Diversity, equity and inclusion principles for custom taxonomies

Received (in revised form): 19th May, 2023

Sharon Mizota
Inclusive Metadata Consultant, USA

Sharon Mizota is a metadata consultant who helps archives, museums, libraries and media organisations transform their metadata to improve diversity, equity and inclusion in the historical record. She has over ten years of experience managing and creating metadata for arts and culture organisations, including Walt Disney Animation Studios, the Academy of Motion Picture Arts and Sciences, Harvard Business Publishing and Curationist. She is also an art critic, a recipient of an Andy Warhol Foundation Arts Writers' Grant, and a co-author of the award-winning book, 'Fresh Talk/Daring Gazes: Conversations on Asian American Art'.

E-mail: sharon@sharonmizota.com

Abstract This paper discusses the common challenges and considerations associated with the development of custom taxonomies for describing people along the lines of race, gender, sexual orientation, nationality and other identity facets. The paper recognises that there are inherent and divisive problems in reducing complex human identities to discrete categories, and that terminology is always changing. Nevertheless, in the absence of a public vocabulary that addresses all facets of identity in a respectful and accurate way, custom taxonomies remain necessary. This paper explores the strengths and weaknesses of existing vocabularies, supporting this analysis with real-world examples, including a discussion of the LGBTQ+ vocabulary, Homosaurus and alternative terminology sources. The paper then proposes a flexible process for researching and developing custom taxonomies that draws on the strengths of existing vocabularies, knowledge of one's content and users, professional best practice literature, journalistic style guides, and first-person consultations. The paper argues that custom identity taxonomies are a way of showing respect for the people represented in one's digital asset management system and the people who use it. Thus, bringing the values of diversity, equity and inclusion into one's metadata are a small way of expressing and fostering a community of care.

KEYWORDS: taxonomy, DEI, controlled vocabulary, metadata, inclusion

INTRODUCTION

In the last few years, interest in and attention to issues of diversity, equity and inclusion (DEI) have surged in response to brutal police killings of Black people, anti-Asian hate, anti-transgender laws and the rolling back of reproductive rights in the USA. With divisions further exacerbated by a global pandemic, it has become glaringly apparent that some of our most revered and beloved institutions do not serve and represent us all equally and respectfully. In response, many organisations have initiated efforts to understand better and improve the ways in which they represent and address their constituencies across intersecting categories of gender, race, ethnicity, nationality, sexual orientation, disability, immigration status etc. (This interest is reflected in efforts like the San Francisco Museum of Modern

 © Henry Stewart Publications 2047-1300 (2023)

Art's 'Connecting People to the Art of Our Time',[1] in which they attempted to create a demographic picture of the artists represented in their collection, as well as staff, audiences and donors. In the archives space, there are multiple reparative and critical cataloguing initiatives, some of which are collected in the Society of American Archivists Inclusive Description Portal.[2] With respect to the DAM space, see 'Inclusive Metadata: Diversity, Equity & Inclusion in Digital Asset Management'[3])

Digital asset management is often instrumental in these efforts as it determines what materials and information are available for research, marketing, outreach, exhibitions and programming. The metadata associated with these assets act as a gateway to this content. If your metadata do not accurately represent — or perhaps do not represent at all — the people you are trying to reach, crafting communications and initiatives that appeal to them will be difficult. If you do not know what people and identities are represented in your digital asset management (DAM) system, you also cannot make effective decisions about how to grow and manage your collections.

As a metadata consultant focused on diversity, equity and inclusion, I work with organisations to bring their metadata in line with their goals for social justice and representation. One of the most common projects I work on is the development of custom taxonomies to describe people. This paper draws on my experience developing custom 'social identity' taxonomies or controlled vocabularies for different clients: a large, international media company, an encyclopaedic art museum and a university publisher. I will discuss some of the common challenges that arise when attempting to describe people according to socially prescribed categories and share lessons learned from balancing the desire to standardise terminology with the need to reflect and respect the nuances of human diversity. The results — custom taxonomies — are always a compromise, but they are a necessary tool for improving the description of people.

