

Our evaluation criteria included a clustering classification model to typify the clusters and an action scoring system to indicate the usefulness of a resultant cluster. We mentioned that it is important that not all redundancy be automatically discarded, as some may represent significant, intentional decisions made by the data creators. In the process of reducing redundancy, the decisions about whether to delete, merge, reorganize, or leave data untouched must be weighed in context with the rest of the data within which it is stored. For this reason, we defined four different clustering models that share related and duplicate data and used those to compare our clustering results. The four models are shown in Figure 4: 1) pair of clusters share a set of duplicate data; 2) one cluster contains a subset of duplicate data; 3) Pair of clusters contains related, but not duplicate data; 4) cluster is formed entirely by duplicate data. A fifth clustering model was identified as not having related nor duplicate data and thus as a non-useful result. For the results evaluation, each clustering result was compared to the models and classified as containing one or more of them.

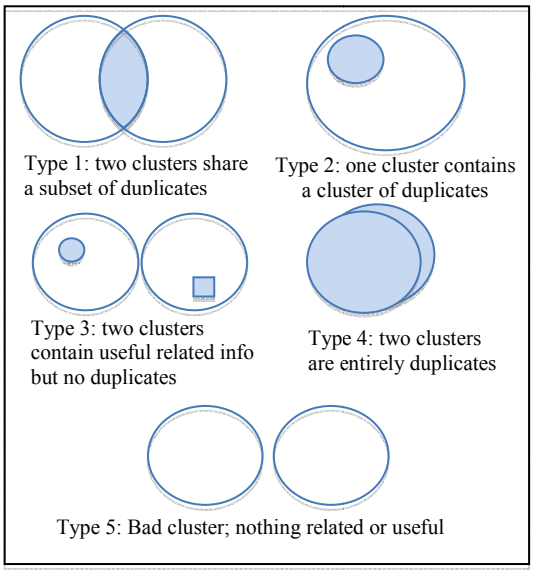

Figure 4. Cluster classification model with five clustering types.

We then proceeded to score the clustering results. This scoring system allowed us to determine the overall performance of the model in relation to data management actions. The scoring represents the level of confidence of the curator in making a data management action (0=ignore cluster; 1= some confidence, but needs further investigation; and 2=complete confidence, take action).

A curator actively working with the test collection reviewed the clustering results after being trained to identify clustering types and clarifying the meaning of the action scoring system. Note that the evaluation done with the GUI (See Figure 2), only involves reviewing directory structure, labels and file names, not reading or observing the files contents. At this point, the curator did not define the type of data management action to pursue, but was asked to make notes about the characteristics of the resulting clusters to improve or refine the model.

B. *Quantitative Analysis of Clustering Comparisons*

We reviewed the three sets of clustering results with a different k-value (21, 50, and 125), with an eye toward which one better achieved the challenges of: a) identifying redundant information for action and b) identifying and merging all data related to one project (e.g excavation site, photographic campaign, publication, etc.). We observed that each set provided different results.

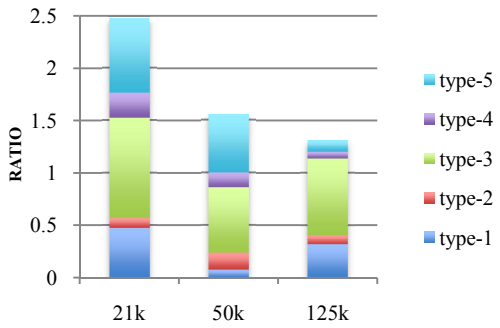

Figure 5. Comparison of different k values in relation to clustering types.

Figure 5 shows a comparison of the different k value clusters, indicating the percentage of individual types of clusters labeled by the curator within the different k results. Because each clustering result was classified as mapping to one or more clustering types, the sum of the percentage of individual types may exceed 1. In fact, a higher value indicates that the clustering is less decisive to the curator as it contains mixed cluster types. Hence, a large number of smaller clusters can further separate out those cases and reduc the number of multiple types in each cluster. In all the clustering results, the type 3, which are clusters containing related directories but no duplicated data, was the most common occurring. We observe that the percentage of type 5 clusters, labeled as useless, decreases as the k values increase. This indicates that with higher k values the clustering results become more homogenous and consequently, more helpful for the curator to make decisions.

