DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

Automated Parsing of Personal Identity Facets for a Collection of Visual Images

Brian Dobreski
University of Tennessee-Knoxville, USA
bdobreski@utk.edu

Melissa Resnick
University at Buffalo, USA
mresnick@buffalo.edu

Benjamin D. Horne
University of Tennessee-Knoxville, USA
bhorne6@utk.edu

Abstract

Collections of digitized, historical images serve as rich primary sources for digital humanities research, though access to these resources has been hindered by inadequate subject metadata. In this study, researchers explored the feasibility of performing subject analysis for a collection of historical images of persons through an automated procedure. Building on previous work that developed a faceted system for representing the identities of persons depicted in 19th century visual images, the present work attempted to automate the process of person and facet parsing for images from the A.S. Williams III Collection at the University of Alabama. A case-based model was built and used to analyze image titles. Compared to a manual control process, the automated model achieved a 95% success rate in parsing persons and an 85% success rate in parsing facets. Errors in parsing were more likely to occur for images of multiple persons, as well as those labeled with incomplete or uncertain names. Findings offer further support for faceted analysis of personal identity in historical materials, and reveal the potentials of automated, text-based methods of enhancing subject access for large visual image collections.

Keywords: facet analysis; case-based models; personal identity; visual images

1. Introduction

Over the past 30 years, advances in digitization, online systems, and metadata standards have opened up access to unique resources in the collections of cultural heritage institutions such as libraries, archives, and museums. Digitized manuscripts, images, and other materials offer a convenient and accessible means of interacting with historical primary resources. This in turn has opened up new areas of research in the digital humanities and many other disciplines as well. Beyond scholarly research, these resources provide new opportunities for teachers, students, writers, artists, and other members of the general public. The growth of online, digitized historical materials has not been without its challenges, though. The sheer amount of material now available can be overwhelming to users. Accurately representing historical materials as well as their original contexts is also difficult, as the perspectives and terminology present differ from those of modern users.

One means of overcoming these challenges is metadata. Complete, accurate, and well-designed descriptive and subject metadata can help users both navigate large and growing collections of digitized resources and better understand these materials. While such metadata is best planned and implemented before a digitization process (Zeng & Qin, 2022), this may not always be possible. For subject metadata, many digital collections have de facto relied on Library of Congress Subject Headings (LCSH), a system that has long been criticized for providing inadequate access for such materials (Walsh, 2011). Enhancing existing metadata to provide deeper subject access is possible, though the size of some collections complicates this task. This is especially true for collections of visual images, which may require individual, manual review to determine who or what is present.

https://doi.org/10.23106/dcmi.953156738

The possibility of automating a process for enhanced subject analysis of visual images could help overcome some of the barriers to increasing access to these collections.

The present study seeks to explore this possibility. Building on a previously established, faceted framework for representing persons in historical visual images (Dobreski et al., 2021), this work examines the feasibility of automating the process of identifying persons and their identity facets based on image titles. Working from coded data from a previously analyzed collection, researchers built a case-based model designed to replicate the manual process of parsing persons and identity facets from image title text. This automated procedure was then tested on a similar but previously unanalyzed collection of visual images. With minor modifications to the cases, the automated process performed with a high success rate when compared with a manual analysis of the same materials. The results, detailed below, further illustrate the applicability of the faceted identity framework for representing persons in historical images, as well as demonstrate the feasibility of rule-based systems for providing deeper subject analysis and metadata for larger image collections.

2. Background

While cultural heritage institutions maintain a variety of historical materials, collections of visual images hold special interest for researchers and members of the general public. Portraits, posters, and postcards provide a unique window into the history of arts, entertainment, and private life. Among types of commonly held visual images are cartes de visite, a form of early photography popular in the late 19th and early 20th century. These small, black and white images depicted persons, groups, or places and were often given to friends and acquaintances as gifts (Rudd, 2016). Cartes de visite of famous performers were quite common, and were used as a means of advertising and generating revenue (Bogdan, 1988). These images were not limited to famous persons and performers, however, and many private citizens also enjoyed producing and distributing cartes de visite of themselves. In this sense, these images have been likened to modern social media, serving as a means of sharing images and connecting with other persons (Rudd, 2016). Today, many historical cartes de visite of both public and private figures have been preserved in the collections of libraries, archives, and museums, and offer a revealing glimpse into the past for modern users.

