Depicting Historical Persons and Identities: A Faceted Approach[†]

Brian Dobreski*, Jian Qin**, and Melissa Resnick***

*University of Tennessee, Knoxville, School of Information Sciences, 449 Communications Bldg,
1345 Circle Park Dr., Knoxville, TN 37996, <bdobreski@utk.edu>
**Syracuse University, School of Information Studies, 311 Hinds Hall, Syracuse, NY 13244, <jqin@syr.edu>
***University at Buffalo, Jacobs School of Medicine and Biomedical Sciences, 955 Main St., Buffalo,
NY 14203, <melissa_resnick@hotmail.com>

Brian Dobreski is Assistant Professor in the School of Information Sciences at University of Tennessee-Knoxville. His research focuses on the social implications of metadata, resource description, and other knowledge organization practices, as well as the concepts of personhood and personal identity. Brian received his PhD in information science from Syracuse University. His dissertation, "Values in Knowledge Organization Standards: A Value Analysis of RDA," received an honorable mention for the ProQuest Best Dissertation Award, as well as runner up for the iSchools 2020 Doctoral Dissertation Award.

Jian Qin is Professor at the School of Information Studies, Syracuse University. Her research areas include metadata and knowledge modeling, knowledge networks in biomedical data, research data management, and scholarly communication. She has published widely in library and information science journals and given presentations at numerous national and international conferences and workshops. Her research has been funded by National Science Foundation and Institute for Museum and Library Services among others. She is the recipient of the 2020 Frederick G. Kilgour Award for Research in Library and Information Technology. Qin holds a PhD degree from University of Illinois at Urbana-Champaign.

Melissa Resnick is currently a post-doctoral fellow at the University at Buffalo, Jacobs School of Medicine, Department of Biomedical Informatics. She holds a PhD from the University of Texas Health Science Center at Houston, a MLS from the City University of New York's Queens College, a MS from Rensselaer Polytechnic Institute, and a BA in biology from the University at Albany. Her doctoral work was in classifying healthcare providers' information needs using Medical Subject Headings (MeSH). Her research interests include terminologies and their use in knowledge organization systems.

Dobreski, Brian, Jian Qin and Melissa Resnick. 2020. "Depicting Historical Persons and Identities: A Faceted Approach." Knowledge Organization 47(8): 668-679. 22 references. DOI:10.5771/0943-7444-2020-8-668.

Abstract: Archives are responsible for presenting historical materials to users while also placing them in context. Historical individuals and their identities pose specific challenges for the archive, including how to negotiate changing cultural perspectives on identity and how to convey and explain contexts to modern audiences. Contemporary subject representation practices in cultural heritage tend to offer a reductivist view of personal identity. Using a collection of visual images of nineteenth and twentieth century sideshow performers, the present work explored a faceted approach to representing personal identity in order to: 1) offer a more holistic view of historical individuals, their identities, and relationships; and 2) help place historical terms and perspectives into context in order to better support user understanding. Utilizing an inductive approach, the researchers developed a framework of seven facets of personal identity. A textual keyword analysis of existing title metadata for images in the collection was implemented in order to identify distinct individuals depicted and any facets associated with them. Finally, collection-based language and closest matches from several controlled vocabularies were used to assign values to the facets for all individuals. The resulting metadata offers a flexible, richer, more holistic analysis of personal identity for these materials, while also exposing some deeper issues concerning identity, performance, and visual imagery. Opportunities exist for further automation and application of this faceted system to other historical collections.

Received: 29 June 2020; Revised: 10 September 2020; accepted: 7 October 2020

Keywords: facet analysis, facets, persons, keywords

† The authors wish to thank Frank S. Shamenek, Ellie Dougherty, and Cassidy Thompson for their efforts in the data collection and analysis for this study.

1.0 Introduction

Advances in collection management, preservation, digitization, metadata, and data publishing have improved access to historical materials and brought them to new and expanding audiences. As stewards of these materials, libraries, archives, museums, and other cultural heritage institutions are responsible not only for presenting them to users but helping the users understand and make sense of them as well (Society of American Archivists 2013). Conveying and explaining the context around historical materials, including the persons, places, groups, and terminologies associated with them is thus a critical task within the cultural heritage domain. This is also a difficult task, however, as changes in conceptualization, language, and perspective over time can leave contemporary audiences either unable to discover what they need or confused or offended by what they may find. This problem is particularly apparent in the subject metadata associated with historical materials. How can subject representation address historical materials that have been seen through changing contexts and perspectives over time? How can subject languages describe resources from a variety of times and places while accommodating modern vocabularies? And, by labelling older materials with contemporary terminologies and perspectives, do cultural heritage institutions actually run the risk of obscuring context and mispresenting these resources and their original cultural settings?

These problems become particularly challenging when dealing with resources that depict historical persons. Recording and using metadata about persons and their identities is a complex but significant area of work in cultural heritage (Dobreski and Kwaśnik 2018), one that becomes even more complicated when representing persons who have been viewed from sharply changing perspectives over time. One group of persons that exemplifies such challenges is American sideshow performers. Sideshows or freak shows were widely accepted forms of entertainment in the United States from mid-nineteenth to mid-twentieth centuries, showcasing individuals with physical abnormalities and promoted through sensational imagery and language. In the twentieth century, public opinion shifted from spectacle and curiosity to embarrassment and sympathy as medical science was able to explain performer conditions in scientific terms (Bogdan 1988). Over time, popular and controlled vocabularies for the conditions these performers embody has grown more reflective of the medical and scientific community's perspectives (Adler et al. 2017). In describing materials that depict these persons, modern subject representation thus runs the risk of reducing historical performers to their clinical conditions and obscuring other important context concerning their identities and lives. In a review of subject analytic approaches to such materials, Rinn (2018) argued that a range of diverse terminology should be used to capture the whole of a person's identity and experience but found that accomplishing this through traditional subject representation strategies was difficult. Indeed, work has shown that traditional cultural heritage metadata generally lacks the capability to fully capture contemporarily relevant aspects of identity, such as race, ethnicity, and culture (Clarke and Schoonmaker 2019). Facet analysis, a longstanding approach to providing a flexible, multi-perspective representation of subject content (La Barre 2010), has been underutilized in this area but may offer one means of addressing these issues.