THE TROUBLE WITH TAXONOMY

First, it is important to acknowledge that when it comes to people, taxonomy itself is a problematic enterprise. Ancient class and caste distinctions, the invention of the concept of 'race', any system that creates an 'us' and a 'them' that holds people at arm's length because they look/act/believe differently, is the result of a hierarchical organisation of knowledge. And that is the definition of taxonomy. It is not difficult to understand how applying a system of classification to groups of people or individuals can be reductive and dehumanising.

The gold standard would be, of course, to represent everyone exactly as they would like to be represented. Gender identity in particular has seen a profusion of terms in addition to 'man', 'woman' and 'nonbinary:' 'agender', 'bigender', 'demigender', 'genderfluid', 'genderqueer', 'pangender', 'trigender' etc. There never has been and never will be a taxonomy complete and nuanced enough to reflect the myriad and continually evolving diversity of people. Such a taxonomy would be like the map in the Jorge Luis Borges short story, 'On Exactitude in Science'.[4] This fable describes a society obsessed with perfection. Its cartographers produce a map so precise and so detailed that it is the exact same size as the territory it describes: it is impossible to wield and thoroughly useless. A taxonomy of similar scope and detail — one that describes everyone exactly as they are — would be just as futile.

Taxonomy is necessarily reductive, but it is a necessary abstraction. If we do not have language to describe groups of people, particularly those who experience oppression or disenfranchisement, those people do not get represented

or accounted for. When people are not represented, they have more difficulty accessing resources and getting their needs met. For example, it is important to know how many women are represented in your DAM system, because even though women are not a minority, they have historically been underrepresented in many spheres of public life. Redressing that underrepresentation requires a benchmark so you can track improvements (or regressions) over time. Although group identification has often been a tool of discrimination and prejudice, it is necessary to ensure representation. We cannot improve people's lives if we do not acknowledge their existence.

That said, the terminology we use to describe groups of people has evolved over time and will continue to evolve. For example, to describe Black people in the USA, we have moved from old-fashioned and sometimes derogatory terms like 'Colored' and 'Negro', to 'Black' and 'Afro-American', then to 'African American' and now back to 'Black'. In the area of disability, we have moved from the derogatory 'cripple', to 'handicapped person', to 'person with disabilities' and from 'person confined to a wheelchair', to 'wheelchair user'. In terms of sexual orientation, the old dichotomy between 'heterosexual' and 'gay' has been thoroughly exploded to include myriad terms for diverse desires and identities. Any taxonomy that attempts to capture these nuances must therefore be a living document subject to regular reassessment and revision. Building DEI principles into custom taxonomies is by definition an ongoing process rather than a destination.

VOCABULARY WOES

If the previous arguments have not convinced you that you need a custom taxonomy, look at the state of existing vocabularies that are widely in use. Perhaps one day there will be a publicly available, broadly supported, linked open data taxonomy that will represent all facets of human identity with respect and accuracy. In the meantime, we are stuck with a patchwork of existing vocabularies, each of which has strengths and weaknesses. Although no vocabulary is perfect, you do not have to start from scratch in building your custom, DEI-informed taxonomy.

Existing, publicly available controlled vocabularies and taxonomies like the Library of Congress Subject Headings (LCSH) and the Getty Art & Architecture Thesaurus (AAT) are great for data aggregation — they make it possible to group similar things or people together under stable, commonly understood concepts. However, they are huge and often slow to change. When it comes to describing humans, they do not necessarily reflect the terms people use today to identify themselves, and changing a term can take a long time. For example, the campaign to change the offensive LCSH term 'illegal aliens' began in earnest in 2014, but the Library of Congress did not officially change it until 2021 (replacing it with the not entirely satisfying 'noncitizens' and 'unauthorized immigration').[5]

The Library of Congress vocabularies also do not always contain the historical terms for people required by many museum or archival collections. Last year, I developed a guideline for describing historical persons at an encyclopaedic museum of art in the US Midwest. The cataloguers wanted to accurately account for the nationalities of artists and subjects like Michelangelo and Roman Emperor Trajan. Today we consider Michelangelo an 'Italian' artist, but in his time, he would likely have identified as 'Florentine'. Trajan was born in what is now Spain, but of course was 'Roman', and no one would ever describe him as 'Spanish'. How we describe individual people differs depending on who they are, and it has as much to do with the moment in which they lived as it does with the one in which we live now.