As cartes de visite typically depict a single person, subject analysis and access for these materials tends to revolve around the use of labels capturing some aspects of the person's identity. Cultural heritage institutions have faced criticism for the ways they have handled this labelling, from their use of LCSH terminology (Walsh, 2011), to the typically reductivist ways in which persons are represented via selective labels (Rinn, 2018). Traditional practices in representing historical persons and their identities may fail to capture contemporarily relevant aspects of identity, such as race, ethnicity, and culture (Clarke and Schoonmaker 2019; Wright 2019). More recently, cultural heritage institutions have explored the use of faceted systems to represent persons. A faceted vocabulary divides all subject content into a set of recurring categories that are meaningful to a set of users (Hudon, 2019). For example, in 2013, the Library of Congress first developed their Demographic Group Terms (LCDGT), a faceted vocabulary for describing characteristics of groups and persons associated with bibliographic resources, featuring nine major facets including age, ethnicity, religion, and nationality (Library of Congress, 2022).

Previous work with the Becker-Eisenman Collection at Syracuse University, a collection of cartes de visite of sideshow performers, has already demonstrated the potentials of faceted subject representation for these kinds of materials. This archival collection holds over 1,400 cartes de visite and other images of sideshow performers from the 19th and 20th centuries, originally intended for promotion and entertainment purposes (Syracuse University Libraries, n.d.). Preliminary work examining these materials and their metadata noted their currently limited subject analysis, often constituting a single LCSH term for a medical diagnosis, while also demonstrating the potentials of a more robust, faceted system based on image title text (Dobreski et al., 2019). This was followed by a full, inductive, faceted analysis of the collection, using title text to parse persons and personal identity characteristics, assigning these characteristics to one of seven facets, and supplementing

https://doi.org/10.23106/dcmi.953156738

these terms with closest matches from several modern controlled vocabularies (Dobreski et al., 2021). This work resulted in the development of a faceted framework for depicting the identity characteristics of age, gender, race, nationality, condition, relation, and role, as well as a robust set of manual coding procedures and a collection of facet dictionaries mapping common cartes de visite title keywords to these seven facets (Dobreski et al., 2021).

3. Methodology

Building on previous work examining faceted approaches to representing personal identity in visual image collections, researchers sought to develop an automated process for facet parsing, evaluate the performance of this process, and draw conclusions on the feasibility and generalizability of automated facet parsing for these materials. Below, the development of the automated procedure is detailed, followed by the process of applying and evaluating this procedure to a previously unanalyzed collection.

3.1. Developing the Automated Procedure

Previous work on the Becker Collection yielded a manual coding process capable of using image title text to determine the number of distinct persons present, as well as a set of facet dictionaries for determining the presence of terms signifying any of the seven identity facets for each person (Dobreski et al., 2021). While the original process also assigned values to these facets based on multiple controlled vocabularies, the present study focuses only on the process of parsing persons and identity facets. Thus, the goal was to create an automated process capable of replicating these tasks. This goal was initially approached with two computational methods in mind: 1. Named Entity Recognition (NER) and 2. heuristic, rule-based systems.

Named Entity Recognition (also known as entity extraction or entity chunking) is a task in information extraction that attempts to extract named entities (named entities include person, location, organization, etc.) that are mentioned in unstructured text (Nasar et al. 2021). Most NER models are built for generic entity extraction from sentences or blocks of text but can be brittle when used on text that has a significantly different structure than the training data. As pointed out by Nasar et al. (2021), "NER is highly dependent on the textual context, and thus word sequence is important in this problem."