In this study, an exploration of faceted subject representation for historical depictions of persons is presented through a case study of the Ronald G. Becker Collection of Charles Eisenmann Photographs. Housed in the special collections of Syracuse University, the Becker-Eisenmann materials constitute an archival collection containing over 1,400 images of sideshow performers from the nineteenth and twentieth centuries. Originally intended for promotion and entertainment, these images now serve as an intriguing but complex set of primary sources for users interested in the areas of history, disability, culture, and performance. Previous work with the Becker-Eisenmann and similar collections has shown current subject analysis to be reductive while also struggling to navigate depictions of disability and outdated or offensive language (Rinn 2018; Dobreski et al. 2019). The present work proposes a faceted approach in order to: 1) offer a more holistic view of these performers, their identities, and relationships; and 2) help put different perspectives and terms into context to help users better understand these materials. This is realized through a multipart analysis designed to determine a framework of useful and easily available facets related to the depicted persons, followed by a demonstration of how a mixture of vocabularies may be used within this framework to accomplish a more complete subject representation. This approach holds potential for the Becker-Eisenmann and similar collections, and stands to help their users through offering a more comprehensive treatment of the individuals and complex identities depicted by historical images.

2.0 Background and literature

Though cultural heritage institutions are responsible for creating and maintaining metadata about information resources such as documents, works of art, and artifacts, many are responsible for recording information about persons as well. Representing persons including authors, interviewees, or biography subjects is a significant area of work in cultural heritage knowledge organization and requires the collection and presentation of metadata about persons and their identities (Dobreski and Kwaśnik 2018). Representing personal identities bears distinct challenges concerning terminology and labelling practices though, and the cultural heritage domain has been criticized for failing to capture contemporarily relevant aspects of identity, such as race, ethnicity, and culture (Clarke and Schoonmaker 2019; Wright 2019). These challenges and failures are particularly apparent in relation to collections of materials associated with American sideshows. Sideshows, or freakshows, were a culturally accepted entertainment form in the nineteenth and twentieth century in which persons with unusual characteristics performed for or were displayed to paying spectators. Though some persons were unusual due to their abilities (e.g., fire eating), more were found unusual due to mental or physical differences that would be considered disabilities in modern culture (Kochanek 1997). Many of these individuals were taken from their families under false pretenses with little information remaining about their real identities and origins (Mateen and Boes 2010). In marketing and promoting these performers, sideshow managers often developed stage personas for them involving elaborate, exotic, and often fraudulent backstories that further obscured their identities (Bogdan 2012). Twentieth century shifts in public sentiment and acceptance of sideshows may be attributed to the growth of modern medical science which changed the conception of sideshow performers from unique spectacles to persons afflicted with documented medical conditions (Kochanek 1997). Sideshows waned in popularity as the American public came to view them as unsympathetic and exploitative. In organizing historical materials associated with these sideshows, contemporary cultural heritage institutions face the significant challenge of accurately and respectfully representing sideshow performers and their identities while allowing users to understand the changing cultural milieus in which they have been situated.

One of the more common types of cultural heritage collections associated with sideshows are photographic collections of performers. These photographs were originally sold to generate revenue and advertise for sideshow performances and were typically composed in such a way as to aggrandize or exoticize the depicted performer through spectacular costumes, posing, and evocative titles (Bogdan 2012). Such depictions were inherently exclusionary and marginalizing and pose challenges to the representation, organization, and discovery of these materials (Altermark and Edenborg 2018). In reviewing descriptions from several collections of sideshow performer photographs, Rinn (2018) found that contemporary subject representation of these persons tended to reduce them to medical diagnoses while neglecting other aspects of their identities. Indeed, Adler et al. (2017) found that common cultural heritage subject languages such as *LCSH* have developed in such a way as to reflect the medical community's perspectives while ignoring other perspectives and terminologies associated with disabilities and differences. Representing sideshow performers in the cultural heritage domain is thus often reduced to a task of retrospective diagnosis, labelling historical persons with contemporary medical conditions in spite of incomplete diagnostic information (Mateen and Boes 2010). In studying the research activities of disabilities scholars, Koford (2014) found that many of them did not utilize subject headings during their search process, citing differences in terminological usage and the presence of offensive or confusing labels. Accordingly, Koford called for knowledge organizers to pursue new tactics for subject representation for these materials. Other researchers have similarly called attention to the ways in which current subject representation in cultural heritage settings has continued to reduce and marginalize people with disabilities and physical and mental differences, suggesting alternative, more inclusive strategies be pursued (Beattie 1997; Adler et al. 2017). While making a similar call, Wright (2019) at the same time cautioned that cultural heritage institutions should continue to acknowledge historical language in a manner that explains and provides context without reproducing historical, oppressive power structures. Cultural heritage institutions are thus tasked with a number of challenges in representing, explaining, and providing access to materials that depict sideshow performers, challenges that may require strategies beyond current subject representation strategies.