These are the kinds of nuances that a vocabulary like the LCSH, which, for the most part contains only present-day nationalities and ethnicities, is not very good at. By contrast, the AAT has a whole branch for 'Roman (ancient Italian culture or period)', including 'Imperial (Roman)' and 'Trajanic'. At the museum, we decided to use the AAT for the nationalities and ethnicities of all artists and subjects, as it includes 'Florentine' among others. Using these terms to describe people is a slight misuse of the AAT, which intends these terms to describe objects rather than people, but it is better than coming up with a list of such terms from scratch. The AAT is sourced from and widely used by museums, so we felt comfortable with this slight tweak in meaning.

The AAT also has a decent selection of terms for gender and sexual orientation, but you may want to consider using a resource like Homosaurus instead. Created by LGBTQ+ information professionals, it is more detailed and more aligned with current terminology because it is solely focused on describing gender, sexual orientation and related concepts. It is also structured as linked data with persistent URIs, so even if terms change, the concepts remain stable.

When it comes to racial categories, AAT and LCSH are not great. Given, race has no basis in biology and was debunked as a system of scientific classification long ago, but it still has a persistent impact on people's everyday lives, so it is important to track it. Here, the Library of Congress Demographic Group Terms (LCDGT) are much more helpful. They seem to be more up to date than the Subject Headings, which are still plagued by old-fashioned terms like 'Indians of North America'. By contrast, LCDGT has 'Indigenous people of America' which aligns better with the way language is moving in the present to describe the original peoples of most parts of the world.

Another dimension of identity that may be important to capture is disability. Not surprisingly, the AAT has very spotty and sometimes offensive terms related to disabilities (such as 'insanity'[6]). Including disability terms in a taxonomy requires nuance because most of the sources for terminology to describe various health-related conditions are medical in nature. For example, the LCSH and the National Library of Medicine's Medical Subject Headings (MeSH) contain lots of very specific terms for various conditions. However, many of these terms cast disability in a negative light, using words like 'disorder' and 'disease'. Although people with disabilities may identify with these terms, the disability rights community often prefers a social rather than medical model of disability. In the social model, people experience disability because society does not fully support their functioning in the world, not because they have certain conditions.[7]

This distinction came up in a recent project for an academic publisher. In that project, we were describing people who were the subjects of case studies. As the authors of the case studies were working with the subjects themselves, the project represented a rare opportunity to collect social identity information directly from the subjects. We wanted to present these people with choices that reflected how they see themselves. But an attempt to represent all major disabling conditions quickly turned into something that looked more like a list of medical diagnoses than identity descriptors. With feedback from a disability consultant, we reworked the taxonomy to include only broad categories: cognitive, developmental, physical, psychiatric, sensory and speech. This felt a lot better and would allow people to indicate what type of disability they experience without having to share their specific medical history. Running alongside this part of the taxonomy were separate terms 'neurodivergent' and 'neurotypical', acknowledging that certain neurological states previously considered disabilities are now attributed to neurodiversity and are not necessarily disabling. And, of course,

the taxonomy included options for 'Other' free-text entry and 'Decline to state'.

All of the major vocabularies are pretty tied to the medical model of disability, so the terms we used for our broader categories were sourced from Wikidata, the crowd-sourced controlled vocabulary for nearly everything on the internet. Wikidata is vast and unruly, but it also provides a great opportunity to share what would otherwise be siloed attempts at custom taxonomies. Anyone can create and edit entries for terms on Wikidata, so if you do not find a term you need, you can add it. You can also add synonyms to existing records, which enables you to use a local version of a term while still linking to the record for the same concept on Wikidata. The main issue with Wikidata for taxonomy is that its hierarchies are unreliable. Different editors understand the relationships between Wikidata records in different ways, so there is no consistent hierarchy. For example, although both are subclasses of 'disability' (Wikidata ID: Q12131), 'physical disability' (Wikidata ID: Q1179623) is described as a 'health problem' whereas 'developmental disability' (Wikidata ID: Q1142806) is called a 'disease'. Wikidata is a great option for providing a persistent identifier for the exact term you need, but it will not necessarily help to create the right relationships between terms.

Although there are many problems with existing controlled vocabularies, changes are afoot. The Respectful Terminology Project of the National Indigenous Knowledge & Language Alliance received seed funding in March 2023 to create an open, online platform for preferred vocabulary describing Indigenous people, places, heritage, traditions, knowledge and culture.[8] And in February of the same year, the Subject Analysis Committee of the American Library Association issued a report on creating an external review board and process for improving the editorial process of the Library of Congress vocabularies.[9] Change is slow, but it is happening.

