MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Exploring the Potential of Digital and Computational Tools for Analysis, Research and Visualization of Art Collections: A Case Study of MoMA's Datasets"

verfasst von / submitted by

Nora Linser BA MA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von / Supervisor:	Rebecca Josina Kahn, PhD

Abstract

This thesis demonstrates that exploring cultural heritage collections using data analysis methods offers new insights, reveals connections, and allows the identification of gaps, limitations, and absences. By utilizing two publicly available datasets from the Museum of Modern Art in New York (MoMA), this study shows the fruitfulness of data analysis for larger art collections, providing a deeper understanding of the collections and the artworks within them, but also the institution’s internal mechanics on how they record and organize objects and the knowledge about them. The thesis addresses the importance of MoMA’s decision to publish parts of its database, which empowers researchers to independently investigate the museum’s collection and to uncover new insights by the use of data analysis methods. This research contributes to disciplines focused on cultural heritage collections, particularly art history, collection studies and digital humanities.

Table of Contents

1 Introduction

Art museum collections – like that of the Museum of Modern Art in New York (MoMA) – are large. Spreading over numerous storage rooms and exhibition halls, the versatile collections of MoMA are home to thousands of diverse objects from various time periods; all of which are distinct in size, display different meanings, and are made from different materials using different techniques. To study every single object and to contextualize them within the collection would fill the working live of multiple researchers. Therefore, how should a large collection, such as the one of MoMA, be examined and investigated?

Typically, this has been done by exploring single objects or subgroups and inferring them to the entire collection, combined with the selection of significant events and persons within the collection's history that are chosen to build up the narrative of the museum. This approach has provided insightful and enlightening results, but it comes with a cost. Selecting objects, events, and people, is always accompanied by the exclusion of the remaining majority of unconsidered objects, events, and people, which are ultimately not part of the story that is told. By implementing data analysis, this thesis proposes an alternative approach to investigating cultural heritage collections, in particular modern art museums. Specifically, two datasets (published by MoMA) will be investigated in this study, with an aim to show that data analysis is a useful tool to examine collections and the institutions responsible for them. The goal of this study is to demonstrate that this approach brings novel insights and prompts the establishment of new research questions and directions.

Today museums manage their collections digitally. By using databases and software, they create plentiful data, multifaceted in formats and structures. This form of storing knowledge about objects, their creators, and the processes that are involved in museums' daily work in relational databases (which is the most often used database structure), is relatively new to the community. The tasks which come with this development, from the analogue filing systems to digital databases, require new sets of skills that are not (yet) commonly part of the education of museum staff, researchers, and scholars. Due to this lack of expertise in the handling of data, many advantages and possible insights are overlooked and/or unknown. This thesis aims to showcase some of those advantages and endeavors to shed light on some insights that are only to be found if this distant reading approach is used to analyze collection. Like one of the most famous artworks in MoMA's collection, Claude Monet's "Water Lilies", a certain distance is helpful to not only see the bigger picture through single brushstrokes, but to also see the appearing holistic image that emerges when the view shifts from the detail to the whole. Nevertheless, it is the combination of distant perception of togetherness with a close-up examination of details that allows for a holistic understanding of Monet's painting, but also of datasets.[1]

This thesis begs the question of whether implementing data analysis along with basic knowledge about the collection as a methodology, enables us to retrieve novel insights. The aim is to not only find answers to common questions regarding the collection and the institution, but finding new directions to explore collections through the examination of their datasets. In this thesis, MoMA's datasets will be examined from two viewpoints in order to achieve this. The first viewpoint examines the collection and summarizes information on the objects within it. The second viewpoint sheds light on the dataset structure itself and possible understandings it may provide on the cataloguing practice of the museum. Data can also never be considered neutral or objective; therefore, part of this thesis will discuss how the creation of data always brings biases with it, even if not intended. Bringing possible limitations and constraints within cultural collection datasets to the attention is another key aim of this thesis. With this, the purpose of the thesis lies in the investigation of available digital methods and their applicability for the fields of digital humanities, art history, and collection studies in particular, but is applicable to further humanities disciplines interested in analyzing datasets. It also aims to showcase that accessing information in datasets can be done with off-the-shelf software and that with the support of this software, almost everyone interested is able to work with cultural collection data.

[1] See https://www.moma.org/collection/works/80220 to examine Monet's artwork in the online collection of MoMA (accessed on 27.7.2023 15:25).

The thesis is structured in two main parts. The first part presents the context for the case study that is conducted in the second part, which is comprised of a literature review married with personal experiences gained from working in a museum and the cultural collection field. A discussion about what the term data infers, how it is understood, and which precautions are to take when working with it, comprises the first chapter and is the entry point into this thesis. This is followed by a short chapter in which the importance of cultural heritage objects, in particular art objects, for the construction of history and memory is put forward. This significance is what makes the museums, archives and libraries the powerful institutions they are – not only in their field but reaching far beyond into politics and society. The third chapter of the first part is dedicated to providing a brief overview of the history of collection and collection catalogues. The structures that were established, along with large collections to manage them, are of particular interest here, since those originating structures build up cataloguing designs today. This is also the place where contemporary collection management systems, international frameworks, and guidelines on how to record cultural heritage objects are introduced. Some terminology and basic principles of databases will be explained here, too. In chapter four of part one, a very important topic is discussed: potential biases within cultural heritage collections, their records, and the authority and power those institutions hold. At the beginning of this chapter, the fondness of the art history discipline for categories is put forward, with an aim to show how deep the Western idea of how to classify objects is engraved in contemporary cataloguing systems. The following part is more practical, displaying examples of possible biases that might be found in collection data and where attention and caution are necessary. The last parts of the chapter provide an optimistic approach on how the biases and limitations are to be tackled, showing that there are ways on how to repair some of the damages that were done.

The second part of the thesis explores the case study of the two MoMA datasets the museums published on their GitHub page. Before the actual data analysis is conducted context information on the museum, though, its history and available information regarding its cataloguing practice are introduced. The collection management system MoMA uses, TMS, will be introduced, followed by a short overview of what the museum publishes on their online collection. Once the available datasets are examined, this section will be of special interest to compare its content with what MoMA published on their website. Prior to the analysis of the data, the documentation on the datasets, in this case consisting solemnly of a short “readme” file on GitHub, will be analyzed. The main part of this thesis section is chapter three: the exploratory data analysis (EDA). This analysis is split into two stages; the first one is the data understanding part, where each available feature in the two datasets is examined and statistical measures are calculated. This is done with the aim to understand what kind of data, in which format, and in which quality the datasets hold. After this step, a summary of what is not included in the datasets is provided, in addition to a comparison of the data MoMA publishes on their web page in their online collection. The second part of the EDA is done in a software called *Tableau.* Topics of interest are investigated through the use of visualizations which allow to explain the data and the findings. At first, the dataset of the artists will be analyzed in more detail, answering questions about the gender, the age, and the nationality the artists identify with, followed by a section in which the artwork records and how the objects are categorized and described is examined. The heart of this section builds on the acquisition history of the museum, showing what was acquired, when and by whom. Coupled with a general overview of the acquisitions, the data will also be separated by each director’s tenure, showing how they influenced the acquisitions. Based on the assumption that historical developments might be distinguishable in datasets like the ones of MoMA, the subsequent section of the EDA will try to find hints or traces of historical global events, such as the Cold War, or more MoMA specific events, such as the donation of an important collection by Lillie P. Bliss or the protests against the underrepresentation of female artists in the 1980s, for example. A summary of the analysis results will round up the second part.

The thesis employs distinct methodologies for its two primary components. The first part, which pertains to the theoretical foundation of collection studies and cataloguing practices, relies primarily on reading and extensive literature review. Key literature is Lisa Gitelman’s “Raw Data is an Oxymoron” of 2013, providing important knowledge on data and imbedded meanings. Johann Drucker’s “Digital Humanties Coursebook” provides an overview on pretty much every topic a digital humanities person might be interested; it is a treasured resource to understand general concepts and ideas of the discipline. Concerning the cataloguing of objects and possible limitations, Blagoy Blagoev, Sebastian Felten and Rebecca Kahn offer in their article on the “Career of a Catalogue” valuable insights in the practice of record keeping in the British Museum.[2] Anthony Griffiths shares his insights on the same museums, giving a unique insight in how large museums operate and they record their collection.[3] With regard to MoMA, Sybil Kantor’s book on Alfred Barr and his legacy is to mention, along with Domínguez Rubio’s fascinating article on the preservation of “docile and unruly” objects within the MoMA collection.[4] Finally, Sandra Zalman shares her insights on the cataloguing practice of the museum, allowing us to get some information on how it conducts this task.[5]

Additionally, the incorporation of personal experiences gained from working in the museum domain and engaging in digital collection management is integrated using autoethnography. This approach is inspired by scholars such as Haidy Geismar, who extensively researches her own interactions with digital objects in various collections.[6]

In the second part of the thesis – the case study on the museum’s datasets – data analysis will be the main method. Summary statistics will be used, in particular, to understand data, gain initial insights into the datasets, and find correlations and dependencies within them. Exploratory data analysis will then be used to further investigate the datasets, the results of which will be visualized with the software *Tableau*, following the methods and principles established by Tamara Munzner.[7]
In art history the method of close reading is the most common approach to investigate object collections, picking specific objects and creating detailed and

[2] Blagoy *Blagoev*, Sebastian *Felten*, Rebecca *Kahn*, The Career of a Catalogue: Organizational Memory, Materiality and the Dual Nature of the Past at the British Museum (1970–Today), Organization Studies 39, no. 12 (12/2018) 1757–1783, doi:10.1177/0170840618789189.

[3] Antony *Griffiths*, Collections Online: The Experience of the British Museum, Master Drawings 48, no. 3 (2010) 356–367.

[4] Sybil *Kantor*, Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art, 2002; Fernando *Domínguez Rubio*, Preserving the Unpreservable: Docile and Unruly Objects at MoMA, Theory and Society 43, no. 6 (11/2014) 617–645, doi:10.1007/s11186-014-9233-4.

[5] Sandra *Zalman*, Unpacking the MoMA Myth: Modernism under Revision, Modernism/Modernity 29, no. 2 (04/2022) 283–306, doi:10.1353/mod.2022.0009.

[6] Haidy *Geismar*, Museum Object Lessons for the Digital Age (2018), doi:10.14324/111.9781787352810.

[7] Tamara *Munzner*, Visualization Analysis and Design, A.K. Peters Visualization Series (Boca Raton 2015); Christian *Chabot*, Chris *Stolte*, Andrew *Beers*, Pat *Hanrahan*, Tableau (2003).

focused examinations of those objects. This thesis will provide a different approach, that is not focused on the individual object but takes a more distant look at the complete collection in order to get insights. This so-called distant reading approach, which was initially established for literary studies by Franco Moretti[8], analyses collections of objects using computational methods with the aim to uncover patterns, trends, and broader insights that cannot be read with a close reading approach. Distant reading allows scholars and researchers to make data-driven observations, lead to the creation of new research questions, and as Johanna Drucker hopes, it "might shift focus away from the established canon"[9] and guide the researchers to areas where then close reading can be applied to closer examine the newfound insights.

The two approaches work best when they are used in combination, each of them leading to points of interest. In this thesis they will overlap in the instances where single datapoints are picked up to closer examine them and their significance for the collection as a whole.

[8] Franco *Moretti*, Conjectures on World Literatur, New Left Review 1, no. 1 (2000) 54–68.

[9] Johanna *Drucker*, The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Scholarship, 1st ed. (First edition. | Abingdon, Oxon ; New York : Routledge/Taylor & Francis, 2021. 2021) 114, doi:10.4324/9781003106531.

2 Part I: From the object to the object record

This chapter comprises several sections that aim to establish the theoretical groundwork for the subsequent case study presented in the second part of the thesis. To fully comprehend the case study, it is essential to first construct a clear understanding of the concept of "data" and its associated implications. Therefore, the initial segment of this chapter delves deeper into the multifaceted nature of the term "data" and explores its significance within contemporary societies. By examining the various connotations and contexts in which the term is employed, we can gain a comprehensive understanding of its meaning and relevance in our societal framework.

Taking this as the foundation, the following sections will discuss the importance of material (and immaterial) objects as representatives of the past, and their use as hooks for narratives and rhetorical interpretations of the past: the history. It will be shown how powerful museums are in their role of creating this narrative by deciding which objects should be part of it and how they are placed within it. A brief history of the establishment of modern collections and museums and how they gained authority and power will be discussed. With a focus on how the roles of the museums shifted over time and how those shifts are apparent in the catalogues of the museums. The cataloguing practices of museums will be examined in detail, too, with the introduction of different frameworks and guidelines. This chapter will discuss the trade-off between systematizing object records to facilitate easier retrieval, and maintaining the uniqueness and descriptive meaning of objects, which will be one of the main topics. This discussion will lead to the final part of the chapter, where limitations in current cataloguing practices, systematic misrepresentation, and absences will be explored. This chapter will demonstrate that the tendency to structure objects within predefined schemas is inherent in art history, and the structures established with the emergence of the discipline continue to be the normative way of discussing art objects, often disregarding anything that does not align with these established lines. The end of the chapter will focus on the challenges these limitations pose for today's researchers, cataloguers, and museum staff in general. It will conclude with the introduction of the "Collection as Data" principles, which serve as a guide for addressing the issues associated with cultural data.

2.1 What are Data?

A look at the etymology of the word data provides a starting point for the discussion on how the term data is used today, what implications accompany it, and how this affects the critical work with data. In his article titled, “Data before the fact” in the insightful publication “‘Raw Data’ is an Oxymoron”, Daniel Rosenberg defines, next to the etymology of the word also those of the term “fact” and “evidence”, which are often used together or synonymously and summarizes:

> *“…facts are ontological, evidence is epistemological, data is rhetorical. A datum may also be a fact, just as a fact may be evidence. But, from its first vernacular formulation, the existence of a datum has been independent of any consideration of corresponding ontological truth. When a fact is proven false, it ceases to be a fact. False data is data nonetheless.”*[10]

The takeaway point here is that data is not inherently true; instead, data can be false and misleading. This lies in the way data is created. Data is not just there, it is always constructed and collected, it is nothing that is “given”.[11]

Before looking at a small example of how data is constructed in the Digital Humanities (DH), we should define what DH researchers mean when they talk about data. In its broadest sense, it refers to any digital information – regardless of its size – that can be computationally processed. This incorporates a wide range of content, including sounds, graphs, texts, images, graphs, etc.. It can exist in various file formats, both structured and unstructured.