One alternative approach to subject representation and access that holds promise in this area is faceted classification, a long-standing means of providing multi-perspective representations of content (La Barre 2010). In a faceted representation system, subject content is divided amongst a specific set of conceptual categories and applied across the resources being organized. Facets can thus be thought of as recurring characteristics that are meaningful to some set of users, and different sets of materials may be analyzed along different sets of facets (Hudon 2019). As an organizational structure, facets can be traced back to Aristotle's categories, and more recently, the works of Ranganathan and Bliss (Mills 2004). Though they have traditionally been considered a rationalist approach to organizing, based on the a priori knowledge of the organizer (Hjørland 2013), facets are not entirely without empirical basis. During the process of facet analysis, terms in a given corpus may be examined in order to determine what general concepts they represent; from this, general classes of concepts and relationships emerge and may be used as the basis for facets (La Barre 2010). Facets have become an increasingly common classificatory feature in many information retrieval tools in web environments, with users growing more accustomed to them, particularly in commercial settings (Broughton 2006). In these interfaces, facets often manifest in the form of search filters, designed to help users narrow down large result sets. Facets are a common feature in cultural heritage information retrieval tools as well, and though many such systems utilize facets to represent more descriptive features of materials, they can be used to offer robust subject representation as well. For example, Broughton (2002) explored the use of faceted representation for a collection of digital objects whose current *DDC* and *LCSH* subject representation was not satisfactory for conveying the collection's complex subject matter. Utilizing vocabulary analysis, the study resulted in the construction and application of subject facets capable of providing more meaningful access for digital humanities researchers. Facets can thus offer an alternative and less reductive means of conveying subject matter and terminology for complex historical materials.

3.0 Methodology

Representing the content of photographs or images may be approached in two ways: text-based and image-based. The text-based approach uses keywords to understand and describe aboutness and of-ness, and most work in this approach is performed by human catalogers or taggers. As such, the text-based approach may bear biases due to differences in interpretation and use of terms. The image-based approach is also called query-by-image content (QBIC) that includes three types: spatial-index-based method, feature-based indexing methods, and object- and pictorial-based methods (Ewald 2009). The QBIC methods employ automatic algorithms to analyze features such as the location of an object or scene, color, texture, or statistic of a scenic in the image, or symbolic image of an array representing the spatial relations among objects. Although the QBIC approach has its advantages, it also has its limitations, e.g., inability to detect and understand nuanced historical and social aspects of image content. Given the pros and cons of both image indexing approaches and the existing textual data in the source collection, we decided to adopt the text-based approach; we acknowledge, however, that future QBIC-based approaches may hold potentials for these materials as well.

The data used for this project included the Becker-Eisenman Collection of digitized photographs and the associated item-level metadata. The data set contained 1,415 records representing unique images held in the collection. These metadata records utilize a modified version of Dublin Core. Current subject representation for these materials consists of a single *LCSH* heading per record. Despite the presence of subject metadata, as well as brief descriptions for each image, we chose to limit our analysis solely to the title element. Title metadata for the images reflects the monikers printed on them at the time of their publication; in the absence of a printed title, archivists had previously devised and assigned a brief title using language from similar objects in the collection (e.g., "Unidentified Bearded Lady"). Data in the title element was thus taken to be the most "primary" and contemporaneous to these materials while being least interpreted or translated by subsequent audiences as subject and description elements may be.

Within the collection, multiple images sometimes shared the same title (e.g., four different photographs with the title "Nora Hildebrandt, Tattooed Lady"). As a result, for the 1,415 records, 957 distinct titles were determined and collected. Using this title data, we devised a multi-part methodology designed to detect facet patterns relating to depicted persons and use these facets to enrich the representation of historical and social aspects of their identities. Analysis proceeded in three distinct phases: facet determination, facet analysis, and facet assignment.

3.1 Facet determination

Prior to full analysis of the collection metadata, the researchers worked to determine the set of available facets relevant to the materials as well as to the goals of the study. During this phase, we reviewed sample materials and their title metadata in order to identify recurring aspects of personal identity. During this process, additional sources of reference for describing persons and historical background about persons were also consulted, including the *Library of Congress Subject Headings* (*LCSH*), which contains many headings relating to groups of persons, *Resource Description & Access* (*RDA*), a descriptive standard that includes instructions on creating personal authority records, and Wikipedia from which we found articles on the lives of some of the sideshow performers. Within these standards and reference sources, certain aspects of identity were common, including gender, ethnicity, age group, language, occupation, and relationship to other persons and groups. The researchers used these standards and sources as a point of reference in their determination of facets. The overall process was inductive and iterative, featuring a recurring process of group discussion followed by review of individual collection items. Due to the overall aims of the study as stated above, we focused on facets whose presence could be signified via textual data rather than only through visual inspection of the images. Details of facet determination within this collection are further covered in a prior, preliminary work (see Dobreski et al. 2019).

3.2 Facet analysis

With the facets of interest determined, we performed descriptive analysis on the title metadata for the 957 unique image titles in the collection. Manual keyword analysis and text extraction was performed on titles to isolate descriptive terms and phrases associated with persons in each of the images (e.g., "Chinese midget," "Bearded lady"). During keyword analysis, researchers focused on two tasks: 1) determining how many distinct persons were present; and 2) determining how many facets could be identified for each of those persons.

To determine the number of persons present in each image, titles were reviewed for keywords and grammatical expressions signaling the presence of distinct persons. In addition to personal names, linguistic features in title phrases provided hints for distinguishing between individual persons. In the original titles in Table 1, for example, punctuations (commas) and symbols (dashes) served as cues for differentiating one person from another. Conjunctions and prepositions were also used as cues to distinguish individual persons. For example, for the title, "Major Little Finger and Wife, midgets," two persons were noted as being present. Wording in some titles made the determination of number of persons difficult. For the title, "Italian midgets and family," no distinct persons could be determined through keyword analysis, only indeterminate groups. As such, no distinct persons were noted here. Other entities were also noted in the images, including objects and animals, though these, along with indeterminate groups, were excluded from the remainder of the present analysis due to the scope of the study. At the conclusion of this process, each title was tagged with the number of distinct persons present, with each indicated by its respective identifying keywords and phrases (see Table 1).