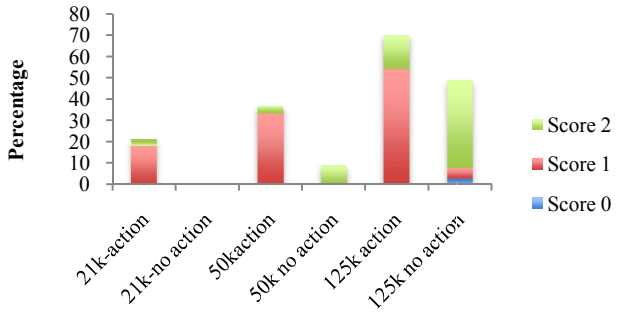

Figure 6. Comparison of action decisions and their confidence in relation to corresponding k value clusters.

Figure 6 shows a comparison between the action decisions made by the curator, their associated confidence levels, and the corresponding k value clusters. The length of each column shows the number of clusters with each decision made. The confidence level of each decision is shown in different colors. 0 indicates no confidence at all and 2 indicates high confidence in the action decision.

When the k value is small, the clusters often contain mixed types of clustering models within. Hence, all clusters with scoring 1 will need close processing. Instead, for the k125 clustering results, we observe that the ratio between action and no action decisions narrows down (70 vs. 49).

C. *Qualitative Evaluation of the Clustering Results.*

The collection's curator completed a qualitative review of the clusters focusing on their content. She observed that in general, across the different k values, the useful clusters (types 1-4) presented data in ways that allowed discovering patterns that would have been impossible to detect through linear review. She was able to derive a list of "action items" and to create new organizational groups based on the clusters. We describe some of observations to illustrate the findings in relation to data management and archival processing practices.

Large sectors of the collection, most of which showed up as types 1 and 4 clusters, were identified as useless old backups that could be immediately deleted. In every case, the backups were located in deeply nested directories. The opportunity to observe this trend in the collection may never have presented itself using close processing methods, as nested directories are confusing to review.

A number of type 2 clusters were of direct interest to the curator. While these were in fact identical copies of other files, they represented deliberate choices by the data creators and, in several cases, allowed for reconstruction of the provenance of files selected out and copied into another directory (either to be processed, reused, or fed into another research project). These duplicates were left as is, but could be described with their provenance information, which would otherwise be lost if they were found out of context.

The curator identified a number of related data as type 3 clusters that never would have occurred as being connected. In more than one cluster, files from the same photographic campaign were identified as related because of file names auto-generated by a digital camera, although they had been separated in two very different directory locations in the archive and labeled differently. This means that the algorithm also discloses different layers of similarity, in this case due to sequencing series that allowed for reconstruction of the relationship between the files as belonging to the same photographic campaign, or sharing the same photographer.

The most "successful" result was the set that used a k value of 125. Although all three k value sets contained useful, and very different, clusters, it performed the best in terms of confidence scores (See Figure 6). For clusters with a score of 2, an immediate decision could be made without close processing. The clusters with scores of 1 were also useful, in that they indicate the need for close processing and as a guide for what to modify in a second iteration of the algorithm. In relation to the latter, the curator observed that, considering the amount of sequencing numbers in this collection, the series score needs to have less relevance in the comparison, and that a dynamic review of the good tags would be useful to data organization goals.

The qualitative evaluation was instructive for: a) establishing priorities for close processing (those action items with confidence score =1), b) reducing unnecessary redundancy, c) informing a deeper understanding of the contents of the archive and d) indicating what needs to be refined in a next iteration of the algorithm. The exercise of reviewing three sets of clusters took a total of ~8 hours. Not only did this "distant" method speed up the process of reviewing and tidying the archive in comparison to reviewing the collection item by item, it also provided significant insight into the formation process and embedded meaning in the collection. The curator noted that actions could be determined after close processing of the cluster results, including directives like "delete directory," "merge with directory Y after deleting duplicates", or "leave as is."

VI. Conclusions and Future Work

As a preliminary attempt to apply general ER framework and techniques in the realm of collection curation, our initial efforts focused on practical benefits for a specific collection. Many aspects of this presentation are specifically tied to the test collection and utilize heuristics from that collection's curator. On one hand, this is necessary due to a lack of prior knowledge or prior models applicable to this problem. On the other hand, the heuristic knowledge, such as series distribution and customized tags lists, simplify the computations and the complexity of the processing workflow. Still, we think that using ER as a framework for data curation processes is a viable approach that can be generalized to other collections with additional learning methods to improve its robustness and effectiveness.

There is an important part missing from the work presented here, namely the scope of the results in relation to the entire collection. Understanding how complete is the representation provided by the distant processing is not yet possible because there is no suitable benchmark set. Creating such a benchmark is not trivial, but is something we will strive for in future work.