A manuscript laying on the table in front of us is not considered data, but as soon as we make reproductive images of the pages and upload them to our machines, we have image data (and metadata of the images produced by the camera). When we then extract the text from the images to a text file (manually or

[10] Daniel Rosenberg, ‘Data Before the Fact’, in *‘Raw Data’ Is an Oxymoron*, ed. by Lisa Gitelman, Infrastructures Series (Cambridge, Massachusetts ; London, England: The MIT Press, 2013), pp. 16–40 (p. 18).

[11] *Drucker*, The Digital Humanities Coursebook, 25.

computationally assisted), we have textual data. If we start counting pages, calculating word counts or sentences in a chapter, we also hold numeric data that can be statistically analyzed. This could be stored in a simple excel sheet, in a .csv file, in a Pandas Dataframe or countless other file formats. On the way to this file – that might be called a "dataset" – a line of decisions was made that might not be obvious when only looking at the "finished" dataset of word counts and average sentence count. Do we photograph the cover page and the first pages, even when there is no or very little text that is not part of the text we are interested in? Do we transcribe the words as they are written, or do we correct spelling errors and typos? What if someone wrote marginals in the book, do we add them as well, and how would we mark them? Do we count stop words? How do we account for abbreviations? These questions and decisions should be part of the documentation of the dataset. Otherwise, all those decisions are obscured, hidden and might not be reproduceable or comprehensible. This is why detailed documentation of datasets and its provenance is crucial and should be considered a main component of all datasets.[12]

Although datasets appear objective, this example shows that datasets are, in fact, not, even within datasets that are typically considered to be more scientific or factual compared to humanist data. Matthew Stanley sheds light on this in his entertaining article titled "Where Is That Moon, Anyway?". Throughout, he delves into the complex challenges involved in accurately determining the timing of historical solar eclipses, revealing that even those astrological datasets are not immune to subjective interpretations.[13]

The concept of the computing-specific term data dates back to the 20th century, but it roots go back longer. It arose with the development of modern concepts of knowledge production and argumentation in the seventeenth and

[12] See Leslie F. *Sikos*, Dean *Philp*, Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs, Data Science and Engineering 5, no. 3 (09/2020) 293–316, doi:10.1007/s41019-020-00118-0. for more information on the difficulties that come with provenance data.

[13] Matthew *Stanley*, Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations, In: "Raw Data" Is an Oxymoron, Lisa *Gitelman* Ed., Infrastructures Series (Cambridge, Massachusetts ; London, England 2013) 77–88.

eighteenth century.[14] Until the beginning of the nineteenth century, the connotation of data changed, while the meaning stayed the same. “It went from being reflexively associated with those things that are outside of any possible process of discovery to being the very paradigm of what one seeks through experiment and observation.”[15] Today, data is broadly assumed to be objective and reliable, to be raw and therefore the truth. It is understood as a “starting point for what we know, who we are, and how we communicate”.[16] It even adds to the argumentative context of every discipline when data is referenced as the starting point of analysis, research, or experiment.[17] But as shown above, these assumptions need to be critically evaluated and questioned. Lisa Gitelman and Virgina Jackson provide us with the helpful comparison of data and photography and how we need to deal with the assumed objectivity:

> *“The presumptive objectivity of the photographic image, like the presumptive rawness of data, seems necessary somehow resilient in common parlance, utile in commonsense—but it is not sufficient to the epistemic conditions that attend the uses and potential uses of photography. At the very least the photographic image is always framed, selected out of the profilmic experience in which the photographer stands, points, shoots. Data too need to be understood as framed and framing, understood, that is, according to the uses to which they are and can be put.”*[18]

It can therefore be stated that “Data has no truth”, as Daniel Rosenberg writes,“[i]t may be that the data we collect and transmit has no relation to truth or reality whatsoever beyond the reality that data helps us to construct.”[19]

[14] *Rosenberg*, Data Before the Fact, 15.

[15] Ibid., 36.

[16] Lisa *Gitelman*, Virginia *Jackson*, Introduction, In: “Raw Data” Is an Oxymoron, Lisa *Gitelman* Ed., Infrastructures Series (Cambridge, Massachusetts ; London, England 2013) 2.

[17] *Rosenberg*, Data Before the Fact, 20.

[18] *Gitelman*, *Jackson*, Introduction, 5.

[19] *Rosenberg*, Data Before the Fact, 37.

The datasets this thesis explores in Part II: Case Study of the Museum of Modern Art datasets are, as all other datasets, also framed and they frame. It was constructed following countless decisions, and cumulated and transformed by different people over a timespan of almost 100 years. It cannot be assumed in any way that it is an objective representation of the collections in the museums. But it can be critically evaluated as the results of how the museum operates and which decisions it made and makes.

2.2 The past, the history and the memory

Section 2.2. analyses the significance of material objects for the construction of past, history, and collective memory and how the description, presence, accessibility, and valuation of those objects matters. The role of the museum as an institution that processes cultural objects as a powerful player in those constructions will be examined.

The past, the history and the memory are mobile things that are constantly negotiated, transformed, and reweighted. The distinction between the terms is fruitful. The *past* is what has happened before today, regardless of how it is remembered and if at all, or not. The *history* is how those events are narrated; it is an “always rhetorical”[20] interpretation of the past. *Memory* is what the individual or the collective (collective memory) recollects of the past experiences and narrated histories.[21] As Blagoy Blagoev, Sebastian Felten and Rebecca Kahn point out, memory is continuously made and remade “by actively bringing the past into the present through narratives, rhetoric and symbols”.[22] It is not a “static repository of knowledge”, but a “collective interpretation of the past devised through practices of remembering”.[23] Material (or immaterial) objects “can not only enable but also actively shape processes of remembering”.[24] The objects are active players within

[20] *Blagoev*, *Felten*, *Kahn*, The Career of a Catalogue, 1759.

[21] Jan *Assmann*, Das kulturelle Gedächtnis: Schrift, Erinnerung und politische Identität in frühen Hochkulturen, 2., durchges. Aufl (München 1997).

[22] *Blagoev*, *Felten*, *Kahn*, The Career of a Catalogue, 1758.

[23] Ibid., 1760.

[24] Ibid.

this process, based on them the past is constructed.[25] Metaphorically, the objects form a container that can be filled with meaning, narration and belonging. They can provide connections to past events, people, cultures, or societies. As those representatives of the past, they are the basis for the negotiations of how history is narrated. And from there their assigned meaning transforms into the larger picture of collective memory.

The presence and accessibility of cultural objects are pivotal factors in shaping the collective memory and narrative. Museums are places where objects are collected and put into the narrative of history. Hence museums can be defined as places where memory is created and negotiated. They act as mediators between the past and the future by providing access to artefacts of human activities.[26] This powerful position also allows them to decide which context the objects should be placed in: "In forming collections, museums recontextualize objects: they remove them from their original contexts and place them in the new context of "the collection". This recontextualization of objects primarily in terms of other objects with which they are related, is a fundamental aspect of the kind of collecting legitimized by the museum.[27]

Be it in museums, archives, public spaces or in the digital space, the presence of objects and how they are described and put into narration is crucial for the collective memory. The selection of the objects and their contextualization profoundly affect the representation and interpretation of the past and form the history. This selection process is fraught with power, as it entails decisions that favour certain narratives, perspectives or historical events over others. Blagoev et al also point out that the selection of certain objects comes with absences and vacancies of other objects: "What is remembered is always highly selective in relation to everything that is simultaneously left out and, thus, forgotten."[28]

One other aspect that is important, especially for this thesis, is about how memory is recorded. Social memory is dependent not only on objects but also on the

[25] Ibid., 1761.

[26] Ibid., 1762–1763.

[27] Sharon *Macdonald*, Collecting Practices, In: A Companion to Museum Studies, Blackwell Companions in Cultural Studies (Malden, MA 2006) 82.

[28] *Blagoev*, *Felten*, *Kahn*, The Career of a Catalogue, 1779.

"technologies of memory"[29] that are used to not forget them. Many scholars have pointed out how the invention of new forms of those technologies, such as printed books, filing cabinets, or digital databases, impacted the social memory functions. Blagoev et al point out their significance:

> *"Importantly, such material technologies of memory […] are not merely containers for stories or external storage separate from the social practices of remembering. Instead, materiality and practice, in their entanglement, constitute memory."*[30]

They argue that those technologies of memory are a kind of repository of organizational memory and emphasize the dual nature of the past, viewed as both an active process of remembering (Gedächtnis) and as material technologies of memory (Speicher).[31]

For the purposes of this thesis, this viewpoint highlights not only the significance of critically examining what museums collect, but also how the objects that are collected are described, narrated, and catalogued, as well as how this knowledge is managed and stored.

The following sections of this chapter will take a closer look at the cataloguing and collecting practices that are implemented in museums' daily work and how they affect the objects meaning, value, and context.

2.3 Collecting and cataloguing cultural objects

Section 2.3. delves into a short exploration of the evolution of museums and collections. Before cataloguing practices and categorization issues are discussed, it is crucial to provide a concise overview of the origins and transformative journey of collections and museums throughout history. Thereby, we gain insights on how the museums got to the power they hold today and how old biases, limitations, and constraints are dragged into today's collecting and cataloguing practices. This

[29] Jeffrey K *Olick*, Collective Memory: The Two Cultures, American Sociological Association 17, no. 3 (1999) 333–348.

[30] *Blagoev*, *Felten*, *Kahn*, The Career of a Catalogue, 1761.

[31] Ibid., 1759.

knowledge will help in the analysis of the MoMA datasets and its contextualization that will follow below.

„Collecting is sometimes seen as a basic urge or instinct, and as a fundamental and universal human (and, indeed, sometimes also animal) activity.“[32] As Sharon MacDonald puts it, collecting is also a set of distinct – though also mutable and varying – practices that not only produces knowledge about objects but also constructs particular ways of knowing and understanding.[33] She continues: "Collecting […] should be seen as a practice in which the intention is to create a collection; and a collection in turn is a set of objects that forms some kind of meaningful though not necessarily (yet) complete 'whole'."[34] Sharon MacDonald adds to that that, even if it may seem redundant to define collecting as an activity centered around the aim to create a collection, for her this distinction highlights the specific object orientation of the task. Objects are not included because of their practical utility or individual significance, but to be part of a specific group of items.[35]

The collecting practice we have in mind when we talk about museums goes back to ancient time. There are recorded collections from Ancient Greece and Rome, from medieval Europe, China, or Japan. The practice grew during the Renaissance in Europe, when the basic framework of modern museum and collections was established.[36] Different forms of private or semi-private collection emerged, such as the "Studiolo" or later the "Kunst- and Wunderkammer". Power has always played an essential role in collecting and exhibiting practices. In the 17th century, it was the power and money which the royals had that allowed them to collect and then exhibit the things they had (to a small circle of people). Being in a position where there was (monetary but also time and intellectual wise) capacity for culture was enough to be

[32] *Macdonald*, Collecting Practices, 81.

[33] Ibid., 94.

[34] Ibid., 82.

[35] Ibid.

[36] Ibid., 83. McDonald provides detailed insights in the collection practices of the Renaissance.

a sign of power. Furthermore, collecting, as the results of these capacities, was the materialized proof of that power. [37]

Public museums were established in the 18th century as social spaces that had to overcome the earlier private, restricted, and socially exclusive boarders.[38] The transformation from former private collections to public ones came along with a decline of the importance to legitimate dynasties through symbols of power, as the move of the Medici collection to the public Uffizi Galleries can show.[39] But power was still the main motivator, especially as royal art collections were used to address the visitors as subjects of the monarchy. As a result, this showed the public that their place within the society was as subordinates.[40] When monarchies and dynasties were replaced by nation states, the narrative of the museums changed again, from representing the history of monarchies museums to now recounting the story of the new states and therefore legitimize them.[41] Sharon MacDonald argues that collections were important tools in the establishment of the new state form:

> *"Collections allowed nation-states to show their possession and mastery of the world – something that colonial powers were especially well able to demonstrate through the accumulation of material culture from the countries that they colonized … They also gave them the opportunity to amass and present evidence of their own pasts, so turning their histories into "objective" fact and legitimizing their right to exist."*[42]

At the same time the general understanding of the world shifted towards a narration where there is a straight line of development from "primitive" to "white, male and middle class", and where everything is connected and could be placed in a general

[37] Tony Bennett, *The Birth of the Museum: History, Theory, Politics*, Culture : Policies and Politics (London ; New York: Routledge, 1995), p. 21.

[38] Ibid., 26.

[39] Ibid., 27.

[40] Ibid., 36.

[41] Ibid., 38.

[42] *Macdonald*, Collecting Practices, 85.

ordering of things.[43] The museums took up those new understandings and displayed and illustrated the generalized, one-dimensional narration with the collection objects. With this unverbalized narrations emerging, the diversity of human experiences became overshadowed, leaving behind a singular dominant narrative.[44] Even though Tony Bennett states that the art museums were the only kind of museums that hold up a different role by continuing to display the "singularity of each objects and its power to dazzle"[45], it should be pointed out that within this uniqueness art museums also put every object in ordered groups and imagined straight lines of developments. The museums played a pedagogic role and predefined how exhibitions had to be read and understood. They went so far that even the walking path of the visitors was prefixed, reducing the mobility within the galleries to an one way walk. This ensured the full control over the narrative that should be taken away.[46]

Public museums were places for men. The involvement of women in public life was reduced largely in eighteenth and early nineteenth century and this also included the museums and art galleries. Throughout the history of museums, the accessibility for women transformed, from not being allowed to enter, to being allowed but not welcomed, to one of the few public spaces were women could safely go (accompanied).[47] It has been claimed that museums may, indeed, have been places that women were allowed to visit, but it will be examined in the case study below, if they truly were and are also places where women's work was and is valued, acquired and exhibited.