In the second task, we attempted to identify the facets indicated by keywords associated with each of these persons, including age, gender, race, family relationship, and other facets established during the first phase of work; keywords and phrases were grouped into one or more of these facets where possible. Family relationships were indicated by keywords such as "wife," "sister/sisters," and "mother." Keywords indicating the gender facet included "boy," "girl," "woman/lady," and "man/men." Forenames and titles were also seen to indicate gender for some persons. Since exact ages were not given, keywords such as "woman/lady," "man/men," and "mother" were taken to indicate an adult age group, while those such as "child," "boy," and "girl" were indicative of a child. Keywords representing nationality and race encompassed such terms as "Borneo," "Chinese," "German," and "Black." Keywords associated with the role facet described a type of performance the person was engaged in, such as "snake charmer" or "dancer."

Single keywords in the titles were often enough to identify the age, gender, race, and role facets for individual persons shown in the photos, but in some instances, single words were too ambiguous to clearly signal the presence of one of the seven facets (e.g., "broken," "sacred," "frizzy"). For single keywords where no facets were clearly indicated, proceeding or succeeding keywords were appended to form a phrase. Marks of punctuation and syntactical clues were used in order to determine whether to check the proceeding or succeeding word. The resulting phrases were then reviewed to see if a facet could be determined. In some cases, even phrases were unable to indicate a facet (e.g., "Unidentified person"). Table 1 provides an example of the in-depth, keyword and phrase analysis of several image titles.

Throughout this process, a dictionary of all distinct keywords and phrases, along with the facets they were associated with, was developed. This would be utilized in the subsequent methodological steps.

3.3 Facet assignment

With facets determined for each person indicated by each title in the collection, the next phase involved assigning keywords into each of these facets. The first task here involved the assignment of collection-based, natural language keywords into facets for all human entities depicted by the items. In many cases, this facet assignment was completed during the facet analysis process described above. In such cases a relevant keyword or phrase from the title that indicated the presence of a facet could be assigned directly into that facet, for example, "giant" into the condition facet, or "black" into the race facet. For other titles, however, a keyword or phrase only implied the value for a facet, for example "Lord" implying a "male" value for the gender facet. In some instances, a keyword had both direct and implied facet values, for example "wife" being directly usable in the relation facet while also implying "female" for the gender facet.

To assign the implied facet values for all human entities represented within the collection, researchers began with a manual review process. Keywords such as "Lord" were reviewed and mapped to value "male." In determining appropriate values, we attempted to use keyword forms already present in titles in the collection where possible. As this mapping process continued, the keyword dictionary was updated to include these mappings, allowing for semi-auto-

Peron Identifying Phrase	Keywords/Phrases	Facets Indicated
Original title: "Admiral Dot, Major Atom - Midgets and Unidentified Normal Sized Man"		
Admiral Dot ... Midgets	Admiral Dot; Midgets	condition (midgets), gender (Admiral Dot)
Major Atom ... Midgets	Major Atom; Midgets	condition (midgets), gender (Major Atom)
Unidentified Normal Sized Man	Unidentified; Normal Sized; Man	age (man), gender (man)
Original title: Moung Phoset and Mah Phoon, The Sacred Hairy Family from Burma		
Moung Phoset ... Hairy Family ... Burma	Moung Phoset; Hairy; Family; Burma	condition (hairy), gender (Moung), relation (family), nationality (Burma)
Mah Phoon ... Hairy Family ... Burma	Mah Phoon; Hairy; Family; Burma	condition (hairy), gender (Mah), relation (family), nationality (Burma)
Original title: Thomas Brockman, black giant		
Thomas Brockman, black giant	Thomas Brockman; black; giant	gender (Thomas), race (black), condition (giant)

Table 1. Examples of keywords and phrase based facet analysis of title metadata.

Facet	Example Textual Indicator	Example Facet Values
Age	Siamese twin babies	babies
Condition	Che-Mah, Chinese Dwarf	dwarf
Gender	Rose Leslie, Fat Lady	lady
Nationality	A German Woman	German
Race	Thomas Brockman, black giant	black
Relation	Pat O'Brien and Wife, Giants	wife
Role	Zoe Zobedia, Snake Charmer	snake charmer

Table 2. Facets of personal identity.

mated mapping of previously encountered terms (e.g., all instances of "Lord" were assigned a value of "male" for the gender facet). At the conclusion of this process, all entities had collection-based, natural language keywords assigned to all of their facets.

In the second task of the facet assignment phase, researchers turned to additional terminological sources to assign alternative values to each facet, thereby depicting multiple perspectives. *Library of Congress Subject Headings* (*LCSH*), SNOMED Clinical Terms (SNOMED CT), and Wikipedia were chosen due to their availability, applicability to collection materials, and ability to represent a range of perspectives. Commonly applied to cultural heritage materials, *LCSH* offers a general controlled vocabulary accommodating a wide range of concepts. SNOMED CT offers a distinct controlled vocabulary reflective of contemporary medical perspectives. While not a formal controlled vocabulary, Wikipedia offers a modern, popular perspective on terminologies; researchers relied on the article titles as representations of the "preferred" terms within this source.

For each facet value, researchers manually mapped the collection-based, natural language keyword value to its closest match within each of the three sources. In instances where a concept had no associated term or close match in a particular source, no facet value as assigned from that source. As in the first task, as manual mapping proceeded the keyword dictionary was updated, recording the chosen mappings for concepts encountered.

4.0 Results

4.1 Facet determination

The facet determination process resulted in seven recurring facets that researchers found to be useful in conveying relevant aspects of personal identity within the Becker-Eisenmann Collection. Table 2 lists these seven facets, along with example textual indicators from title metadata and sample facet values. These specific facets were used as the framework for facet analysis and assignment in subsequent phases of research. Though additional facets such as time period may be derived from further knowledge about the collection, the current study remains limited to only those facets that could be determined from title metadata.