In the future, we want to investigate what kind of results might be produced by further modifying the algorithm as a consequence of observations during the first clustering iteration evaluation and of the emergence of new requirements as the data are cleaned, reorganized, and filtered. We also plan to visualize the cluster review process including other metadata such as identical checksums, and file format identification that can help the curator validate his cluster evaluation and improve action decision making.

Acknowledgments

Funding for this work was provided by grants from the National Archives and Records Administration and the Packard Humanities Institute.

References

[1] R. Serlen, "The Distant Future? Reading Franco Moretti," *Literature Compass,* vol. 7, pp. 214-225, 2010.

[2] M. L. Jockers, *Macroanalysis: Digital Methods and Literary History*: University of Illinois Press, 2013.

[3] C. L. Borgman, "The conundrum of sharing research data," *J. Am. Soc. Inf. Sci. Technol.,* vol. 63, pp. 1059-1078, 2012.

[4] F. Naumann and M. Herschel. (2010). *An introduction to duplicate detection*. Available: http://dx.doi.org/10.2200/S00262ED1V01Y201003DTM003

[5] L. J. Henry, "Appraisal of electronic records," in *In Thirty Years of Electronic Records*, B. I. Ambacher., Ed., ed Maryland, USA: The Scarecrow Press, 2003, p. 216.

[6] P. R. Office. (2000). *Guidance for an Inventory of Electronic Records: a Toolkit*. Available: http://www.nationalarchives.gov.uk/documents/inventory_toolkit.pdf

[7] M. Esteva and H. Bi, "Inferring intra-organizational collaboration from cosine similarity distributions in text documents," presented at the Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, Austin, TX, USA, 2009.

[8] P. Botticelli, "Records appraisal in network organizations," *Archivaria,* vol. 1, 2000.

[9] J. Trelogan., A. Rabinowitz, M. Esteva, and S. Pipkin, "What do we do with the mess? Managing and preserving process history in evolving digital archaeological archives," presented at the 38th Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, 2010.

[10] J. E. Beaudoin, "Specters in the Archive: Faculty Digital Image Collections and the Problems of Invisibility," *The Journal of Academic Librarianship,* vol. 37, pp. 488-494, 2011.

[11] S. R. Jones and H. Petrie, "ChartEx: Discovering spatial descriptions and relationships in medieval charters paper," presented at the Digital Humanities 2013, Lincoln, Nebraska, USA, 2013.

[12] W. Xu and M. Esteva, "Finding stories in the archive through paragraph alignment," *Literary and Linguistic Computing,* vol. 26, pp. 359-363, 2011.

[13] W. Xu, M. Esteva, and S. D. Jain, "Visualizing personal digital collections," presented at the Proceedings of the 10th annual joint conference on Digital libraries, Gold Coast, Queensland, Australia, 2010.

[14] W. Xu, M. Esteva, S. D. Jain, and V. Jain, "Analysis of large digital collections with interactive visualization," presented at the 2011 IEEE Conference on Visual Analytics Science and Technology (VAST'11), Providence, RI USA, 2011.

[15] M. Esteva, J. A. Trelogan, W. Xu, A. J. Solis, and N. E. Lauland, "Lost in the Data, Aerial Views of an Archaeological Collection," presented at the Digital Humanities 2013, Lincoln, Nebraska, USA, 2013.

[16] J. R. Heard and R. J. Marciano, "A system for scalable visualization of geographic archival records," in *Large Data Analysis and Visualization (LDAV), 2011 IEEE Symposium on*, 2011, pp. 121-122.

[17] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 19, pp. 1-16, 2007.

[18] T. N. Herzog, F. Scheuren, and W. E. Winkler. (2007). *Data quality and record linkage techniques*.

[19] H. Kopcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems," *Proc. VLDB Endow.,* vol. 3, pp. 484-493, 2010.

[20] P. Christen and SpringerLink (Online service). (2012). *Data matching concepts and techniques for record linkage, entity resolution, and duplicate detection*.

[21] M. Bilgic, L. Licamele, L. Getoor, and B. Shneiderman, "D-dupe: An interactive tool for entity resolution in social networks," in *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, 2006, pp. 43-50.

[22] L. Getoor and C. P. Diehl, "Link mining: a survey," *ACM SIGKDD Explorations Newsletter,* vol. 7, pp. 3-12, 2005.

[23] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 1, p. 5, 2007.

[24] C. P. Diehl, G. Namata, and L. Getoor, "Relationship identification for social network discovery," in *AAAI*, 2007, pp. 546-552.