Access to museums and galleries is not only dependent of gender, but also based on class. Bennett suggests differentiating the different kind of collections based on who has access to them and who has the possibility to understand the meaning that is imbedded in the objects, their arrangement, and the narrative they

[43] *Bennett*, The Birth of the Museum, 39.

[44] Rebecca *Kahn*, Laura *Gibson*, Digital Museums in the 21st Century: Global Microphones or Universal Mufflers?, Museological Review, no. 20 (2016) 39–51.

[45] *Bennett*, The Birth of the Museum, 44.

[46] Ibid.

[47] Ibid., 29.

are telling. Following Pierre Bourdieu’s critique of the modern art gallery, Bennett states:

> *“The art gallery’s capacity to function as an instrument of social distinction depends on the fact that only those with the appropriate kinds of cultural capital can both see the paintings on display and see through them to perceive the hidden order of art which subtends their arrangement.”*[48]

Bennett concludes that museums have been co-opted by the social elites and serve as influential factors in the distinguishment of the elites from the general population. However, it is not accurate to describe museums purely differentiating. Instead, he suggests that their societal role is shaped by the contradiction of the differentiating force, on the one hand, and the homogenizing tendencies, on the other hand, resulting in an interplay between them.[49]

But what is a museum then when we put it in one sentence? There are probably as many definitions as there are museums in the world. But there is one that bridges to the next aspects that will be discussed: how those collections were and are managed. George Brown Goode, a 19th century zoologist and museums administrator, describes the museum as being the place of a “well-arranged collection of labels illustrated by specimens”.[50] Although Goode was referencing what we today may call a *natural history museum*, this definition brings forward crucial elements of museums. Firstly, it emphasizes the importance of ordering things systematically, something that will be discussed extensively below. Secondly, it emphasizes the significance of labels and the textual descriptions of things as integral components. Lastly it draws attention to the hierarchical positioning of labels above the objects themselves.

The significance of the label shows that objects are replaceable by something similar while the label can stay the same. This is probably truer for natural history museums where objects of the same species, for example, can be replaced without

[48] Ibid., 35.

[49] Ibid., 28.

[50] *Geismar*, Museum Object Lessons for the Digital Age, 11.

change in meaning, but this also goes for the general narration of art museums. Let us say we have an exhibition that focuses on 20th century sculpture, and we have a sculpture by Alberto Giacometti exhibited. The general story of development can be told even when we replace the original sculpture with a similar one. This goes even more so when we think about exhibition copies, facsimile, re-prints, or recasts. The story that is told is on the labels, the objects dissolve behind them. This also highlights the importance of labels – when they are not read, the narration of the exhibition is probably missed. What is also striking about Goode's quote is how early there was awareness of those topics within the museum world.

Following the discussion on collections and museums, it is essential for this thesis on MoMA's datasets, to delve into the topic of collection management. When an object enters a museum, it transforms and gains a new meaning and contextualization. Or as Chiara Zuanni puts it: "The musealization process transforms 'things' into 'objects' […], or better 'museum objects' ('musealia')".[51] Museum's primary tasks are, as also stated by ICOM, the International Council of Museums, to collecting, conserving, interpreting and exhibiting tangible and intangible heritage.[52] In order to do so, the first thing that is needed to successfully fulfil these tasks is the ability to recall the inventory of objects held by the museum, along with their acquisition details and, crucially, their current location. If this basic information is not available the museum is not able to fulfil its primary purpose, and as Blageov et al state, therefore jeopardizes its own existence.[53]

Museums (and Archives and Galleries and Libraries and so on) need a filing system that allows them to manage the large number of objects in their care. Basic lists and catalogues of collection items were a byproduct early one, so for example lists of the collections of Ferdinand II in Schloss Ambrass from the 16th century.[54]

[51] Chiara *Zuanni*, Theorizing Born Digital Objects: Museums and Contemporary Materialities, Museum and Society 19, no. 2 (07/30/2021) 186, doi:10.29311/mas.v19i2.3790.

[52] https://icom.museum/en/news/icom-approves-a-new-museum-definition/ (accessed on 23-06-2023 19:25)

[53] *Blagoev*, *Felten*, *Kahn*, The Career of a Catalogue, 1765.

[54] Sabine *Haag*, Veronika *Sandbichler*, *Kunsthistorisches Museum Wien*, *Schloss Ambras Innsbruck* eds., Ferdinand II: 450 Jahre Tiroler Landesfürst: Jubiläumsausstellung: eine Ausstellung des

Part I: From the object to the object record

The first analog databases of art objects arose together with the discipline of Art History, as Matthew Battles and Michael Maizels bring forward:[55] Art History has relied, from its early stages and in most cases continuing to this day, on the comparative analysis of images of artworks or objects. With the introduction of photographic prints and their accessibility at affordable prices, extensive image archives were established in the academic hubs of art history across Europe. These archives played a significant role in forming classifying systems and labeling schemes of photographic reproductions of artworks and images and location management. These structures are to an large extend still used by museums today. Haidy Geismar talks in detail about Aby Warburg's image archive and how this form of collecting created a new way of remembering: "His project pushed the poetics and the philosophical underpinnings of the picture library into new territory, recognizing the collection of images as the foundation for a new kind of knowledge practice and a new way to understand the ways in which images are embedded themselves within the reproduction of human culture."[56] She also mentions Andre Malraux's *Museeè imaginaire* and argues that his perspective of knowledge, which, in his understanding, emerges from juxtapositions and analogies made between images of objects, was internalized in digitization processes of today's online museum catalogues and how they provide the functionality to search, compare and find links between objects.[57] The idea of how to present objects through images and setting them within a grid next to each other was already blue printed by him in the 1960s and was picked up by the majority of online collection presentations.[58]

Kunsthistorischen Museums Wien in Kooperation mit der tschechischen Nationalgalerie und dem Institut für Kunstgeschichte der Akademie der Wissenschaften der Tschechischen Republik, 15. Juni bis 8. Oktober 2017 (Innsbruck 2017).

[55] Matthew *Battles*, Michael *Maizels*, Collections and/of Data: Art History and the Art Museum in the DH Mode Chapter Author(s): MATTHEW BATTLES and MICHAEL MAIZELS, In: Debates in the Digital Humanities 2016, Matthew K. *Gold*, Lauren F. *Klein* Ed. (2016) 326, doi:10.5749/j.ctt1cn6thb.

[56] *Geismar*, Museum Object Lessons for the Digital Age, 51.

[57] Haidy Geismar, *Museum Object Lessons for the Digital Age* (UCL Press, 2018), p. 52 <https://doi.org/10.14324/111.9781787352810>.

[58] See Windhager and others for more information on how cultural heritage data is displayed and visualized in the web. Florian *Windhager*, Paolo *Federico*, Gunther *Schreder*, Katrin *Glinka*, et al., Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges, IEEE

2.3.1 Categorization and Classification

This subsection delves into the examination of collection management systems, exploring how knowledge about object is transformed into text and other data formats, and subsequently standardized and uniformized to fit the structure of these systems. It highlights the significance of this process, while also looking at its inherent limitations, absences, and the drain of knowledge it entails. The chapter also provides a concise overview of taxonomies, thesauri and how they bring standardization into the cataloguing process, elucidating how they help to make records interoperable within one cataloguing system, and beyond. By introducing the most common frameworks and guidelines for museum object cataloguing, light will be shed on the challenges that come with the implementation of those frameworks in the day-to-day operations of museums, illustrated by a case study from the British Museums that showcases that ideal practices are not always achievable.
This understanding of terminology, collection management systems and frameworks will allow to contextualize the MoMA datasets and make the challenges and considerations that came along with its creation comprehensible.

As we have seen above is one of the key tasks of museums is to manage their objects and their whereabouts. This becomes a considerably challenging task as soon as the number of objects grows. Large museums, like The Victoria and Albert Museum, the Kunsthistorische Museum, the Smithsonian, or the Metropolitan Museum (even MoMA), each take care of millions of objects. Only knowing where each individual object is at any given time, distributed among various storage locations, exhibition halls, loans, restoration- and treatment workshops, is an enormous, logistical task. In order to manage this, a cataloguing system, where every object is represented with an individual record and where locations can be assigned to them, is necessary. Throughout the twentieth century, this was done in analogue filing systems. At the end of the century, they got replaced by digital databases. The scope of these systems broadened over time, from software that mirrored the filing cards, to very complex databases with user friendly interfaces and

Transactions on Visualization and Computer Graphics 25, no. 6 (06/01/2019) 2378, doi:10.1109/TVCG.2018.2830759.

many additional functionality.[59] The collection management systems (CMS) of the market leaders, (TMS Collections by Gallery Systems and Axiell Collections by Axiell) but also smaller local software providers, allows users to record very detailed information.[60] In standard systems fields to record detailed information about the provenance, acquisitions, creation and dating, external characteristics as dimensions, material and techniques, labels and markings, iconographic descriptions and associations of the objects are available. They also provide modules to record involved persons and institutions, relevant exhibitions, loans, transports, and publications, and all those modules can be intertwined and linked to the object records. Modules to manage images, videos, audio, or other media data are also available. Those CMS are based on relational databases, a system that was only invented 50 years ago, that allows to split information into separate tables and connect them with each other, and which allows us to distinguish information that varies from that which stays consistent.[61] The entry fields within the software interface can be of different kinds, so that the person entering data is limited in terms of what to enter, a date field for example can be configured to only accept numbers, a text field can have a character limit imbedded, enumerative fields allow only to select one term from a fixed list of values and some fields are linked fields, they retrieve information from other tables in the database. It is also possible to configure fields as mandatory that need to be filled in order to be able to save a record. These configurations inherent in collection management systems can significantly impact the data collection process and the resulting data. This is why consideration for how the original entry mask of features was configured might help in understanding the structure, format, or potential limitations of recorded information.

[59] Johanna Drucker, *The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Scholarship*, 1st edn (First edition. | Abingdon, Oxon ; New York : Routledge/Taylor & Francis, 2021.: Routledge, 2021), p. 75. <https://doi.org/10.4324/9781003106531>.

[60] https://www.gallerysystems.com/solutions/collections-management/ (accessed on 31.07.2023 09:50) and https://www.axiell.com/about-us/ (accessed on 31.07.2023 09:51). A list of collection management systems, validated by the UK collection trust can be viewed here: https://collectionstrust.org.uk/software/ (accessed on 24.06.2023 19:45)

[61] *Drucker*, The Digital Humanities Coursebook, 78–79.

Imagine your collection already has multiple objects of Merit Oppenheimer in the collection and your job is to catalogue a new acquired artwork of the artist. You create a new object record in the CMS and when you edit the creator field a list of all already existing creator records (in the people database that is an own but connected table in the relational database) is presented. You choose Merit Oppenheimer out of this list and save the record. With this, you not only entered the text “Merit Oppenheimer” to the creator field but made sure that the new object record is linked to the same person record the already existing artworks of Oppenheimer are linked to. When queering the database for artworks made by Oppenheimer, all objects, also the new one, are retrieved. Another huge advantage of this relational system is, that you only need to record static information on the person once. The birth and death dates of Oppenheimer are only recorded once in her person record. It is not necessary to manually add the dates to every object record (in case you wish to see them there), but you can easily display these dates as merged in fields from the linked person record within the object record. Another advantage of those linked tables: if you find a spelling error or someone changes his/her/their name, you can make the edit once in the person record, and every linked object record is automatically updated to the new version.

Having gained a general understanding of contemporary collection management systems, we can now explore the impact those systems make on how information of objects is recorded. We will also delve into the implications of describing cultural objects within a framework predominantly influenced by Western perspectives for Wester collections.

As we’ve observed above, cultural objects play a vital role in shaping our understanding of history and the collective memory. It is imperative to recognize that the manner in which we describe, categorize, and group those objects contributes to constructing of the overreaching narrative of history. Classification systems construct knowledge, and they dictate the structure within which this knowledge needs to fit in order be recognized.[62] Johanna Drucker points out how cautious we therefore need to be with those systems:

[62] Ibid., 57.

> *"We use classification systems to identify and sort, but also to create models of knowledge. Knowledge models expose and embody cultural differences and values. These are implied in every act of naming or organizing. No classification system is value neutral, objective, or self-evident. All classification systems bear within them the ideological imprint of their production. A system of identifying works of art by their creators might be inappropriate in a community where practices are tied to tradition and repetition, rather than originality and invention."*[63]

With the example of identifying artworks by their creator, Drucker demonstrates that the structures established in art history are so deeply ingrained in how we talk and write about objects that, in practice and besides critical academic discourses, they do not get questioned or challenged. This detail alone highlights the extent to which the system rooted in Western academia, imposes its thinking and understanding onto everything else. Often without even realizing it. This will be discussed in more detail in the chapter on Bias and Power Structures.

Münster and colleagues highlight that there are multiple interpretive moments in the lifecycle of an object record, every one of them adding and shifting the narrative the object is put in. Those moments are for example when one system is replaced by another. This could be from an analog filing cabinet to a first digital cataloguing system but also when one relational database is migrated from one system to another. At these moments already existing data is interpreted, valued and, when the available data fits the new data structure, migrated. If the data does not fit, it either gets unified or restructured or otherwise transformed to make it suitable. Where this is not possible or if information is categorized as unimportant from the start, the data gets discarded and is therefore lost.[64]

Migration projects are complex and take up a lot of labor and time. From my personal experience, I was part of a project team that managed the migration of one collection database to another; it is more work than expected and it always comes

[63] Ibid.