4.2 Facet analysis

Out of the 957 unique titles in the collection, 800 of these were found to contain one or more distinct persons. These 800 titles represent separate 1,219 images in the Becker-Eisenmann Collection. Within these 800 titles, 1,004 persons were identified. No attempts were made to deduplicate recurring persons across titles, meaning that some individuals are represented multiple times due to their association with multiple, different titles.

For these 1,004 persons, a total of 2,245 facet values were determined. The number of facets per person ranged from zero to four, with an average of 2.23 facets per person identified. Gender, condition, and age were the most frequently identified facets. Table 3 shows the frequency of facet occurrence among the 1,004 depicted persons.

As also shown in Table 3, the facets varied in terms of how many distinct keywords and phrases they were associated with. While the age facet was identified 460 times, only forty different keywords and phrases made up these 460 occurrences. In contrast, the gender facet was made up of 381 distinct keywords and phrases; much of this may be attributed to the large number of personal forenames that could be reliably associated with a gender (e.g., "Mary," "John").

It should be noted that some distinct keywords and phrases appeared in multiple facets. For example, in the context of this study, "wife" is a combination of both gender and relation, and thus is associated with values for both of these facets. In fact, of the 630 total distinct keywords and phrases included in the dictionary at the conclusion of this phase, 206 were associated with more than one facet. The presence of such compound concepts is reflective of the nature of the analytical depth of the framework employed here; a larger number of facets would be expected to reveal additional compounds.

4.3 Facet assignment

The facet assignment phase of this project saw researchers using different terminological sources to assign facet values for the persons represented within the Becker-Eisenmann Collection. The first step of this phase involved the assignment of collection-based, natural language keywords into facets for all human entities depicted by the items. To accomplish this, we utilized the dictionary of all extracted keywords and the facets to which they corresponded. As such, much of the data needed for this step was already available at the conclusion of the facet analysis process described above.

Keywords in the dictionary corresponded to values in one of two ways. The first was where keywords were directly usable as facet values, i.e., overt facets. For example, "hairy" was a valid value for the condition facet, while "Sweden" was a valid value for the nationality facet. The second was where keywords were not directly usable as a facet value, but implied one, hence a "covert facet." For example, "enchantress" implied a value of "female" for the gender facet. Some keywords fell under both scenarios. While the term "father" signified a valid value for the relation facet, it also implied an "adult" value for age, and a "male" value for gender. All keywords were manually reviewed again in this phase for both overt and covert facet values.

Example terms from the keyword/facet dictionary are depicted in Table 4, with covert facet values denoted by the use of square brackets. "Girl" and "Princess Pontus" offer examples of phrases associated with multiple facets.

At the conclusion of this process, the dictionary of 630 keywords and phrases contained 445 directly usable facet values and 422 covert, implied facet values.

With the dictionary of collection-based language established, it was then used to assign facet values to each title, for each person depicted by the title. Title keywords were matched with keywords in the dictionary for each identified person, with the corresponding facet values then assigned. Table 5 shows the results of this process for three example titles, depicting a total of six persons.

In the second part of this phase, we turned to additional terminological sources to demonstrate how other perspectives, beyond those represented by collection-based language, could be added into our facet system. For each keyword in the dictionary, researchers reviewed *LCSH*, SNOMED CT, and

Facet	Total Occurrences	Distinct Keywords/Phrases
Age	460	40
Condition	629	149
Gender	852	381
Nationality	45	24
Race	13	5
Relation	139	17
Role	105	54

Table 3. Facet occurrence and related, distinct keywords and phrases.

Keyword	Facets						
	Age	Condition	Gender	Nationality	Race	Relation	Role
Major			[male]				
Midgets		midget					
Chinese				Chinese			
Couple						couple	
Girl	[child]		girl				
Princess Pontus			[female]				Princess

Table 4. Examples from dictionary of keywords and corresponding facet values.

Human Entity	Cumulative Facets	Facet Values
Original title: "Admiral Dot, Major Atom - Midgets and Unidentified Normal Sized Man"		
Admiral Dot … Midgets	condition, gender	condition: midget gender: male
Major Atom … Midgets	condition, gender	condition: midget gender: male
Unidentified Normal Sized Man	age, gender	age: adult gender: male
Original title: Moung Phoset and Mah Phoon, The Sacred Hairy Family from Burma		
Moung Phoset … Hairy Family … Burma	condition, gender, relation, nationality	condition: hairy gender: male relation: family nationality: Burma
Mah Phoon … Hairy Family … Burma	condition, gender, relation, nationality	condition: hairy gender: female relation: family nationality: Burma
Original title: Thomas Brockman, black giant		
Thomas Brockman, black giant	race, condition, gender	race: black condition: giant gender: male

Table 5. Examples of facet values assigned with collection-based terminology.

Wikipedia for closest matches for all corresponding facet values, both overt and covert. For example, for the keyword "father," matches were determined for the values "father," "adult," and "male." Only authorized terms from *LCSH* and SNOMED CT were considered, while article titles in Wikipedia were taken to represent the authorized term forms in this source. In cases where multiple matches existed in a single source, multiple terms were recorded. If no suitable matches were present, no terms were recorded. At the conclusion of this process, the keyword/facet dictionary was enhanced with a total of 690 terms: 249 from *LCSH*, 188 from SNOMED CT, and 253 from Wikipedia. The enhanced dictionary thus took the form of a crosswalk, with values from each of the three controlled vocabularies matched to the value from the original item.

With terminology from multiple sources correlated to the keywords and their facet values, researchers were able to enhance the facet metadata to include terms beyond the collection-based language depicted in Table 5. Table 6 shows one of the example titles, depicting two persons. For each person, four facets were assigned (condition, gender, relation, nationality), and values for these facets were supplied from collection-based language, *LCSH*, SNOMED CT, and Wikipedia.