For cartes de visite collections, pictures are labeled with titles that rarely use complete sentences and often only include ambiguous names with one-to-two-word descriptions. Given the unique structure of the titles in the Becker Collection, an NER model was determined to be infeasible. Hence, a simpler but typically less generalizable method was employed, the heuristic/rule-based approach. Rule-based systems are made up of human-crafted rule sets created for a specific group of automated tasks. Broadly, these systems have two parts: 1. A rule base and 2. A parser that can take input and map that input to the rule base (Dunstan 2008).

To build this model, the results of the previous manual analysis of the Becker Collection were examined, forgoing part-of-speech tagging and focusing instead on syntactical patterns used when naming people and their descriptors. After multiple rounds, six main cases emerged. Using these identified cases, researchers built a parser and rule-based function in Python designed to automate the procedure of parsing distinct persons and their identity facets. Comparison of the results of the automated procedure on 800 distinct titles from the Becker to the previous manual analysis results showed 97% accuracy in parsing distinct persons. On a sample of 100 of these titles, the automated procedure matched manual parsing of facet type and number with 72% accuracy.

3.2. Testing the Automated Procedure

To test the automated person and facet detection procedure, researchers first sought a comparable collection for analysis. The A.S. Williams III Cartes de Visite Collection at the University of Alabama was selected due to its accessibility and similarity in scope. This collection contains digitized images of 3,343 cartes de visite, the majority of which are portraits of single persons; the

https://doi.org/10.23106/dcmi.953156738

original images date from 1854 to 1910 (University of Alabama Libraries Special Collections, 2019). While it is similar to the Becker Collection in terms of time period and material type, it should be noted that the A.S. Williams Collection contains images of private citizens as opposed to performers, and were likely intended for personal use as opposed to advertisement. These differences afford an opportunity to examine the generalizability of the previously developed faceted framework to a less specialized collection of historical images of persons.

First, titles for all 3,343 digitized images were collected from the University of Alabama Libraries website. It was found that many images shared the same exact title (e.g., "Portrait of Unknown Man"). As such, titles were deduplicated, resulting in 767 unique titles. Next, following the procedures and facet dictionaries previously developed for the Becker Collection (Dobreski et al., 2021), researchers manually analyzed each of these 767 titles. For any title that signified the presence of at least one distinct person with at least one detectable identity facet, researchers recorded all facets present. For example, for the image titled "Lieutenant J. W. Parrish," the word "lieutenant" signifies three facets: gender (male), age (adult), and role (lieutenant). Facets could occur multiple times for a given image, depending on the text of the title and the number of persons present.

At the same time, the automated model was applied to the same 767 unique titles. During the process, researchers adjusted the rule base and parser to better fit the A.S. Williams Collection. For example, syntactical features noted in the Becker Collection such as dashes and separated descriptors were not present in the A.S. Williams Collection. This resulted in an overall simpler set of rules. For comparison, Table 1 represents the main cases modeled for the Becker Collection, while Table 2 reflects the simplified set of cases used to adapt the model to the A.S. Williams Collection. The accompanying facet dictionaries developed from the Becker Collection were reused for the A.S. Williams Collection without any modification.

TABLE 1: Main cases for the Becker Collection.

Case	Description
1	Title contains a person's name with no description
2	Title contains a person's name with descriptor separated by a dash
3	Title contains multiple people separated by 'and', no descriptions
4	Title contains multiple people separated by 'and' with descriptors for each separated by dashes
5	Title contains multiple people separated by 'and' with one descriptor that applies to all people separated by a dash
6	Title contains multiple people separated by 'and', some people have descriptors separated by a dash, others have no descriptors

TABLE 2: Main cases for the A.S. Williams Collection.