[64] S. *Münster*, F. I. *Apollonio*, P. *Bell*, P. *Kuroczynski*, et al., Digital Cultural Heritage Meets Digital Humanities, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W15 (08/23/2019) 818, doi:10.5194/isprs-archives-XLII-2-W15-813-2019.

with some loss. This can be because there was a specific field in the old system that has no counterpart in the new one and the values must be mapped to a "notes" field. Or because information that was in the former system recorded in a free text field now should be split into designated structured fields to fit the new structure. For example, we had to deal with measurement data that was recorded in one free text field. The values would look something like this: "30x40x50cm, frame measurement". The new system organized measurements in a different structure, providing fields to specify which part was to be described (the frame, the object, the passe-partout etc.), the dimension (height, widths, depth, weight), followed by the value itself and the unit. To record three dimensions as in the example three occurrences of this group of fields would be needed. During migration it was not possible to automate the split of the available data into those fields, especially because the original field had no regulations, and the values were too inhomogeneous to automatically split it (at least not in a justifiable expenditure of time within the project). In the end, the values were mapped to a notes field next to the designated dimension fields. The information did not get worse through this process; again, it was a free text field that could be queried with a full text search, but since new records are catalogued according to the new systems, a discrepancy between the migrated records and the new records emerged. Since the new system showed how well structured (and therefore searchable) dimension data could be, the limitations of the old data became apparent.

As has been shown with the example on the person record of Merit Oppenheimer, standardization is crucial to make large databases work. Databases are there to find stuff. So, if you can't find what you are looking for, for example because the one object you are looking for has "Merit Openheimer" (instead of Oppenheimer) listed as creator, its main purpose is not fulfilled.

Next to people databases, other so called authority databases are common parts of contemporary collection management systems. They allow us to name and describe objects in consistent ways, for example with term taxonomies. "Taxonomies are, quite literally, naming systems. They are comprised of selected and controlled vocabulary for naming items or objects."[65] This selective vocabulary can be about different things, for example places, subject terms, materials and techniques, object names or groups, but also languages or roles of object creators. The taxonomies should be, in the best case, be agreed by the whole institution and consist of hierarchical organized terms. For example, there could be the object name term "painting" with subordinate child terms "oil painting" and "acryl painting". If a record was assigned "oil painting" but the query in the database is looking for "painting" the results would still also include the "oil painting" because it is recognized as the narrower term of "painting". In an attempt to make data interoperable beyond one institution, there are huge endeavors to agree on term vocabularies. One example is the Arts and Architecture Thesaurus of the Getty, who provide terms within hierarchies with unique IDs and descriptions that should clarify what the meaning of the term is. The AAT is in largest parts only available in English, but more and more languages are following suit. By using terms form those external sources (and also stating that), museums can make sure to use terms that were already defined and structured – A task they don't need to repeat by themselves. They also make their data better interoperable.

The interoperability of museum collection databases and the ability of different systems to exchange and utilize data, is one of the pillars of FAIR data (findability, accessibility, interoperability and reuse) and part of most museum mission statements. Several standards have been developed specifically for museum collections in order to achieve this, and the large CMS providers implemented these structures within their software.[66] To name only the most prominent ones: CIDOC-CRM, a formal ontology and conceptual framework, LIDO a XML-based harvesting schema, Dublin Core, a widely used metadata standard that provides a basic set of recourses. Additionally, SPECRTUM is a collection management standard that provides guidelines and workflow suggestions for managing museum collections, not only covering the object cataloguing but also processes like loans, treatments, or deaccessioning. One other important standard is the Object ID of INTERPOL, is it

[65] *Drucker*, The Digital Humanities Coursebook, 58.

[66] See for example Axiell Collection by Adlib: https://help.collections.axiell.com/en/Topics/Standards.htm (accessed on 24.6.2023 19.38) and TMS by gallery systems: https://www.gallerysystems.com/solutions/collections-management (accessed on 24.6.2023 19:38)

aims to provide a guideline on the minimum information that is necessary to identify cultural objects, in case they are stolen or missing. The standard was created together with the Getty Information Institute and is also advertised by ICOM. Object ID is defined by nine information categories: Type of object, materials and techniques, measurement, inscriptions and markings, distinguishing features, title, subject, date or period and the maker. This information should be accompanied by photographs. INTERPOL has created an database for stolen art (Stolen Works of Art Database), which is structured along those fields of information. In the case study below, in Chapter 3.3.2, a comparison of the fields published by MoMA and the Object ID will be done in order to see if the available data would fit the criteria.

There is a significant tradeoff between the extend of how structured, findable, and standardized data is and how much space there is for unique and object specific information. As Gibson and Kahn write, "[t]here is no wiggle room"[67], either an object is an oil painting or an acryl painting, to pick up our example from above, but if the object was created by the use of oil and acryl paint, there is no suitable term available. Adding a new term to the thesaurus would be a possibility, but that only corresponds to the meaning of the thesaurus system if the term will be used regularly. Otherwise, we could end up with as many terms as there are objects and the structuring of those terms in a hierarchy would become an impossible task, leading to the logic of a thesaurus becoming lost.[68]

One interesting thing about CMS and the work of museum professionals with them is that, although the systems are text based, the people working with them rely greatly on the link to images. Picking up again on Andre Malraux's *Museè imaginaire* as introduced briefly in Chapter *2.3*, it becomes obvious that a lot of knowledge and memory lies in images and the photographic reproduction of objects accompanying textual object records.[69]

[67] *Kahn*, *Gibson*, Digital Museums in the 21st Century: Global Microphones or Universal Mufflers?, 42.

[68] See Griffith and his experiences on how standardization is used in describing prints in the collection of the British Museum: *Griffiths*, Collections Online: The Experience of the British Museum, 360.

[69] André *Malraux*, Das imaginäre Museum, Reihe Campus, Bd. 1017 (Frankfurt New York 1987).

Antony Griffiths, from the British Museums, names the ability to add images to the object records as the most important development in the journey of the museums transition to digital catalogues.[70] From my own experience as an employee of a large museum in Vienna, I can agree with him on the significance of images. Or let us say, the significance of missing images or other visual representations. What I have experienced is that if there are no images linked to the object record, the chance that this object is further investigated or worked with, for whatever project, is very small. Let us consider here a situation whereby an object does not have an image attached to it, it tells us something about the status the object has within the collection. It probably was never exhibited, never on loan and never part of an educational program since it came to the museum. Those processes are typically the points within an object life in a museum when their records are enriched, additional information is added, photos are taken, and restaurateurial care is given. So, the records of objects without images are in most cases very marginal, only holding the data that was entered when the object was accessioned. And, if since this moment and today the transition between multiple cataloguing systems took place, this information might have even been boiled down to the absolute minimum through the migration process. The missing image and the marginal record together create vicious circle. The marginal record will not appear in the search statements the stuff is entering, and even if the record was retrieved, the people won't know what the object is with the limited information available and no image visually representing it. Consequently, object records with no images dissolve into the graveyard of the database, the limited data available on the object will even loose significance over time because all other object records are enriched constantly, in comparison to the untouched object record that will become more and more neglected over time. This also shows that object records are never finished, they will typically get enriched over time, new research is added, existing data revised and at times corrected. "In this way, it is fair to say that a museum record is never truly complete, since the information contained within it is liable to be constantly changed and updated."[71]

[70] *Griffiths*, Collections Online: The Experience of the British Museum, 358.

[71] Rebecca *Kahn*, Rainer *Simon*, Feast and Famine: The Problem of Sources for Linked Data Creation, In: Graph Technologies in the Humanities - Proceedings, 2020 91.

As seen above, we touched upon one other very important aspect of record-keeping in museums: it is done on a human scale, by individuals.[72] Large museums, like MoMA, which exist for a long time develop distinctive cataloguing practices and rules which lead to the creation of data that is unique to these institutions and sometimes even unique to specific catalogers or departments within the institution.[73]. The process of acquiring objects requires significant investment by the museums, and, alongside all other tasks in daily business, cataloguing (and/or cataloguing resources) gets too little attention sometimes, as Thomas Padilla further points out.[74] It is important to stress, especially before we delve into the biases that are inherent in collection management systems, that museum staff are doing their best in their abilities and knowledge and that databases are grown things that build on work that might be decades old. Where there's muck, there's brass as Anthony Griffith exemplifies through the cataloguing practices of the British Museums and as we will see in the case study below.[75]

The next chapter will examine how those categorizations and classifications imprint structural imperatives of the Western world on every object. That is, to describe and how misrepresentative, insensible and in cases harmful this can be.

2.4 Bias and Power Structures

The process of systematizing, structuring, and grouping cultural assets often involves constraints. It involves decisions about what to include or exclude, what is considered important or significant, and relies on Western ideologies, which can overshadow alternative perspectives. This chapter will provide a brief overview of

[72] See Coleman who talks about the similar difficulties for libraries and their stuff that can also be applied for museum workers: Catherine Nicole *Coleman*, Managing Bias When Library Collections Become Data, International Journal of Librarianship 5, no. 1 (07/23/2020) 10, doi:10.23974/ijol.2020.vol5.1.162.

[73] *Kahn*, *Simon*, Feast and Famine: The Problem of Sources for Linked Data Creation, 90. For Blagoev see: *Blagoev*, *Felten*, *Kahn*, The Career of a Catalogue.

[74] Thomas *Padilla*, Responsible Operations: Data Science, Machine Learning, and AI in Libraries 11, doi:10.25333/W8SG-8440, (05/05/2023).

[75] *Griffiths*, Collections Online: The Experience of the British Museum, 366.

how the structures of databases are interconnected to the systematics of art history. When exploring the prevailing Western perspectives in art history, we will observe a tendency to undervalue or downplay the importance of non-Western cultures and artistic traditions. Moreover, we will uncover the need to question and critique what is considered to be the "natural" order of things and way of talking about objects. Some examples will illustrate how difficult it will be to acknowledge naturalization, especially when examining the systematic from within the same ideology and framework.

As a positive ending of the chapter, some of the existing efforts of art historians, digital humanists, museum staff, archivists, librarians, and many others will be introduced. They will show various initiatives that are aimed to open up the institution walls, to democratize the decision-making process and to redistribute the power that lays withing institution's hands in a more widespread and ethical way.

2.4.1 Art history and its love for categories

Art history has predominantly been shaped by the perspectives and principles established by male scholars and researchers originating from Western countries within the last two centuries. The discipline's methodologies and understandings are based on Western ideologies, the patriarchal system, and colonial ideas. And even if not intended, they are ingrained within their very foundations, so deep down that they are sometimes hard to recognize. One aspect that is inherent of the discipline is to place objects in a line of development, to put them in relations to other objects, and to put them together in groups. The founding figures of art history of the last century, Heinrich Wölfflin (1864-1945), the already mentioned Abi Waburg (1866-1929), Erwin Panofsky (1892-1968) or Ernst Gombrich (1909-2001), to mention just a view, all took significant effort in systematizing the discipline. They also established the general canon of what is part of art history and how the objects were to understand. Even if Panofsky and others embedded the artworks into broader historical, social, and intellectual context in the aim to understand their meaning, they created clear distinctions of what was considered to be art and what was not. They also established the manner of talking about artwork and what the key information is that allows an artwork to be identified. The creator's name, the title and the creation date are the most important facts about an artwork and are always part of common art catalogues, publications, and labels on museum and gallery walls. This might be enriched with information on material and technique, dimensions and possibly the current owner of the object.

The placement of artwork within the storyline of development of styles, schools, and groups was considered the main art historian task. This can be recognized in the chapter titles of famous books like "The story of Art "[76] by Gombrich, or in the works of Panofsky that focus on putting objects or oeuvres in relation to other objects, time periods or artists oeuvres.[77] And while this is an explanation of why the catalogues look like they do, this also already shows that a lot is missed, when only little information on artworks is collected that is regarded as key information. The fact that objects are connected to each other and also to historical or social developments is not represented in catalogue systems. It also puts enormous significance on the individual who created the object, something that might not be of relevance for all cultures.

Today it is made possible in digital databases – through hierarchical or flat links, or through shared characteristic – to create references and connections between records, but the intrinsic meaning of context (social, historical, economic, cultural, etc.), that is an important aspect of artwork is still only (and if at all) recordable in unstructured free text fields in databases.[78]

It is valid to assert that cataloguing systems inherently fall short in capturing a complete picture of objects and their meanings. They have always been limited, but these limitations were not recognized or examined critically. As data, in general, is widely assumed to hold truth or be consistent of facts, catalogues of museums are also assigned credibility and truthfulness. With the increase in power of the museums as an institution, museum catalogues have also acquired an authoritative position within society that can hardly be questioned.

[76] Ernst *Gombrich*, The Story of Art (London 1979).

[77] Erwin *Panofsky*, Early Netherlandish Painting, 9. [print], The Charles Eliot Norton Lectures 1947–48 (New York 1987); Erwin *Panofsky*, Irving *Lavin*, William S. *Heckscher*, Three Essays on Style (Cambridge, Mass 1995).

[78] *Kahn*, *Simon*, Feast and Famine: The Problem of Sources for Linked Data Creation, 90.

2.4.2 Bias in collection data

Johanna Drucker asks questions like, “Is there such a thing as a “feminist” table? Or a “racist” form for data entry? Can format embody bias, or only content?” [79] And she comes to the conclusion that “[t]he structure of data is always fraught with values and the very act of defining categories can create exclusionary results.”[80] For Drucker already the design of databases is based on conceptual decisions and every implementation step is based on choices that might incorporate biases. [81] Coming back to the introductory chapter on data, it should be repeated before we continue, that data is always produced, never something that exists on its own. It is necessary to recognize data always as an “expression of a point of view and value system.”[82]

Before we advance to some examples on biases in collection data, it is an important side note, that also language is something that is constructed. Rhiannon Mason brings forward the motion within cultural theory and museum studies to start from Ferdinand de Saussure’s theory of semiotics and the significant realization that language is not objective but socially constructed, learned, and negotiated.[83] This leads us to the first example on how museum catalogues misrepresent communities. Computational text is expressed predominantly in English worldwide;[84] however, other languages of the Western world were predominantly used to describe cultural objects, too, even then when those languages are missing terms to properly describe objects. An example of objects from Māori culture can show this exemplary. Heidy Geismar introduced in her book “Museum Object Lessons for the Digital *Age*” multiple examples of objects that cannot properly described in English, if not terms from Māori are incorporated in the dictionary. One of them is the noun “taonga” which means a cultural treasure and which significance does not lie in the object itself but also in the, and more importantly, relationship and connection to other

[79] *Drucker*, The Digital Humanities Coursebook, 76.