This facet assignment test followed a workflow in which a number of steps were taken to extract facet-bearing keywords from collection-based terminology, assign keywords to facets, and normalize keywords under each facet based on multiple controlled vocabularies. The fact that multiple vocabularies were used in facet assignment creates a novel ap-

Title: Moung Phoset and Mah Phoon, The Sacred Hairy Family from Burma	
Person: Moung Phoset	CONDITION: hairy (term from item); hypertrichosis (LCSH, Wikipedia); Congenital hypertrichosis (disorder) (SNOMED CT)
	GENDER: male (term from item, SNOMED CT, Wikipedia); males (LCSH)
	RELATION: family (term from item, Wikipedia); families (LCSH); Family (social concept) (SNOMED CT)
	NATIONALITY: Burma (term from item); Burmese (LCSH); Burma (geographic region) (SNOMED CT); Myanmar (Wikipedia)
Person: Mah Phoon	CONDITION: hairy (term from item); hypertrichosis (LCSH, Wikipedia); Congenital hypertrichosis (disorder) (SNOMED CT)
	GENDER: female (term from item, SNOMED CT, Wikipedia); females (LCSH)
	RELATION: family (term from item, Wikipedia); families (LCSH); Family (social concept) (SNOMED CT)
	NATIONALITY: Burma (term from item); Burmese (LCSH); Burma (geographic region) (SNOMED CT); Myanmar (Wikipedia)

Table 6. Examples of facet values assigned from multiple terminological sources.

proach to deep-subject representation. Both overt and covert facet assignment enrich the semantic breadth and depth in representing historical materials which can enable sophisticated, facet-based retrieval and increase the discoverability for these materials.

5.0 Discussion

As described above, a faceted approach to subject representation in the Becker-Eisenmann Collection allowed for a more holistic view of the persons depicted, their identities, and their relationships. By drawing out data such as gender, nationality, and family relationships, we were able to provide richer, clearer insight into aspects of personal lives and identities that are often obscured or overlooked in more traditional approaches to cultural heritage metadata. Beyond representing these persons more fully, the availability of facets could provide new search and retrieval options for users in online interfaces, many of whom are already familiar with facets and filters in contemporary web design (Broughton 2006). These facets, in the context of an information retrieval system, would allow users to ask more complex questions about these materials and the performers associated with them.

The facet determination and analysis employed in this study leveraged existing metadata to add further context around these historical materials. While the descriptive, evocative title metadata for these materials provided useful insight, use of title metadata alone is somewhat limiting and does not represent everything the archive knows about these materials. Metadata from the description field could prove particularly useful, though its analysis would require new considerations and analytic rules. Other facets may be drawn out from knowledge of the collection as a whole and its provenance. For example, time period and occupation represent additional facets that, while not explicit in the metadata, could be inferred from the materials. Further visual inspection of the images, though labor-intensive and infeasible for many large collections, could provide further insight as well. This could be particularly helpful in clarifying family relations. Regardless, the procedures of this study show the potential of enriching subject metadata through the use of pre-existing metadata even without re-analysis of each item.

It must be stated that the set of facets derived in this study unavoidably reflect the modern, academic perspectives of the researchers. Additional work should be undertaken to understand archive users' notions of personal identities and which facets of identity are meaningful to their research and use of historical materials. The determination of values for the facets used in this study was also done from a modern perspective on the past, leading to a certain level of interpretation on the part of the researchers. For instance, terms like "Major" were taken to indicate male values for the gender facet, an assumption that may hold true for late nineteenth century materials but not other contexts. Additionally, some terms must be taken as historical euphemisms rather than at face value; for example, "baby" was a term used to imply obesity. Ultimately, whatever framework of facets is developed and implemented today may not match with future assumptions and understandings, rendering these subject representations less meaningful to future audiences. It may be worth further consideration of how to anticipate the need for materials to be reclassified by future historians who must always endeavor to interpret the past from their present milieu (Jenkins 1991).

In addition to the use of facets, this study also employed the application of multiple terminological sources in order to show variations on perspective and labelling in regards to these materials. This strategy shows potential for providing richer insight into historical identities and helping modern audiences place individuals in cultural context. It also helps mitigate some of the reductivist subject representations that have been noted in analyses of sideshow related collections

(Rinn 2018). In the present study, we chose *LCSH*, SNOMED CT, and Wikipedia as additional terminological sources, though any number of controlled or uncontrolled vocabularies could be utilized. As can be seen above, the three sources we chose showed little variation overall in the terminology applied to gender and relation facets, though further variation was notable in facets such as condition and nationality. While sources such as SNOMED CT offer a deep, fine-grained approach to labelling medical conditions, this can be difficult to leverage in practice as conditions and their underlying pathophysiology cannot be fully assessed from the images and metadata. This can paradoxically result in closest matches to more general terms in SNOMED CT than in *LCSH* (e.g., "short stature disorder" vs. "dwarfism"). Variations in nationality terminology often reflected historical and ongoing geopolitical conflicts (e.g., "Burma" vs. "Myanmar").

In both of these cases, the individuals depicted in this collection may be labelled with non-contemporaneous terms for their identities. The use of the original, collection-based language should thus be maintained as a means of helping users understand cultural and historical differences between contemporary and past perspectives. While Table 5 gives an example of what this might look like, care must be given to the presentation of this metadata in order to support user understanding. The original, collection-based language employs obscure historical euphemisms as well as terms found offensive by contemporary audiences, and further thought must be given toward how to utilize this language while acknowledging the marginalizing, exclusionary milieu under which it was employed. Differences among the multiple vocabularies may also result in user confusion, though providing background information on the terminological sources, their perspectives, and the reasons behind employing them may help in this regard. Any terminological source harbors its own potential biases (Knowlton 2005); transparency may be the best means of helping users navigate these. Overall, however, we believe the use of contemporaneous and multiple controlled vocabularies will be beneficial to the user. As Svenonius (2000) observes, much knowledge organization work is intended to bridge the language of resources with the language of users. In this study, we built multiple bridges from multiple facets of identity to help modern users connect with historical materials.