Case	Description
1	Title contains a person's name with no description
2	Title contains multiple people, names separated by 'and', no descriptions
3	Title contains multiple people, names separated by 'with', no descriptions
4	Title contains multiple people, names separated by commas and 'and', no descriptions
5	Title contains one person and a descriptor separated by a comma

The modified automated procedure was used to determine which of the 767 titles had at least one distinct person. For this subset of titles, the process also produced the type and number of personal identity facets present. Following both the manual and automated analysis processes,

https://doi.org/10.23106/dcmi.953156738

researchers compared the results of the two to determine the success of the automated process in parsing both persons and facets.

4. Results

The person parsing task involved identifying which of the 767 unique titles had at least one distinct person present with at least one identity facet determinable. The manual review process determined 639 titles met this criteria. In comparison, the automated process resulted in 666 titles, including all 639 titles that had been manually identified, representing a 95% accuracy. Of the 27 titles found solely by the automated process, all were judged to be incorrect, i.e., no facets should have been determined based on the title; as such. The majority of these cases were names in which initials took the place of forenames, situations in which not even age or gender could be determined. Other errors stemmed from missing forenames, surnames that held other meanings in the facet dictionary, and the presence of place names. Table 3 summarizes these findings and provides examples.

TABLE 3: Person parsing errors.

Error Type	Example Title	Count
Initialism	Portrait of E. B. Lee	21
Surname only	Portrait of [illegible] Coleman	3
Surname meaning	Portrait of H. C. King	2
Place name	Image of Capitol Oyster Saloon and Restaurant	1

The facet parsing task involved identifying the type and number of personal identity facets that could be determined for each title. For the 639 titles with at least one person facet, the automated process was able to exactly replicate the manual coding of both type and number of facets for 544 of them. This represents an 85% success rate on facet parsing. Given the absence of a clear automated baseline for this task, the researchers considered this result successful.

For the 95 titles in which the automated results did not match the manual results, subsequent review determined 2 of these were attributable to errors in the manual coding process. The remaining 93 titles represented errors in the automated facet detection process. The most common error involved the incorrect assignment of the relation facet to one or more persons, for example, incorrectly labeling two persons as married. The second most common error type involved duplicate facets for the same individual, such as when multiple names or phrases identify the same facet for the same person. In these cases, the automated process did not identify all possible facets. Other errors included incorrect age parsing, missing facets, and problems with facet parsing for groups of persons. These errors, alongside examples, are summarized in Table 5.

https://doi.org/10.23106/dcmi.953156738

TABLE 5: Facet parsing errors.

Error Type	Example Title	Count
Incorrect relation assignment	Portrait of Mrs. Tribby and an unidentified man	43
Duplicate facet error	Portrait of a woman named Meg(?) [unknown last name]	23
Incorrect age assignment	Portrait of Alice Auferman	14
Multi-person group	Portrait of a group of unidentified people, five women and two men	3
Missing age	Portrait of R. G. Hull and wife	3
Person with alternative name	Portrait of Elle Re Dickson or J. C. Wheat	3
Facet assigned to indeterminate group	Portrait of Mrs. H. G. Kimball and her daughters	2
Missing gender	Portrait of Mannie [unknown last name] with an unidentified man and an unidentified woman	1
Incorrect gender assignment	Portrait of Lou Rollins	1

Both the manual and the automated analysis processes did not parse any nationality, race, or condition facets in the collection. All facets parsed were from the age, gender, relation, and role categories. The count of all facets determined across the 639 titles are summarized in Table 6.

TABLE 6: Comparison of facet counts.

Facet	Manual Process	Automated Process
Age	183	191
Gender	709	683
Relation	13	61
Role	44	43

Manual and automated processes performed almost identically for the role facet. Performance was similar for age as well, though incorrect age assignments by the automated process resulted in some discrepancies. Differences in gender facet parsing largely reflect the duplicate facet errors described above, while differences in relation facet parsing stem from titles with incorrect relation assignment errors.