[80] Ibid.

[81] Ibid., 77.

[82] Ibid., 26.

[83] Rhiannon *Mason*, Cultural Theory and Museum Studies, In: A Companion to Museum Studies, Sharon *Macdonald* Ed., Blackwell Companions in Cultural Studies 12 (Malden, MA 2006) 18.

[84] *Padilla*, Call to Action, 16.

objects.[85] This example also shows that the real meaning of the object cannot be represented in the existing database structures, as they all focus on the object, their creator, its technique, and materiality, which are assumed to carry the meaning. And there is no possibility for adding other levels of meaning, such as that of connectivity. Picking up from the example made above that a system that identifies object by their creators might also be inappropriate; for example, when the objects are results of traditions or habits and not objects intended to hold originality.[86] This and also the example on taonga can show types of naturalizations that are imbedded in cataloguing strategies and how the Western museums think about objects. For Haidy Geismar, de-naturalization as one of the key tasks together with the general questioning of the assumption that the digital and data is something natural in order to diversify collects. [87] Geismar states that "[b]y thinking of digitization as a cultural process of interpretation and meaning making, we can open up what has often been radically naturalized in both museum and digital environments."[88]

Things that are considered to be natural is also what Gabrielle Foremand and Labanya Mookerjee talk about when they argue that the available protocols of collecting and cataloguing cultural objects do not meet the need of all users and that they are specifically frightening to Black communities: "Not only do the protocols and developments of digital collections, of interacting with objects, not meet the needs of various users – let's call them people or communities – who interact with "objects in digital spaces," the lexicon itself reproduces particular freighted ideas for Black communities of researchers and students, many of whose ancestors entered the West as chattel property, as people who were both called objects and "leveraged," that is bartered, mortgaged, sold and *listed* as such. In the US, this is true for the almost 250 years of municipal, census, and other records which make up collections and archives during slavery, for records that document the dept peonage that characterizes Jim Crow, and, one might argue, for ways in which Black people are accounted for in a prison industrial complex that again treats members of

[85] *Geismar*, Museum Object Lessons for the Digital Age, 24.

[86] *Drucker*, The Digital Humanities Coursebook, 57.

[87] *Geismar*, Museum Object Lessons for the Digital Age, 23.

[88] Ibid., 27.

communities as things to be categorized, as surveilled and recorded objects."[89] They conclude that the methods of data collection and structurization needs to be critically evaluated and adapted in order to not repeat institutional discrimination that lies within them.

> *Datasets, when constructed using conventional methods of data collection and organization, run a similar risk of activating institutional power and defining 'credibility', especially when the data is procured from traditional archival sources that too often excise, anonymize and erase certain subjects, transmogrifying them in turn into (almost invisible, ghosting) 'objects' and 'items'."*[90]

One of the most obvious but maybe also most urgent bias in collection data and collections in general lies in which artists are represented and which objects are selected.

The collections of major art museums have historically focused on collecting artworks created predominantly by male artists. This bias is evident in the case study of the MoMA dataset (see below), as it shows that only a small percentage of museum objects were created by female artists. Furthermore, non-binary artists are severely underrepresented or nearly absent from the collection. These inequalities reflect the patriarchal structure in the Western world and the ongoing struggles of women* artists in their quest for recognition in the art world. It also highlights the enforced binarity of gender classification in data collection, most cataloguing systems only hold two distinct gender values, male and female. Although most states today accept multiple gender classifications, this is not yet reflected in the datasets of art collections. During the aforementioned migration project I was part of, from one CMS to a product provided by one of the global market leaders, it was an additional configuration task to add the in Austria recognized gender categories to the

[89] Gabrielle *Foreman*, Labanya *Mookerjee*, Computing in the Dark:
Spreadsheets, Data Collection and DH's Racist Inheritance, Always Already Compuational: Collections as Data, 2019 108.

[90] Ibid., 109.

enumerative term list that is imbedded in the gender field. The default version only held three values: male, female, unknown.[91]

Discrimination and exclusion were not limited to gender bias but encompassed every form of discrimination dictated by the socially accepted canon. This pervasive system marginalized artists based on factors such as race, ethnicity, sexual orientation, socioeconomic status, and more. And, as we have seen how powerful players museums are in establishing narratives and memory all those perspectives and viewpoints were excluded and silenced. Thomas Padilla summarizes the already mentioned aspects:

> *"Historic and contemporary biases in collection development activity manifest as corpora that overrepresent dominant communities and underrepresent marginalized communities. Where marginalized communities are represented, that representation tends to be within the context of narratives that dominant cultures sanction."*[92]

2.4.3 Challenges and chances

In the latter part of the twentieth century, museums progressively challenged the validity of established categorization systems for organizing collections and reassessed their educational function, adopting more inclusive and participatory approaches.[93] Although Heidy Geismar argues that is wasn't a growing awareness on biased or misrepresentative data, but the emergence of art objects that did no longer belong to the canonized mediums of painting, drawing and sculpture, this therefore prompted a shift in the conversation on cataloguing.[94] [95] Today the field agrees on the necessity of action, as Thomas Padilla writes: "[a]ll agreed that the challenge of doing this work responsibly requires fostering organizational capacities

[91] The following gender categories are legally available in Austria: inter, divers, open, female, and male. See also: https://www.wien.gv.at/menschen/queer/intersexualitaet/anerkennung-oesterreich.html (accessed on 26.6.2023 11:19)

[92] *Padilla*, Call to Action, 15.

[93] *Macdonald*, Collecting Practices, 88.

[94] *Geismar*, Museum Object Lessons for the Digital Age, 55.

[95] See also Bruce *Altshuler* ed., Collecting the New: Museums and Contemporary Art (Princeton, New Jersey 2007).

for critical engagement, managing bias, and mitigating potential harm."[96] The goal is not to eliminate biases but to recognize them and manage them responsibly. "Managing bias rather than working to eliminate bias is a distinction born of the sense that elimination is not possible because elimination would be a kind of bias itself—essentially a well-meaning, if ultimately futile, ouroboros."[97] Also Catherine Nicole Coleman points out that the attempt to de-bias data is not the goal but that we need to accept bias as inherent of data: "Bias is an unavoidable consequence of situated decision-making that we have to reckon with." [98]

In summary, biases are deeply rooted in both data and databases and present a challenge when it comes to standardization and classification as they are a basic feature therein. A key challenge is to find a balance between fitting existing data structures and models while making room for expanded descriptions and meanings. One of the points at which movement toward greater inclusion and fairness can be seen, are in the attempts to decrease gender bias in art collections in art museums around the world. This is also evident in the case study of MoMA (see below), where efforts to achieve gender balance are also evident in the dataset.

2.4.4 Collection as data

To exemplify a practical but also ethical and mindful approach on how computational driven methods of research and teaching can be done in the field of cultural heritage collections, the so called "Collection as Data" framework can be drawn upon.[99] While Thomas Padilla and his colleagues' framework is primarily rooted in library studies, it can be effectively applied to various other fields.

> *Collections as data development must critically engage with bias in collection and description, archival silences, and assumptions about collection use. The viability of collections as data effort demands*

[96] Thomas *Padilla*, Laurie *Allen*, Hannah *Frost*, Sarah *Potvin*, et al., Always Already Computational: Collections as Data, 2019 6.

[97] *Padilla*, Call to Action, 9.

[98] *Coleman*, Managing Bias When Library Collections Become Data, 10.

[99] Thomas *Padilla*, On a Collections as Data Imperative, 2017.

> *critical engagement - especially as collection practices leveraging computational means like machine learning, computer vision, and more hold as much potential to harm as to help. Archival approaches to provenance, with their focus on documenting the custodial and contextual history of objects, provide one path forward. Ethical fault lines are often easier to see when trying to develop new policies and workflows. Examination of policies and workflows should support changes in practice. Prior harms should be acknowledged and remediated to the extent possible.*[100]

The framework is based on ten principles that can be summarized as such: They aim to promote the computational use of digitized and born digital collections while addressing historical and current inequities in the scope, description, access, and use of collections. They strive to lower barriers to use and emphasize the need for customized collections that meet the specific needs of users. Shared documentation is encouraged to facilitate work, and collections should be freely accessible by default unless ethical or legal obligations prevent it. Interoperability is emphasized, and transparent procedures are used to develop trustworthy and long-lasting collections. Both primary data and associated metadata are considered, and the development of collections as data is an ongoing process with no definitive end point.[101]

This example should illustrate, before we turn to the MoMA case study in the next part, that work can be done even when the task seem too big, and the limitations of the available data have been recognized as considerable. It is work that will never be finished, but even small changes are the step in the right direction. Recognizing and owning up to absences, misalignments, misrepresentations, or just plain mistakes is an important first step.

[100] *Padilla, Allen, Frost, Potvin, et al.*, Always Already Computational: Collections as Data, 14.

[101] The ten principles are listed in: *Padilla*, On a Collections as Data Imperative, 20.

3 Part II: Case Study of the Museum of Modern Art datasets

The following case study of the MoMA datasets shows how the exploration, analysis and interpretation of a large art collection dataset can provide meaningful insights in collection and cataloguing history and practice. The first part will provide a short introduction on the Museum of Modern Art to provide context to the datasets. Furthermore, the cataloguing practices and the current online presentation of the museum will be examined briefly, followed by an examination on how the datasets that will be worked with are published by the museum, and what information the added to accompany it. After a statistical analysis of the datasets, an exploratory data analysis (EDA) is performed to dig deeper into the dataset. During the EDA, a systematic exploration of the dataset will be undertaken, and in the subsequent step, selected research questions regarding the MoMA collection are analyzed in depth to obtain meaningful results and provide comprehensive answers. The resulting findings are summarized and concisely presented to provide a clear and comprehensive understanding of the data set. Visualizations will be used to convey the results to enhance their understanding and accessibility.

3.1 About the Museum of Modern Art

Prior to conducting a collection analysis based on the datasets, it is pertinent to introduce the museum whose datasets will be worked with in detail. A look at the museum's official presentation provides valuable insights into how it sees itself, the mission it has chosen for itself, and its transparency in terms of cataloging practices and classifying objects. This chapter therefore begins with a brief overview of the museum's beginnings, with a look at influential personalities who have played an important role in the development of the museum and contributed to its current international recognition. This is followed by a look at the information that the museum publishes about its own policies and workflows regarding data entry and cataloging of objects in their databases. Very briefly, the current online presentation of the collections will also be discussed, primarily in order to compare it below with the information available in the datasets used here.

The Museum of Modern Art was established by three women in 1929 in New York. Abby Aldrich Rockefeller (1874-1948), Lillie P. Bliss (1864-1931) and Mary Quinn Sullivan (1988-1939) were, as Sybil Kantor puts it, “three ambitious, socially aware women who were conscious of the gap in American museology resulting from the absence of readily available European modern art.”[102] From the three Abby Rockefeller was the most active one, her motivation and also money, together with the in 1931 inherited collection of Lillie P. Bliss, put the young museum on its path of success,[103] The museum was not created out of a vacuum, rather it was the results of many developments, ideas and needs that accumulated in its foundation. One large influence was the “Harvard Society of Contemporary Art” at the University of Harvard, founded one year before MoMA in 1928 and which had a similar program and followed similar ideas as MoMA then did.[104] The Bauhaus in Germany was another important source of inspiration, its organizational concept got imbedded in the “multidepartmental” structure of the museum and the ideas and the art of its associates got incorporated in the collections.[105] The first museum director was Alfred Barr (1902-1981), he was, although very young and unexperienced, he only just finished his studies, chosen by the three founding partners after being suggested by Paul J. Sachs (1878-1965).[106]

[102] *Kantor*, Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art, 191.

[103] Ibid., 193.

[104] Ibid., 197.

[105] Ibid., 254.

[106] Ibid., 210.

Figure 1 Jay Leyda, Alfred H. Barr, Jr., New York, 1931–1933. Courtesy: The Museum of Modern Art, New York.[107]

One difficulty the museums faced from the very first exhibition on was the critique they would not focus on American art and artists enough but rather only display foreign, in most cases, European art. Following the first exhibition that displayed European postimpressionists, the second one was solemnly showing American artists in order to work against these claims.[108] "Paintings by 19 Living Americans"[109], which was more or less a second edition of an exhibition that took place at "The Harvard society of Contemporary Art" a year earlier, was the first of many attempts to counter this criticism.[110] Although Alfred Barr had dreamed of a permanent collection for the museum from the beginning (see his iconic whale drawing of the perfect

[107] Ibid., 3.

[108] Ibid., 219. See the online presentation of the exhibition on MoMA's website, with scans of publications, installation views and additional archival material: https://www.moma.org/calendar/exhibitions/1767 (accessed on 28.06.2023 14:20)

[109] See the online presentation of the exhibition on MoMA's website: https://www.moma.org/calendar/exhibitions/1912 (accessed on 28.06.2023 14:20)

[110] *Kantor*, Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art, 202.

collection)[111], in the early decades the museum had actually intended to transfer works that no longer are regarded as contemporary to other museums. Contracts were drawn up with the Metropolitan Museum for this purpose, but they were never implemented, because in the 1950s, the museum had to realize that the exhibition of prominent and canonized works of art was necessary to keep visitors coming in. This led, on the one hand, to the conception of a permanent exhibition to make the already famous works permanently accessible and, on the other hand, to the abandonment of the plan to pass on works in return for money to acquire new ones.[112] After Barr's directorship ended in 1943 and a period of times with only interim directors, five other men directed the Museum of Modern Art until today. Rene d'Harnoncourt (1901-1968) from 1949 until 1967, Bates Lowry (1923-2004) for the short time from 1968 until 1969, John Brantley Hightower (1933-2013) for only two years 1970 until 1972, followed by Richard Oldenburg (1933-2018), brother of the Pop Art sculptor artists Claes Oldenburg, who directed until 1995, when Glenn D. Lowry (born 1954) took over, who holds the position until today.[113] In the glorified story about the rise of the MoMA Alfred Barr plays the most significant role of all the directors. His primary goal was to expand the public's perception of art to new mediums like photography, architecture, film, and industrial design. Under his directorship influential and today famous exhibitions like "Machine-Art" in 1934 or "Photography: 1839-1937" in 1937 were put together and resulted in the incorporation of those fields in the classical art canon.