While a language-based approach to analyzing and representing the subjects of this collection affords a rich, flexible means of depicting complex identities, it also raises further issues around language, performance, and reality. With some exception regarding stage names, the current study takes title language literally and truthfully, though given the inherently performative nature of these materials, this is not always the case. Many individuals performing in sideshows were presented and marketed with fabricated origin stories, obscuring their original identities and the sometimes fraudulent means by which they were separated from their earlier lives (Mateen and Boes 2010; Bogdan 2012). Within the Becker-Eisenmann Collection are instances of unrelated persons depicted as siblings or spouses due to their similar conditions or appearance. Similarly, some conditions were completely "performed" by the persons being depicted, with the performers not, in fact, being conjoined twins or intersex persons. The literal approach taken in this study could undoubtedly perpetuate such misrepresentations, many of which are marginalizing and harmful in nature. Given the information we have, however, how are we to determine what was performance and what was real? For all of the work undertaken with the Becker-Eisenmann Collection, many individuals' real identities, as opposed to their performed identities, remain obscured; our hope is that increased subject access to these materials can prompt renewed interest in this area and facilitate work being done by historians and researchers that may shed further light on the matter.

This dilemma also points to a deeper issue related to the analysis of persons and images. For instance, if a photograph is taken of a person at a costume party during which they are dressed as a doctor, the image does indeed depict a doctor, but should that person be labelled as a doctor when they are simply "performing" one? In some instances, pictures are themselves performances, not accurate depictions of individuals, and in these cases, some subject terms are perhaps more associated with the image/performance itself rather than the person. Assuming we can pick apart which images are intended to be truthful depictions of reality and which are performances, how would we convey this to users? The use of genre and form headings to denote performance or fiction may be worth further consideration in this regard. Just as films are labelled with fictive genre headings, the images could be labelled with similar terminology in order to help users more clearly understand the materials and prevent the ontological collapse of the person and the performance.

Though the overall approach and procedures described here are most immediately applicable to other collections of sideshow performer images, we believe there is potential for generalizability to other types of collections as well. Given the language-based nature of the present analysis, collections of textual materials concerning sideshow performers could benefit quite directly from similar procedures. Collections with more textual data would likely yield greater context about these performers and could further enhance our framework as well. Other collections of visual images depicting persons (e.g., postcards, portraits, carte de visite, family albums) could also be enhanced through an application of faceted subject representation of personal identity. It should be noted, however, that some level of interpretation was required of the researchers during the present keyword analysis, and the decisions made here may hinder generalizability to less closely related collections of materials. Also, images in the Becker-Eisenmann Collection carried descriptive, evocative titles that were useful in a large-scale reanalysis, though many other collections of visual images may not. Using textual data rather than visual inspection in our analysis allowed us to leverage existing metadata and proceed more efficiently than revisiting over 1,400 images individually. While metadata beyond the title field could be used in the analysis of other collections, there may be some image collections where further subject analysis is only possible through visual inspection. In such cases, query-by-image content (QBIC) (Ewald 2009) may prove useful, though further development of such methods would likely be necessary in order to capture the social and cultural aspects of these images. Ultimately, a combination of textual analysis and visual analysis may offer the most robust opportunities for in-depth subject analysis and representation of image collections, though this could represent substantial work.

No matter the route taken, scaling up the procedure described here to other collections would be greatly facilitated by further automation, particularly in the facet analysis and assignment phases. In the facet analysis phase, the objective was to determine which facets were present and what linguistic and semantic patterns those facets might be indicated by. The facets and patterns identified provided opportunities to build a mapping between facets and the keywords in the form of a dictionary. This dictionary may enable the construction of some algorithms to automate the facet analysis of other collections. Doing so would also enrich the current dictionary with additional keywords, though further consideration is needed of how appropriate this dictionary could be for other kinds of collections from other cultural settings.

In the facet assignment phase, the keywords are analyzed to determine which overt and covert facet values are associated with them (e.g., father signifying "father," "male," and "adult"). In addition, other terminological sources are consulted for closest matches on these facet values. Though much of this initial work was done manually, automated, linked data approaches hold potentials for scaling up. For example, during the facet assignment phase, mapping could be performed and written into structured statements, as the examples show in Table 7. Envisioning the dictionary as an ontology, facets are classes and keywords or phrases are instances of these classes. Some facets may have sub-facets while others would have only direct instances such as those for the nationality facet. Each facet has an array of keywords that can be used to construct programs for automating the facet assignment.

This ontology could be enhanced by including terms/entities from various controlled vocabularies as further instances of each facet. For example, in *LCSH*, the term "Dwarfism" (http://id.loc.gov/authorities/subjects/sh85040186) represents a concept that serves as an instance of the condition facet. Fully leveraging this kind of linked data set would require well-designed procedures to build up the relationships between various terminologies and instances. The results, however, could prove useful in applying faceted subject analysis to other collections of materials dealing with sideshows and other aspects of American performance in the nineteenth and twentieth centuries. In addition, this ontology could be used to enhance user search interactions, providing search suggestions and augmenting results with additional terminologies. The overall applicability and potentials of this ontology for a wider array of collections and systems warrant separate investigation.