THE PROCESS

You do not have to wait for the controlled vocabularies to catch up to start doing a better job of representing the people associated with your collections. As the above section demonstrates, custom taxonomy does not have to be original — it can be a patchwork of terms sourced from different places. Of course, this can complicate maintaining and updating vocabularies in your DAM system, but it is necessary to represent people appropriately. If nothing else, the resurgence of interest in DEI issues has foregrounded a fact of library science we have not always confronted: the process of creating, reviewing and revising descriptive terminology is an ongoing, evolving process. The idea of a single vocabulary that meets all of our needs is a fiction of the past.

But if you are picking and choosing among different terms, how do you know which terms to use? This is where research comes in. First, it is important to understand the nature of your collections: Who is represented within them? Who is likely to be? And who are your prospective audiences? Metadata form the bridge between assets and users — how can the terminology you use most accurately and respectfully present the people represented in your assets to the audiences you are trying to reach? How will they recognise themselves in your content? If you can answer these questions, at least partially, that will provide a good idea of where to focus when developing your taxonomy.

Once you have identified the groups of people involved, do some research on issues of representation that affect those groups and the ways in which people describe themselves. Scholarly research can be useful, in the form of journal articles or books, but if you do not have easy access to scholarly databases, this avenue can be time-consuming and expensive. Within the archives community, there is a growing library of free resources published about reparative description and inclusive metadata.

The Society of American Archivists (SAA) recently endorsed 'Archives for Black Lives in Philadelphia' as an outside standard,[10] and the SAA's Description Section has an Inclusive Description Portal with links to other free case studies, guidelines and articles.[11,12]

Another research option is to look at journalistic style guides. These guides are published by different organisations and affinity communities to instruct journalists on how to write about people from a given community. For example, the GLAAD Media Reference Guide[13] provides guidance on how to describe LGBTQ+ individuals; the National Center on Disability and Journalism publishes the 'Disability Language Style Guide' with a glossary of terms to use (and not use) when describing people with disabilities[14]; and the National Association of Hispanic Journalists has created a 'Cultural Competence Handbook' as a guide to representing people of Latin American descent.[15] In the UK, meanwhile, Carissa Chew and the National Library of Scotland have created a shared Google Drive called the 'Inclusive Terminology Glossary' with research on historical and contemporary terms across multiple categories.[16] These are just a sample of the resources available; there are style guides for just about any community identity you can think of. Although these guides probably cannot answer every terminology question you may have, they are more up to date than most controlled vocabularies and provide a relatively quick way to 'take the temperature' of issues around a particular term or community. Be warned however, that these guides do not always agree with one another. It is important to look at more than one and to understand the background, purpose and perspective of the organisations that author them.

Finally, if you have the time and resources, the best way to find out what terms people prefer is to ask them. This could take the form of casual one-on-one conversations, focus groups, or perhaps an online survey. In a corporate setting, you might be able to collaborate with employee resource groups that represent various communities and have a vested interest in DEI-related initiatives. In an academic setting, you might be able to take your questions to a professor or class focused on the group you are describing. For museums, it might mean consulting a curator who has worked with specific communities or collaborating with the education department to poll docents or visitors. In a taxonomy I designed for a large international media organisation, I was struggling with what terms to use for the Indian caste system as a dimension of people's socioeconomic background. They were able to connect me with a co-worker in India who explained that the terms I was working with were for the Hindu castes only. Operating from my limited Western perspective, I did not realise that there are different caste structures and terms for Muslim and other ethnic and religious communities in India. This one conversation led us to shift our emphasis away from enumerating all the castes, to a higher-level view indicating the difference between the so-called 'Forward' or 'General' castes, which are considered economically advantaged, and the 'Backward' or less advantaged castes. Although these terms sound rather harsh, they are in fact the terms used by the Indian government, with members of 'Backward' castes being eligible for specific forms of government assistance. As these terms are in wide use in India, we decided to go with them, with the caveat that we would revisit them in the future.

In any case, if you are asking for the time of people from marginalised backgrounds, it is important to offer some kind of compensation, preferably monetary, in exchange for their time and expertise. In making our descriptive tools more respectful and accurate, it is important not to replicate relationships of exploitation in sourcing this information.