He also was responsible for the installation of specialized departments, each of which were responsible for one category of art, and he created exhibitions, publications, and managed acquisitions. Not all of the today existing departments were established from the beginning, but they evolved over time, along with the growth and valuation of the art type they represented. The department of architecture was the first one Alfred Barr was to create in 1932, only three years

[111] Kirk *Varnedoe*, Introduction, In: Modern Contemporary: Art since 1980 at MoMA, Kirk *Varnedoe*, Paola *Antonelli*, Joshua *Siegel*, *Museum of Modern Art (New York, N.Y.)* Ed., 2nd ed (New York 2004) 11.

[112] *Zalman*, Unpacking the MoMA Myth, 287–288.

[113] Glenn D. *Lowry*, Introduction, In: MOMA Highlights: 350 Works from the Museum of Modern Art, New York, *Museum of Modern Art (New York, N.Y.)*, Harriet Schoenholz *Bee*, Cassandra *Heliczer* Ed., 2nd ed (New York 2004) 16–21.

later, in 1935, the film library, which was later upgraded to its own department, was established. Departments of drawings, prints, and photography followed.[114] There from the beginning was the department for painting and sculpture. This very broad and overlapping divisions in six departments (although, as we will see the dataset shows eight) reflect a basic tendency in Barr's approach to art and created clear responsibilities within the museums management. His aim was to assign all art objects to clearly definable categories, and to create superordinate narratives of developments, connections, achievements, and assigning key roles, almost creating geniuses, to specific men withing those narratives. This endeavor can be found today, in his idealized diagram of cubism and abstract art which provides the impression that the development of art can be looked at from an almost scientific and analytical viewpoint that declared clear dependencies, see Figure 2.[115] The departments exist today and still form the buildings blocks of the museums organization and the narration they build within their productions. They get introduced on the museum's website and are part of the story the museums tells about itself.[116]

[114] *Kantor*, Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art, 212.

[115] Sybil Kantor, *Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art*, 2002, p. 317; Sandra Zalman, 'Unpacking the MoMA Myth: Modernism under Revision', *Modernism/Modernity*, 29.2 (2022), 283–306 (p. 284) <https://doi.org/10.1353/mod.2022.0009>.

[116] *Varnedoe*, Introduction, 13; *Lowry*, Introduction, 16. See the description of the six departments on the museum's website here: https://www.moma.org/about/curatorial-departments/ (accessed on 30.6.2023 09:53)

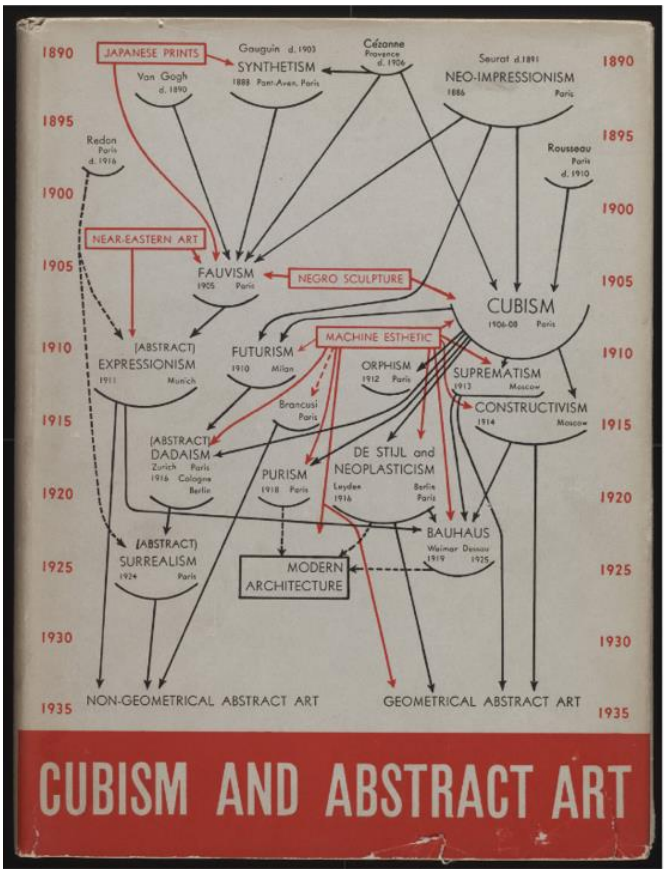

Figure 2 Dust jacket with chart prepared by Alfred H. Barr, Jr. of the catalogue, "Cubism and Abstract Art" by Alfred H. Barr, Jr. 1936, offset, printed in color, 10 1/8 x 7 3/4 in. (25.7 x 19.7 cm.).The Museum Modern Art Library, New York. The Museum of Modern Art © The Museum of Modern Art/Licensed by SCALA/Art Resource, NY.[117]

Figure 2 shows some interesting aspects that reflect the social and political self-understanding and the Western mindset of Barr's time. While the chart goes on individual level, when naming Van Gogh or Cézanne as important influences for fauvism and cubism, it is tremendously oversimplified and generalized when adding "near-eastern art" or "negro sculpture" as sources of inspiration. It reflects the racist implementation of art history, that acknowledges the craftsmanship and quality of those sculptures but on the other hand did not care enough to further investigate these objects, the circumstances of their creation, or their meanings. Rather generalizing everything based on the attributed race or skin color of the assumed creators. Sybil Kantor talks about the large impact the chart had and uses it as one example on Barr's outstanding role in 20th century art understanding. In her text, she replaces, without any comment "negro sculpture" with "African sculpture", showing that even in 2002, although recognized that the original term cannot be repeated and

[117] *Zalman*, Unpacking the MoMA Myth, 285.

is changed to a less offensive and harmful one, the generalization and the categorization stays.[118]

As Sybil Kantor writes, “Barr’s interest in precise terminology was the basis from which he categorized the various art movements chronologically, forming classifications similar to the taxonomies of zoologists and botanists…”[119] he created a system of evolution and innovation that is prioritizing European and American art.[120] Canonizing what is considered to form the highlights, “the classics”, of modern art until today.[121] In 1931, when Lillie Bliss inherited her collection to the museum, it came in possession of artworks from “some of the most important artists of the modernist movement”, beneath them works from Redon, Degas, Toulouse-Lautrec, Picasso, and Matisse.[122] This was, as the MoMA story is told, the start of a successful history of donations and acquisitions that form today’s collection with hundreds of thousands of objects. On their website, MoMA lists in its mission statement, “200,000 paintings, sculptures, drawings, prints, photographs, media and performance art works, architectural models and drawings, design objects, and films” additionally to two million film stills, the museums library and archive as in their possession.[123]

Although the museum enjoys international recognition, inequalities, and disproportions and absences within the museum’s collection, program, and politics were and are criticized and protested. Since the 1980s protests against the gender-based exclusion of non-male artists from the collection and the exhibitions take place, started by the “Women’s Artists Visibility Event” (W.A.V.E) in 1984. This protest was lighted by the enormous underrepresentation of female artists in the

[118] See *Kantor*, Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art, 327.

[119] Ibid., 321.

[120] Page xx in the Introduction in: *Kantor*, Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art.

[121] As Kirk Varnedoe, a curator at MoMA, wrote about the museum. See: *Varnedoe*, Introduction, 12.

[122] *Kantor*, Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art, 240.

[123] The mission statement is available online on the museum’s website: https://www.moma.org/about/mission-statement/ (accessed on 28.06.2023 17:01)

exhibition “An International Survey of Recent Painting and Sculpture” (91.5% male artists).[124]

Since the 1930s the absence of black and Native American artists was a topic of discussion in MoMA. Artwork from nonwhite artists were exhibited in MoMA under two possible narrations, either being a representative of so called “primitive” or “naïve” art, or, as an inspirational source for modern art, as can also be seen in Barr’s chart (see: Figure 2) or exemplary on the titles of the exhibitions “American Sources of Modern Art (Aztec, Mayan, Incan)”[125], “African Negro Art”[126] or “Understanding African Negro Sculpture”. The first solo exhibition of a nonwhite artists was the “Sculpture of William Edmondson” exhibition in 1939, showing only 12 objects of Edmondson, framing him and his work in stereotypical racist ways, focusing primarily on his skin color and allegedly naivety and lacking professionality. And although the exhibition was a success, the museum did not acquire a single object at the time, although to acquire objects after exhibiting them was the common practice.[127]

To date, MoMA is in critique for following a narrative that prefers artist and artworks that belong to the Western standard canon, a canon the museum itself created and manifested over the last almost 100 years of existing. In 2019, after renovation and expansion, MoMA reopened its doors to new permanent exhibitions, claiming to have “expanded the way of thinking” and the promise to change the program frequently to allow different stories and histories of modern art to be seen and told.[128]

[124] Sabra *Moore*, Openings: A Memoir from the Women’s Art Movement, New York City 1970-1992, First edition (New York, NY 2016). For the presentation in the online exhibition history of MoMA see: https://www.moma.org/calendar/exhibitions/2220 (accessed on 29.06.2023 09:25).

[125] https://www.moma.org/calendar/exhibitions/2932 (accessed on 29.6.2023 12:24)

[126] https://www.moma.org/calendar/exhibitions/2937 (accessed on 29.6.2023 12:26)

[127] Charlotte *Barat*, Darby *English*, The Artist Wasn’t Present: On MoMA’s Fumbled First Showing of Black American Art, ARTnews, 07/17/2019, online at <https://www.artnews.com/art-news/news/among-others-blackness-at-moma-excerpt-12972/>; *Museum of Modern Art (New York, N.Y.)*, Charlotte *Barat*, Darby *English*, Mabel *Wilson*, et al. eds., Among Others: Blackness at MoMA (New York 2019).

[128] William S. *Smith*, Dissident Modernism Meets Peak Philanthropy at the New MoMA, Art in America, 10/25/2019, online at <https://www.artnews.com/art-in-america/features/moma-reopens-

In 2021 protests against some of MoMA board members and its “toxic philanthropy” in its so called “Strike MoMA” campaign endured for ten weeks and resulted in the resignment of Leon Black as chairman.[129] And during this year’s fund-raising event at MoMA climate activists protested against one other board members, whose husband invests in oil and gas projects.[130] The decision of who is on the board is crucial for the development of the museums. This is because of internal rules the institution follows. The board decides who forms the committees, which in turn decide which objects are acquired for the collections.[131] It can be argued that the worldview and ethics of the board members have a direct impact on the development of the collections and the acquisitions.

Those different protests can again underline how significant museums are for the narration of the history and contemporary developments and how serious the communities take them in that role. The analysis of the datasets won’t allow to examine issues like the personnel on the board or other executive organs of the institution, but it will allow to see general trends, and if, for example the gender distribution of artists represented in the collections, changed over time.

modern-art-politics-protests-63665/>. See also the museums statement regarding the reopening: https://www.moma.org/interactives/moma_through_time/2010/moma-reopens/ (accessed on 29.06.2023 13:05)

[129] Zachary *Small*, MoMA Survived Ten Weeks of Protest. But Inside the Museum, Some Employees Are Feeling the Strain. A Protest Movement Questioning the MoMA Board’s Ties to “Toxic Philanthropy” Came in the Midst of a Staffing Crisis., Artnet News, 07/19/2021, online at <https://news.artnet.com/art-world/moma-survived-ten-weeks-protest-strike-moma-1990049>.

[130] Elaine *Velie*, Protesters Crash MoMA Gala Over Board Chair’s Fossil Fuel Ties. Climate Activists Are Asking the Museum to Remove Board Chair Marie-Josée Kravis, Whose Husband’s Private Equity Firm Has Invested Billions in Oil and Gas Projects., Hyperallergic, 06/06/2023, online at <https://hyperallergic.com/826458/protesters-crash-moma-gala-over-board-chairs-fossil-fuel-ties/>.

[131] See the Collection Management Police of the museum for a details description of the process and also Kirk Varnadoe how describes the process in a brief way: *Museum of Modern Art*, Collections Management Policy (04/20/2020) 5, online at <https://www.moma.org/momaorg/shared/pdfs/docs/about/Collections-Management-Policy-2020-04-20.pdf>; *Varnedoe*, Introduction, 14.

Before we continue to the next section devoted to the cataloguing practices of MoMA one project should be addressed: The museum devotes itself to a major ethical challenge: provenance research of works of art that were expropriated during the NS Era in Europe. The project started in 2001 and examines the provenance of around 800 objects that could have been looted.[132] They also provide an overview of the objects in question in a similar as the regular online collection is fashioned, with the ability to download basic information of the artworks in a spreadsheet. Although they claim that research is still ongoing and that there are regular updates, the file was based on its title edited for the last time in January 2020.[133] However, the museum has not restituted any artworks, but there are ongoing lawsuits and claims that might lead to that.[134] The analysis of the dataset below might show if there other objects of questionable provenance the museum should look into. This could be objects of communities that were colonized or suppressed and who's objects were taken without permission.

[132] See he projects website here: (accessed on 29.06.2023 10:22)

[133] The list of artworks can be downloaded as .xslt sheet and its named "prp_objects_1march2016_updated_january2020". The features that are included are the Artist, Title, Date, medium, Dimensions, URL, Object Number and Department. There is no additional information on the provenance and even the credit line, which is provided in the dataset the museum provides via GitHub (see below), is not included.