6.0 Conclusion

While any subject representation of historical materials runs the risk of oversimplifying or obscuring original context, the faceted approach pursued in this paper offers one means of moving beyond traditionally reductivist approaches to representing persons, presenting instead a more holistic, multi-faceted view of personal identity. By leveraging existing metadata alongside perspectives from additional terminological sources, our reanalysis of the Becker-Eisenmann collection works to build bridges between historical vocabulary and the language of contemporary communities, and in doing so, more clearly conveys the contexts around these materials. Beyond the identity framework developed in this paper, other contributions include the keyword/facet/value dictionary that could be utilized in the analysis of other archival collections, as well as the beginnings of an automated procedure that can be expanded upon and applied in future research.

As we look toward applying the findings of this study more broadly, several limitations of the present work must be

Facet	Relation	Sub-facets	Keywords/Instances
Age	hasGroup	child adult	{boy, girl, baby, child} {woman, lady, man, couple, wife}
Nationality	hasIndividual		{Australian, Burmese, Chinese, Circassian, Egyptian, German, Russian}
Condition	hasGroup	natural decorative	{dwarf, midget, giant, hairy, dog-faced, albino, fat, legless, armless, hunchbacked} {tattooed}

Table 7. Examples mapping between facets and keywords as instances.

acknowledged. As expressed above, the construction of our faceted framework reflects our own contemporary perspectives on identity; persons from other times or cultures might arrive at a different allotment of identity facets. Similarly, facet analysis and assignment required a certain level of interpretation on the part of the researchers. Further work should acknowledge and attempt to address such limitations. Most critically, gathering user perspectives on these resources and the meaningfulness and usefulness of the metadata generated in this study would allow for a better assessment of the procedures presented here. Other opportunities exist to further explore the generalizability of this work, including the application of this faceted metadata to resource description and discovery in similar archival collections of visual materials. Finally, additional work to further progress the automation of these procedures could allow for more efficient, large-scale implementation of faceted identity metadata.

References

Adler, Melissa, Jeffrey T. Huber and A. Tyler Nix. 2017. “Stigmatizing Disability: Library Classifications and the Marking and Marginalization of Books about People with Disabilities.” *Library Quarterly* 87: 117-35.

Altermark, Niklas and Emil Edenborg. 2018. “Visualizing the Included Subject: Photography, Progress Narratives and Intellectual Disability.” *Subjectivity* 11, no. 4: 287-302.

Beattie, Diane. 1997. “Retrieving the Irretrievable: Providing Access to 'Hidden Groups' in Archives.” *The Reference Librarian* 26, no. 56: 83-94.

Bogdan, Robert. 1988. *Freak Show: Presenting Human Oddities for Amusement and Profit*. Chicago: University of Chicago Press.

Broughton, Vanda. 2002. “Facet Analytical Theory as a Basis for a Knowledge Organization Tool in a Subject Portal.” In *Challenges in Knowledge Representation and Organization for the 21st Century: Integration of Knowledge Across Boundaries: Proceedings of the Seventh International Conference of the International Society for Knowledge Organization 10-13 July 2002 Granada, Spain,* ed. María J. López-Huertas and Francisco J. Muñoz-Fernández. Advances in Knowledge Organization 8. Wurzburg: Ergon, 135-41.

Broughton, Vanda. 2006. "The Need for a Faceted Classification as the Basis of All Methods of Information Retrieval." *Aslib Proceedings* 58, nos. 1/2: 49-72.

Clarke, Rachel Ivy and Sayward Schoonmaker. 2019. "Metadata for Diversity." *Journal of Documentation* 76, no. 1: 173-96.

Dobreski, Brian and Barbara Kwasnik. 2018. "Changing Depictions of Persons in Library Practice: Spirits, Pseudonyms, and Human Books." *Knowledge Organization* 44: 656-67.

Dobreski, Brian, Jian Qin and Melissa Resnick. 2019. "Side by Side: The Use of Multiple Subject Languages in Capturing Shifting Contexts Around Historical Collections." In *Proceedings from North American Symposium on Knowledge Organization. Vol.7,* 16-26. doi: 10.7152/nasko.v7i1.15615

Ewald, Mark. 2009. “Content-Based Image Indexing and Retrieval in an Image Database for Technical Domains.” *Transactions on Machine Learning and Data Mining* 2, no. 1: 3-22.

Hjørland, Birger. 2013. “Facet Analysis: The Logical Approach to Knowledge Organization”. *Information Processing and Management* 49: 545-57.

Hudon, Michèle. 2019. “Facet.” *Knowledge Organization* 47: 320-33.

Jenkins, Keith. 1991. *Re-Thinking History*. London: Routledge.

Knowlton, Steven A. 2005. “Three Decades since Prejudices and Antipathies: A Study of Changes in the Library of Congress Subject Headings.” *Cataloging & Classification Quarterly* 40, no. 2: 123-45.

Kochanek, Lisa A. 1997. "Reframing the Freak: From Sideshow to Science." *Victorian Periodicals Review* 30, no. 3: 227-43.

Koford, Amelia. 2014. “How Disability Studies Scholars Interact with Subject Headings.” *Cataloging & Classification Quarterly* 52, no. 4: 388-411.

La Barre, Kathryn. 2010. “Facet Analysis.” *Annual Review of Information Science and Technology* 44, no. 1: 243-84.

Mateen, Farrah J. and Christopher J. Boes. 2010. “‘Pinheads’: The Exhibition of Neurologic Disorders at ‘The Greatest Show on Earth.’” *Neurology* 75, no. 22: 2028-32.

Mills, Jack. 2004. “Faceted Classification and Logical Division in Information Retrieval.” *Library Trends* 52: 541-70.

Rinn, Meghan 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.

Society of American Archivists. 2013. *Describing Archives: A Content Standard*. Chicago.: Society of American Archivists.

Wright, Kristen. 2019. “Archival Interventions and the Language We Use.” *Archival Science* 19, no. 4: 1-18.