5. Discussion

Overall, the automated procedure performed well in comparison with the manual process. The adjusted case-based model was able to parse distinct persons from image title text in the A.S. Williams Collection with 95% accuracy, roughly the same as for the Becker Collection. Given the relatively brief, formulaic titles for cartes de visite, several cases were almost identical between the two collections (see Tables 1 and 2); additional cases for the Becker Collection reflected both the syntactical idiosyncrasies of some title text as well as the distinct ways in which promotional (as opposed to personal) cartes de visite are titled. The automated procedure showed a higher success rate of parsing facets for the A.S. Williams Collection than for the Becker Collection. There are likely several explanations for this finding. First, titles in the A.S. Williams Collection tended to be shorter and less evocative than those of the Becker Collection, resulting in less text and less complex syntax. Second, given the personal as opposed to promotional nature of the cartes de visite, image titling in the A.S. Williams Collection used more literal language, as opposed to the euphemisms and performative monikers prevalent in the titles of Becker Collection materials. Finally, while seven identity facets were observed in the Becker Collection, only four were present in the A.S. Williams Collection (age, gender, relation, role). The absence of nationality, race, or condition facets may have simplified the facet parsing tasks.

https://doi.org/10.23106/dcmi.953156738

While this leaves some uncertainty over how well these facets could be parsed in other collections, performance for the age, gender, relation, and role facets demonstrate the potentials of applying the automated procedure for some enhanced subject analysis tasks for other historical visual images. The overall findings of this study are promising, and though generalizability is limited by the small number of datasets examined, further opportunities for automated analysis of additional collections can be pursued in the future. While the faceted framework and procedures were developed from a collection of specialized images of sideshow performers, they show applicability to more general collections of images of persons as well, opening up new avenues for testing and implementation. For any large collection of historical images of persons, the automated approach developed here could save time on the part of resource describers and organizers. A hybrid approach, in which records not matching any rule cases are marked for manual review, could be particularly effective.

Some distinct challenges were noted during the analysis the of the A.S. Williams materials that are worth further consideration, though. While images of multiple persons posed challenges in both collections, they proved difficult to successfully parse in the more general A.S. Williams Collection due to the absence of rich descriptors and full names. While the Becker Collection materials featured performers with distinct and evocative names, materials in the A.S. Williams Collection featured private individuals whose images were often labeled with initials, incomplete names, or uncertain names. This made parsing gender more difficult, as facet dictionaries for the Becker Collection relied frequently on first names for this. While many terms in the facet dictionaries reflect the euphemisms and stage names of performers, language used in the A.S. Williams Collection is more ordinary. This led to terms like King and Bishop being incorrectly treated as stage names rather than surnames.

This also points to larger issues faced when performing subject analysis for cartes de visite and other historical images of persons. Due to the time period and material type, researchers were able to make certain assumptions about identity. For instance, in both collections, terms of military rank were used to parse a male gender for the individual. While this holds true for 19th century America, this is not an assumption that can be made across all times and settings. Similarly, marriage was used to parse both gender and age, under the assumptions that marriages were between male and female adults. Thus, the facet dictionaries employed here not only reflect 19th century language, but also 19th century understandings and social practices. Due to the relatively limited time frame in which cartes de visite were produced, the assumptions made within the facet dictionaries are generally applicable to collections of these materials. Applying these to other image types, images from different time periods, and images from different cultural settings, however, would prove more problematic. Still, cartes de visite are a distinct and commonly held visual resource in cultural institutions, and are of particular interest in understanding both entertainment and private life, as well for their resemblance to the social media practices of today (Rudd, 2016). The automated procedure developed here stands to provide deeper access to the individuals depicted in these unique materials.