[134] See this master thesis for further details on the topic and also the following newspaper articles on the subject:: Tiffany-Quan *Le*, MoMA and Nazi-Era Art Restitution: Contexts and Thoughts for the Future (Master Thesis Concordia University 2017), online at <https://spectrum.library.concordia.ca/id/eprint/982899/1/Le_MA_F2017.pdf>; With New Lawsuits Against MoMA and the Santa Barbara Museum of Art, the Heirs of a Holocaust Victim Are Seeking to Reclaim a Pair of Schieles. The Pieces Were Once Owned by Austrian Jewish Performer Fritz Grünbaum., Artnet News, 12/21/2022, online at <https://news.artnet.com/art-world/with-new-lawsuits-against-moma-and-the-santa-barbara-museum-of-art-the-heirs-of-a-holocaust-victim-are-seeking-to-reclaim-a-pair-of-schieles-2234437>; Patricia *Cohen*, Family's Claim Against MoMA Hinges on Dates, The New York Times, 08/23/2011; Isabel *Vincent*, These Famous Artworks Were Looted by Nazis — and Are on Display at Met, MoMA, New York Post, 08/22/2022, online at <https://nypost.com/2022/08/22/new-law-requires-new-york-museums-to-label-nazi-looted-works/>.

3.1.1 Cataloguing practices in MoMA

MoMA unfortunately does not provide specific information about its cataloging practices, including details about the individuals responsible for cataloging or the specific policies or frameworks used. The provided “readme” file (see above) in the associated GitHub repository does not convey any insights into the specific cataloging practices either. The museum also does not provide comprehensive information on this topic on its website, leaving the details of its cataloging practices in the dark. Based on the information provided in the "Collection Management Policy" disclosed on the museum's website, there are indications that the institution aligns its cataloguing practices with the rules and guidelines established by the “American Alliance of Museums” (previous American Association of Museums, short: AAM). Although the museum makes multiple references to the AAM throughout their mission, an explicit statement confirming this alignment is not explicitly articulated.[135]

Some insights can be learned from scholarly articles by researchers who have worked with MoMA's collection. Fernando Domínguez Rubio states that MoMA utilizes TMS (The Museum System) for cataloguing their collections.TMS is a comprehensive software solution employed by museums worldwide to facilitate various aspects of collection management, including cataloguing, documentation, and location management.[136]

The software will be briefly described after summarizing some other insights we can gain from Rubio. Through his work with the MoMA collections, he is able to describe the cataloguing process almost like an insider and for him, the main task for cataloguing new acquisitions is to divide the object and its “constituencies” between the category of art and non-art. [137] Constituencies is the term for everything that

[135] *Museum of Modern Art*, Collections Management Policy. The American Alliance of Museums is a prominent organization that supports museums and museum professionals, providing resources, guidelines, and advocacy for the museum field. Their mission includes promoting excellence, ethical practices, and public engagement in museums across the United States.

[136] *Domínguez Rubio*, Preserving the Unpreservable, 629.

[137] Ibid., 637.

comes with the art object, crates, frames, electronic parts, contracts, certificates, installation instructions, spare parts, exhibition copies and so on.[138]
He describes this process as follows:

> *"The first task when acquiring an artwork is to transform the complex constituencies in which they are inserted into legible, manageable, and unified 'objects of knowledge.' To achieve this, the first operation upon receiving a new artwork is to classify and separate those components containing aesthetic value—i.e., the art 'proper'—from those 'non-art' components containing other forms of value, like 'research' or 'legal' value. Tracing the boundary between these forms of value—a task reserved at MoMA to curators—defines the physical location of each component within the museum and, more importantly, its location within different realms of knowledge and expertise."*[139]

This allows for some assumptions about the workflow that provides valuable insights for our matter at hand. First, the curators are directly involved in the cataloguing practices, they are the ones who decide what is considered art and what is handed over to the archive or other repositories. This is not a trivial task; for example, the packing material which the objects arrive in the museum can be artistically valuable, which could be, to name an example, drawings by the artist on envelops, packing paper, maps or on the frame. To draw the line what is accessed to the art collection and what is left out has long lasting consequences. The climate and security standards for the art collections are mostly higher than those for archives, and the chance for objects to be exhibited or worked with is, as has been shown in the chapter on Collecting and cataloguing cultural objects, significantly higher, if recorded in the object catalogue and photographed. We can additionally assume that the curators also decide in which department the objects are assigned to, this might even already happen earlier in MoMA's case, since the commissions of the departments are the ones suggesting an object for acquisitions, and therefore responsibility within the museum structure. How deep their saying for the other

[138] Ibid., 628.

[139] Ibid.

descriptive classification goes remains unsure. Based on my own experience from working in art museum collection, I suspect that the task of further describing and classifying is handed over from the curators to the experts in the registrar department.

One additional fact Rubio is informing us on is that for objects that consist of different parts specific records are added. Each of them representing one part and set in hierarchical relation to each other.

> *"For example, in the case of an oil painting, the canvas is assigned a unique number (e.g., 300.456), while the other components are assigned a suffix identifying their function within the constituency, for example, FR, for main frames (300.456.FR) and TR for the travel frames (300.456.TR)."*[140]

If the links and dependencies is also recorded in other ways additionally to the semantic within the object number is not known. It will be investigated in the case study below (see page 65) if those constituency records are part of the datasets.

TMS The Museum System

As we have learned MoMA is using TMS for cataloguing their collections. A brief description of the system is therefore of need: "Collections Management with TMS Collections" is a product by "GallerySystems". "GallerySystems" was founded in 1981 and following their website, they have more than 800 clients in over 30 countries worldwide, under them famous and large museums like the Metropolitan Museum of Art, the Tate, The National Gallery, Getty, and LACMA.[141] The company describes "TMS collection", the new web-based version of TMS, as such:

> *"TMS Collections is a sophisticated, easy-to-use relational database application, designed specifically for collections, content, media, exhibition, and loan management. Our web-based CMS is comprised of interrelated modules with supporting functionality for*

[140] Ibid., 629.

[141] https://www.gallerysystems.com/about-us/ (accessed on 17.06.2023 16:57)

> *entering and tracking all collections data and management activities.“*[142]

We don’t know if MoMA is using the web-based version already, but nevertheless there will not be a huge difference to the client-based version. TMS Collections offers ten different modules: “Objects”, to catalogue information on objects; “Bibliography”, for cataloguing publications; “Exhibitions”, to manage exhibitions; “Loans”, to manage incoming and outgoing loans; “Shipping”, to manage all transports of object to and from the museum; “Media” to record media files and their metadata, “Constituents” for cataloguing persons and institution records; “Sites”, to manage geographical locations; “Events”, to record significant events in an objects history, and “Insurance”, to track everything related to insurances.

Add-on software is available, for example “eMuseum”, a tool that published data from TMS Collection to open accessible websites, a Digital Asset Management System, and an audit manager.[143] The software is adherent to international standards, one of them being SPECTRUM, but also CIDOC CRM, LiDO, Dublin Core Metadata Standard or Getty vocabularies (AAT and TGN) are supported.[144] We can assume that MoMA is using most of the available tools to manage its extensive collections and attached workflows.

3.1.2 The online presentation of MoMA’s object and artists records

MoMA publishes parts of their data in various ways. Through well-known printed publications, like catalog raisonnés or exhibition catalogues but also in various digital formats. The datasets that will be used for the case study is one example, one other is the online collection on the museum’s website. Although it was not possible to find any recourse that would state it, it is assumed that the current online collection pulls its data from the collection management system. There are multiple possibilities how the data is retrieved, through a specific API, by the use of the “eMuseum” tool

[142] https://www.gallerysystems.com/solutions/collections-management/ (accessed on 17.06.2023 17:06)

[143] *GallerySystems*, GallerySystems. TMS Collections Guide, Gallerysystems.Com, 06/17/2023, online at <https://ideas.gallerysystems.com/rs/962-HZY-660/images/TMS%20Collections_web.pdf>.

[144] https://www.gallerysystems.com/solutions/collections-management/ (accessed on 17.6.23 12:04)

described in the chapter above, or by using a static export of the data in a structured file format (might be .xml). While this technicality is not of high importance for us, it is interesting to see what and how the museum publishes its data online, especially because we can from there draw up comparisons on what they share in the GitHub files.

The starting page of the online collection of the museum is dominated by a search field that offers a free text search, additionally to filters. One filter is already pre-set, the “Has image” filter reduces the results from currently 101,747 works online to 88,044 records that have an image.[145] Below the search options, images accompanied by the creator’s name, the title of the object and its date are displayed in a grid view. When one object record is selected a new page opens showing more detailed information. There are some fields that are always published (Artist, Title, Date, Medium, Credit, Object number, Copyright and Department) but others are specific to the departments the object is assigned to. For example, the fields “Publisher”, “Printer” and “Edition” are only visible for objects of the “Drawings and Prints” department. The current exhibitions status is available to each object, either displaying “Not on view” when the object is not exhibited, or “On view” with the detailed location.[146] This detail is also one of the available search filters. Another filter option is to show “uncatalogued objects”.[147] Those records include a disclaimer on the bottom of the object’s web page: “Research in progress; information about this work may be incomplete.” There is an index on the “Classification” feature imbedded in the web presentation of the collection. This enables the users to see

[145] https://www.moma.org/collection/ (accessed on 30.6.2023 10:43)

[146] The object “Entryways” by Diamond Stingily was currently on display in “MoMA, Floor 2, 201” on 30.06.2023. See: https://www.moma.org/collection/works/425481?artist_id=134038&page=1&sov_referrer=artist (accessed on 30.06.2023 11:11)

[147] See for example Pope L.’s series “How much is that Nigger in the Window a.k.a. Tompkins Square Crawl” from 1991: https://www.moma.org/collection/works/292282?artist_id=37145&page=1&sov_referrer=artist (accessed on 20.6.2023 11:04)

which classification the current object is linked to and how many other objects of the same classification are online.[148]

From object records it is possible, through similar indexes as for the "Classification" feature", to go to the associated artists webpages. Next to the name, the nationality and basic biographical data, an overview of all objects of this artists, his/her/their participation in exhibitions, reference to publications and further links to additional material (in the case of Pope.L a link to an interview) are shown, when available.[149] Details of the person are provided based on external resources, in Pope.L's case data is retrieved from Wikipedia and Getty and displayed directly on the MoMA website. The links to the external websites are imbedded as well. For some artists biographies are added and if available also links to audio records and other media files.[150] If records are related to other objects this is represented in the online collection. For example, one print by Siah Armajani is listed as "part of a portfolio with 11 other works" which are online.[151] Specific records are accompanied by texts that describe the artwork together with a note when this label was used, for example Jasper Johns "Target with Four Faces" from 1955 is described by a gallery label from 2009.[152]

On the bottom of every webpage in the MoMA online collection a disclaimer is added stating that the records are work in progress and inviting the user to get in contact with the museum if an error was found.

[148] In Rosemarie Trockel's "Copy Me" object record the index shows that "There are 1,626 sculptures online". By clicking on that text line, the results of all records of the classification "sculpture" are displayed. https://www.moma.org/collection/works/180368?classifications=10&direction=fwd&include_uncataloged_works=1&page=2 (accessed on 30.6.2023 11:46)

[149] Artist page of Pope.L https://www.moma.org/artists/37145#audio (accessed on 30.06.2023 11:16)

[150] See for example Carolee Schneemann's website that includes a biography and an extensive list of additional material. https://www.moma.org/artists/7712 (accessed on 30.06.2023 11:25)

[151] https://www.moma.org/collection/works/430811?classifications=8&include_uncataloged_works=1 (accessed on 23.07.2023 20:17)

[152] https://www.moma.org/collection/works/78393?artist_id=2923&page=1&sov_referrer=artist (accessed on 23.07.2023 20:20)

3.2 The *Artists* and *Artworks* datasets

MoMA provides three datasets via GitHub named *Artists*, *Artworks* and *Exhibitions*.[153] This thesis focuses on the analysis of the first two datasets, while excluding the *Exhibitions dataset*. Although the *Exhibitions dataset* holds valuable insights, two primary reasons underlie its omission from this project. Firstly, the museum itself has undertaken efforts to present the exhibitiona history on their website, featuring "MoMA exhibition history list" encompassing a comprehensive list of exhibitions, accompanied by creator information, publication, archival material like press releases or checklists and installation photographs. Thus, the museum's thorough work in showcasing this dataset makes it less pertinent for immediate analysis within this project.[154]

Secondly, the most pressing inquiries for this research involve a closer examination of the objects featured within the exhibitions. To answer questions such as what kind of objects are on display repeatedly and in which context? Can they be considered as block busters? Do those objects have similarities or shared characteristics, for example are they very easy to handle and not very sensible in terms of climate and light exposure? Who are the creators of those objects and what do they have in common? However, the *Exhibitions dataset* does not encompass information of specific artworks within the exhibitions but only holds information on the artists represented within the exhibitions. Consequently, the comprehensive scrutiny of the *Exhibitions dataset* to extract detailed information on the artworks themselves will be deferred to future analyses, as it falls outside the scope of the current research endeavor.

[153] Cited here once for all references within this thesis: *MoMA*, MoMA Collection - Automatic Monthly Update (11/01/2022), doi:10.5281/ZENODO.7269353.

[154] See Jonathan Lill's description of the Museum of Modern Art Exhibition Index in: *Padilla, Allen, Frost, Potvin, et al.*, Always Already Computational: Collections as Data, 71. See also: https://www.moma.org/research/archives/about-exhibition-history-project (accessed on 04.06.2023 16:11).

The museum uses a public GitHub account.[155] Because of the large size of the datasets the files are treated as large files (LFS) and are stored outside of the main repository. This makes managing the files more efficient and reduces the overall size of the repository. The datasets are available as .json and as .csv files. The account also provides a "readme" file, where the museum's collection is introduced briefly, together with general information on the datasets and how to use them.
The repository has two contributors, "Momadm"[156] and "john-halderman"[157], Momadm has two repositories linked, the collection repository of Tate Gallery and the collection repository of the Carnegie Museum of Art in Pittsburgh, Pennsylvania. john-halderman has two other repositories, both of them not relevant for the MoMA datasets.

Following his LinkedIn profile, John Halderman is "Software Engineering Manger" at the Museum of Modern Art, he is also mentioned in the credits of other digital projects of MoMA, for example the "Modern Women – Women artists at the museum of modern art" website.[158] Although they contributors state that there will be monthly updates, the last update was done on first of November 2022, at current moment this was already eight months ago.