A COMMUNITY OF CARE

It is important to compensate people for their time because, in the end, we are trying to build a system of reciprocity and care. It is important to bring DEI principles into the management of DAM systems not only because it is the right thing to do, but because it is an expression of care for the people in our community, whether they are designers, students, cataloguers, photographers, museum visitors or anyone else who interacts with the DAM system and its assets. Too often we get carried away with all the things we need to do with assets — ingesting, documenting, transforming, migrating, downloading — that we forget about the people on the other end. A single, carefully chosen term can say, 'I see you', or 'You are welcome here'. It may sound corny, but it can be the difference between a person feeling erased or disrespected and feeling seen and included. We have spent so much time investing in processes and systems that separate us from one another and put efficiency and productivity ahead of everything else. In the name of efficiency and productivity, we have excluded or overlooked certain groups of people, leaving them out of the collective narrative or historical record. Creating a DEI-informed custom taxonomy is only one part of a larger effort to make our institutions more equitable and inclusive, but it is a step in the right direction.

References

(1) San Francisco Museum of Modern Art, (2022), 'Connecting people to the art of our time', available at: https://www.sfmoma.org/connecting-people-to-the-art-of-our-time/ (accessed 15th July, 2023).
(2) Society of American Archivists, (2023), 'Inclusive description portal', available at: https://www2.archivists.org/groups/description-section/inclusive-description (accessed 15th July, 2023).
(3) Mizota, S., (2023), 'Inclusive metadata: Diversity, equity & inclusion in digital asset management', available at: https://tenovos.com/resources/guides-dei-inclusive-metadata/ (accessed 15th July, 2023).
(4) Borges, J. L., (1946), 'On exactitude in science', available at: https://kwarc.info/teaching/TDM/Borges.pdf (accessed 23rd March, 2023).
(5) Dartmouth Library, (2019), 'Change the subject: Dartmouth Library's statement on potentially harmful content', available at: https://www.library.dartmouth.edu/digital/digital-collections/change-the-subject (accessed 15th March, 2023).
(6) J. Paul Getty Trust, (2004), 'Insanity', Getty Art and Architecture Thesaurus Online, available at: http://vocab.getty.edu/page/aat/300055132 (accessed 15th July, 2023).
(7) Buder, S. and Perry, R., (2023), 'The social model of disability explained', available at: https://www.thesocialcreatures.org/thecreaturetimes/the-social-model-of-disability (accessed 23rd March, 2023).
(8) National Indigenous Knowledge & Language Alliance, (2023), 'NIKLA Projects', available at: https://www.nikla-ancla.com/projects (accessed 15th July, 2023).
(9) Cataloging Lab, (2023), 'SAC Working Group on external review of LC vocabularies', available at: https://cataloginglab.org/externalreview/ (accessed 15th July, 2023).
(10) Society of American Archivists, (2022), 'Archives for Black Lives in Philadelphia (A4BLiP) Anti-Racist Description Resources', available at: https://www2.archivists.org/standards/archives-for-black-lives-in-philadelphia-a4blip-anti-racist-description-resources (accessed 15th July, 2023).
(11) Society of American Archivists, ref. 2 above.
(12) Mizota, S., (2023), 'Free resources about bias and metadata', available at: https://www.sharonmizota.com/metadata/resources/ (accessed 15th July, 2023).
(13) GLAAD, (2020), 'GLAAD Media Reference Guide — 11th Edition', available at: https://www.glaad.org/reference (accessed 15th July, 2023).
(14) National Center on Disability and Journalism, (2021), 'Disability language style guide', available at: https://ncdj.org/style-guide/ (accessed 15th July, 2023).
(15) National Association of Hispanic Journalists, (2020), 'Cultural competence handbook', available at: https://nahj.org/wp-content/uploads/2020/08/NAHJ-Cultural-Competence-Handbook.pdf (accessed 15th July, 2023).
(16) Chew, C., (2023), 'Inclusive terminology glossary', available at: https://drive.google.com/drive/folders/1JlZG0zmzlzPauwqJ5JxxUajf5hYkD0ta?usp=share_link (accessed 15th July, 2023).

 © Henry Stewart Publications 2047-1300 (2023)