Though fewer identity facet types were present in the A.S. Williams Collection, this study provides further support for the use of faceted representation of personal identities. Providing metadata on age, gender, relation, and role for materials in this collection could enhance access and increase the kinds of questions users could ask of such datasets. It should be noted that the work undertaken in the present study was limited to the subject analytic phase; facets were parsed and filled using terms from the titles themselves. Actually assigning controlled terminology into these slots is a separate task, but the facet dictionaries allow mapping to closest matches in several controlled vocabularies, as well as Wikipedia (Dobreski et al. 2021). Further expansion of the facet dictionaries could allow any number of other vocabularies to be used. Modifications of the facet dictionaries might also make the automated procedure more generalizable to other collections of visual images of persons, though caution would need to be exercised in attempting to accommodate varying linguistic and cultural perspectives together. Ultimately, the facet dictionaries themselves might be further developed into an ontology better capable of doing this. Even so, in their current

https://doi.org/10.23106/dcmi.953156738

state, the facet dictionaries, along with the rule-based system developed here, represent a viable automated option for improving access to and understanding large collections of cartes de visite without requiring intensive manual review.

6. Conclusion

Collections of digitized, historical images offer modern users a unique look into the past, though additional subject metadata is needed to enhance access to and understanding of these primary source materials, particularly those that depict persons. In this study, researchers adopted a previously established, faceted framework for depicting various aspects of personal identity, and developed an automated, case-based model for parsing these facets from the titles of a collection of cartes de visite. The results displayed high accuracy compared to a manual control process. These findings demonstrate the feasibility of rule-based systems for accomplishing subject analytic tasks in cultural heritage collections, as well as the potentials for faceted identity representation for visual images of persons. This process could serve as a less time-consuming alternative to costly manual review of visual images while offering users improved opportunities to find and understand historical resources.

References

Bogdan, R. (1988). *Freak show: Presenting human oddities for amusement and profit*. University of Chicago Press.

Clarke, R. I., & Schoonmaker, S. (2019). Metadata for diversity: Identification and implications of potential access points for diverse library resources. *Journal of Documentation 76* (1), 173-196.

Dobreski, B., Qin, J., & Resnick, M. (2019). Side by side: The use of multiple subject language in capturing shifting contexts around historical collections. In *Proceedings from North American Symposium on Knowledge Organization (7)*, 16-26. doi: 10.7152/nasko.v7i1.15615

Dobreski, B., Qin, J., & Resnick, M. (2021). Depicting historical persons and identities: A faceted approach. *Knowledge Organization 47*(8), 668-679.

Dunstan, N. (2008). Generating domain-specific web-based expert systems. *Expert Systems with Applications 35*(3), 686-690.

Hudon, M. (2019). Facet. In B. Hjørland & C. Gnoli (Eds.), *Encyclopedia of knowledge organization*. ISKO. http://www.isko.org/cyclo/facet

Library of Congress (2022). *Library of Congress Demographic Group Terms*. https://www.loc.gov/aba/publications/FreeLCDGT/freelcdgt.html

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, *54*(1), 1-39.

Rinn, M. R. (2018). Nineteenth-Century depictions of disabilities and modern metadata: A consideration of material in the PT Barnum Digital Collection. *Journal of Contemporary Archival Studies 5*, article 1.

Rudd, A. (2016). Victorians living in public: Cartes de visite as 19th-century social media. *Photography and Culture*, *9*(3), 195-217.

Syracuse University Libraries (n.d.). *Ronald G. Becker Collection of Charles Eisenmann Photographs*. https://scrconline.syr.edu/xtf/search?brand=scrcdc;f1-collection=Ronald%20G.%20Becker%20Collection%20of%20Charles%20Eisenmann%20Photographs

Walsh, J. (2011). The use of Library of Congress Subject Headings in digital collections. *Library Review 60*(4), 328-343.

Wright, K. (2019). Archival interventions and the language we use. *Archival Science 19*(4), 1-18.

University of Alabama Libraries Special Collections (2019). *A. S. Williams III Cartes de Visite Collection*. http://archives.lib.ua.edu/repositories/3/resources/4732

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

https://doi.org/10.23106/dcmi.953156738