3.2.1 readme

A "readme" file is a fundamental component of most code repositories on platforms such as GitHub. Typically presented as a plain text file, it serves as a critical guide for users, developers, and contributors. The primary function of a "readme" file is to provide an overview of the project or datasets, elucidate the code or files contained in the repository, and furnish instructions on how to use the software or data, as well as guidelines for contributing to the project or repository.

[155] https://github.com/MuseumofModernArt/ (accessed on 23.07.2023 16.08). For more information on GitHub and its use in Digital Humanities see: *Drucker*, The Digital Humanities Coursebook, 207.

[156] https://github.com/momadm (accessed on 23.07.2023 16.08)

[157] https://github.com/john-halderman (accessed on 23.07.2023 16.08)

[158] See the linked in profile: https://www.linkedin.com/in/john-halderman-56523a260?original_referer=https%3A%2F%2Fwww.google.com%2F (accessed on 04.05.2023 16:35). And the website of the MoMA project: https://www.MoMA.org/interactives/modern_women/credits/ (accessed on 04.06.2023 16:41)

Given their vital role in establishing a repository's context and usability, “readme” files are often the first point of contact for users and contributors. Thus, it is imperative to carefully read the information they provide before engaging with the dataset. By doing so, researchers can gain a comprehensive understanding of the project's scope and purpose and ensure that they are utilizing the software or data as intended.

The significance of descriptive information of data was discussed above (see the chapter What are on page 8). Metadata, as it is called, is “used to classify, describe, organize, and connect records to artifacts and documents and to each other.[159] The first paragraph of MoMA’s „readme“ file gives basic information on the museum and the type of objects it houses. They state that, compared to the museum website the datasets contain all object records that are part of MoMA’s database as well as all linked artist records. They add the motivation for sharing the datasets: “MoMA is committed to helping everyone understand, enjoy, and use [their]our collection”, a motif that is inherent of the museums mission since its foundation.[160] The datasets are placed in a public domain using a CC0 1.0 Universal License, it therefore is reusable for everyone.[161] Further down in the “readme” they state how they want the datasets to be referenced, when it is used and provide a unique digital object identifier to do so: DOI 10.5281/zenodo.7269353. They then provide additional usage guidelines. They explain why images are not part of the dataset (because of copyright issues) and give contact details if errors or mistakes are found in the datasets. Once again, they add a disclaimer: “This data is provided ‘as is’ for research purposes and you use this data at your own risk. Much of the information included in this dataset is not complete and has not been curatorially approved. MoMA offers the datasets as-is and makes no representations or warranties of any kind.“[162] This makes clear that the museum is aware or at least suspects their own data to be flawed, unethical or even unsafe. The final paragraph askes the users not to misrepresent the dataset:

[159] *Drucker*, The Digital Humanities Coursebook, 60.

[160] https://github.com/MuseumofModernArt/collection#readme (accessed on 10.04.2023 10:23)

[161] For further information on the creative commons license: https://creativecommons.org/publicdomain/zero/1.0/ (accessed on 10.04.2023 10:38)

[162] https://github.com/MuseumofModernArt/collection#readme (accessed on 10.04.2023 10:23)

> *“Do not mislead others or misrepresent the datasets or their source. You must not use MoMA’s trademarks or otherwise claim or imply that MoMA endorses you or your use of the dataset.*
>
> *Whenever you transform, translate or otherwise modify the dataset, you must make it clear that the resulting information has been modified. If you enrich or otherwise modify the dataset, consider publishing the derived dataset without reuse restrictions.“*[163]

What they don’t provide is a complete list of all the features of the datasets and their descriptions. This is why this will be tried to achieve as the first step of analysis in the section below.

3.3 Exploratory Data Analysis

Exploratory Data Analysis is an approach to analyze data and to gain insights and understanding of the characteristics, connections, and patters within it. EDA was introduced as a field by John Tukey in the late 1970s and his vision was to use “visuals based on descriptive statistics as the primary entry point to new datasets.”[164] It serves as an investigation into unknown data, uncovering what is present, and what is absent. Together with possible limitations or challenges that come with the available data. EDA forms the starting point for all data related work, ranging from a brief look at the data to an extensive and detailed analysis, as demonstrated in this thesis.

There are multiple techniques that are common for EDA, two of them are used in this thesis: summary statistics and data visualization. Summary statics provides knowledge about the values that are available, their type, their range, their occurrences, and their overall quality in terms of classification, standardization, and regulation. On the other hand, data visualizations support the understanding of

[163] https://github.com/MuseumofModernArt/collection#readme (accessed on 10.04.2023 10:25)

[164] Christopher *York*, Exploratory Data Analysis for the Digital Humanities: The Comédie-Française Registers Project Analytics Tool, English Studies 98, no. 5 (07/04/2017) 462, doi:10.1080/0013838X.2017.1332024; John Wilder *Tukey*, Exploratory Data Analysis, Addison-Wesley Series in Behavioural Science (Reading, Mass 1977).

available data through transferring numbers and strings (in our case) into visual representations. For instance, the transformation from a table that is too big to be (close) read, to a graph visualization or box plot offers an overview, a summary of what is there and how it connects to different other features in the datasets.[165] Data visualizations are very common within Digital Humanities and its importance for explaining scientific research to peers as well as to non-experts cannot be underestimated.[166] The advantages of EDA and the results that can be obtained by applying its methods will be shown in this case study.

The following chapter will describe the data understanding part of the thesis that is done by summary statistics. At this stage, the goal is to understand how the dataset is structured, what the features describe and in what kind of data format they are in. If not stated otherwise throughout the whole analysis decimal numbers are rounded, either to two decimal numbers or to integer numbers (if applicable).

3.3.1 Data Understanding

This section describes the process of data understanding. The first insights into the datasets were gained using the software *KNIME*.[167] *KNIME* is a low-code tool for data analysis, data mining and visualizations. It is modular in its structure, allowing the users to choose so called nodes to build up data flows/pipelines in the user-friendly interface. There are hundreds of nodes to choose from, they cover tasks for data integration, transformation, statistical methods, data mining, exporting and much more. Each of the node's properties can be configured to match specific needs and they can be either executed individually, or the complete pipeline as one.
For the task of getting the basic measures on the MoMA dataset only three nodes were used to create an occurrence and a statistics table for both datasets.

[165] Johanna *Drucker*, Graphical Approaches to the Digital Humanities, In: A New Companion to Digital Humanities, Susan *Schreibman*, Ray *Siemens*, John *Unsworth* Ed. (Chichester, UK 2015) 238, doi:10.1002/9781118680605.ch17.

[166] *Windhager, Federico, Schreder, Glinka, et al.*, Visualization of Cultural Heritage Collection Data, 2316.

[167] Michael *Berthold*, KNIME The Konstanz Information Miner, Java (Zürich 2004), online at <https://www.knime.com/>. See also https://www.knime.com/ (accessed on 23.06.2023 16:28)

- File reader node. It reads in the .csv file. The default settings were unchanged, only the limit for the data rows that are scanned was increased to 10 million, otherwise not every record would have been read in correctly.
- Statistics node. All columns were included in the settings, and the number of possible values per column were increased, so that also for columns with only unique values a result was achieved (e.g., IDs).
- Excel Writer Node. The Node is linked to two different ports from the Statistics node, and it based on the port it created a .xslt table with the statistical measures of the dataset or the occurrences of the values within the dataset.

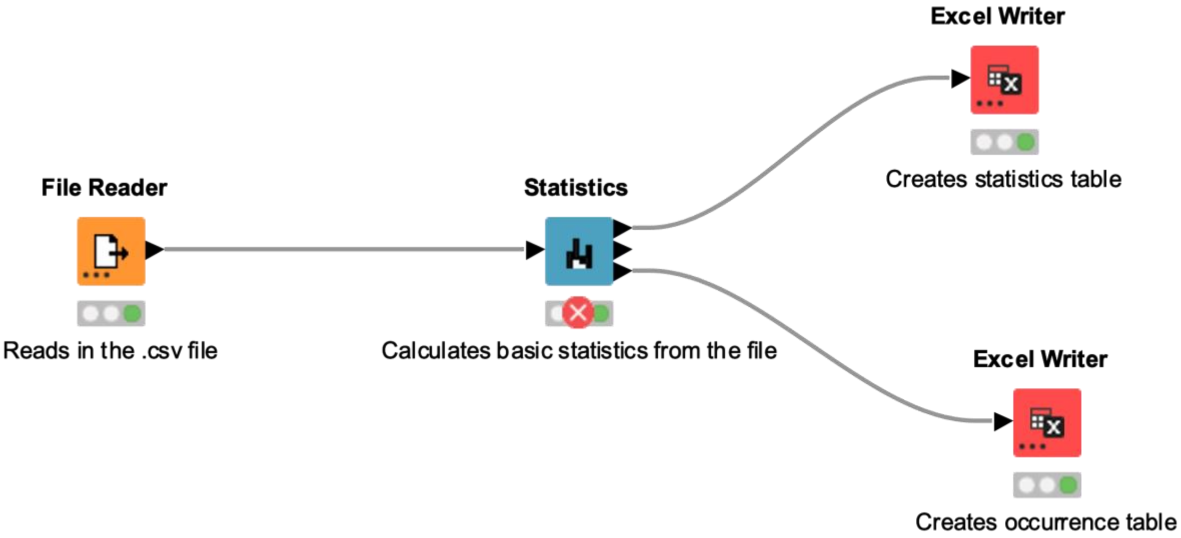

Figure 3 Export from KNIME Workspace. The three different nodes that were used to create the statistics and occurrence tables.

The goal of this step is to understand each feature of the datasets, to get a feeling for the quality of the data, and what steps would be necessary to use the data. At the end it should be clear what data there is and if and how it can be relevant for our endeavor. Furthermore, it should help to decide from there what research questions are possible to be answered based on the available data. This chapter has a note-taking character to it but is summarized at the end to provide a better readable version of the results.

The following information on the *Artists* and *Artworks dataset*s were extracted from the occurrence and statistical tables created by *KNIME*.

The *Artists dataset*

In the following analysis, each feature of the *Artists dataset* will be examined and summarized. This serves as the starting point for the subsequent data visualization process and aims to stimulate the identification of possible research questions that might be answerable through and with the dataset.

The *Artists dataset* contains 9 features: *Constituent ID*, *Display Name*, *Artist Bio*, *Nationality*, *Gender*, *Begin Date*, *End Date*, *Wiki QID* and *Ulan*. Each artist record represents one row.
On 10th of April 2023, the artist file was 976 KB large, consisting of 15243 rows.

Constituent ID

All the *Constituent ID*'s are unique within the dataset. The values are numeric.

Display Name

There are 23 nonunique names, "Unidentified Designer" appears 22 times, "Various Artists" nine times, "Unidentified Artists" four times, John Wood and "Unidentified designer" 3 times, 18 other names appear twice.

Artist Bio

This feature has 7248 unique values. The values consist of the *Nationality* value, a comma and either the word "born" followed by the birth year (*Begin Date*) or, if also a death year (*End Date*) is available, the years separated by a binding line. If there are no dates available, only the *Nationality* is repeated. Some entries use the abbreviation "est." followed by a year. Assumingly, those values belong to artist groups or organizations and not to artists records representing single people.
The entries get more complicated if the artists were born in a different country, then where he/she/they are citizens. This is for example listed as "American, born Lithuanian". Because of these exceptions it is assumed that the data is not a standardized combinations of preexisting data fields but a free text field that is entered individually.
2215 records don't hold values in the *Artist Bio* Column, which is 14% of the records.

Nationality

There are 119 values for the *Nationality* feature assigned to the artist records, the largest group is “American” with 5181 records, followed by records without any *Nationality* (2472 records, 16%). 161 records hold the assigned value “Nationality unknown”. 37 values only appear once or twice within the whole dataset.

Gender

There are 6 different *Gender* values: Male, Female, male, Non-Binary, female, Non-binary. The different spelling of the values can be unified, so that we end up with three categories: Male, Female and Non-Binary. 9731 artist records have “Male” assigned as *Gender*, 2342 “Female”, and 3 “Non-binary”.
3165 artists have no *Gender* category assigned to them, which is little over 20%.

Begin Date

Numeric data. 3642 records don’t hold a *Begin Date*, which is understood to be the birth year of the artists. The remaining 75% hold data.
1942 is the *Begin Date* that occurs the most often, 186 times.

End Date

Numeric data. 10074 records, 66%, don’t hold data in the *End Date* feature, which is understood to be the death year of the artists. The date that occurs the most times is 1991 with 88 appearances.

Wiki QID

11994 records don’t hold a *Wiki QID*, a combination of letters and numbers, which is 78% of the dataset.

Ulan

12311 records don’t hold an *Ulan* id, a combination of numbers, which is 80% of the dataset.

Summary

Every record has a unique ID, but not every name is unique. Furthermore, there are names that represent unknown names, some of them specify the type of medium the creator’s name is linked to, like “unknown designer”. It is not yet clear how the values of *Artist Bio* Feature get created, they seem to be combined from different source fields, but it was not possible to find a pattern in how it is put together from the available features. More than 80% of all creators have a *Nationality* assigned to them. There are some values that should be examined more closely, for example “Native American” or “Canadian Inuit”, also nationalities that stem from states that don’t exist anymore, like “Czechoslovakia” should be revisited. The three different *Gender* values appear each twice, when they are merged, we can see that almost 65% of the creates are categorized as male. *Begin Date* and End Date are considered to be the birth and death year of the creators. Almost 70% of the artists have no death year value assigned to them, it will be examined further if all those creators are still alive, or if the death year was not catalogued. *Wiki QID* and *Ulan* reference external sources. *Wiki QID* is the unique identifier of Wikidata, *Ulan* stands for the United list of Artist Names, an authority vocabulary that is provided by the Getty Research Institute.[168] Both provide permalinks that can be used to consistently link to the records within their databases, and, potentially with other databases, should the MoMA decide to do so. As we have seen above on page 50, MoMA uses the links to retrieve external data to their artists web pages. The advantages of authority files were already discussed in the chapter Categorization and Classification. It will be a task in the further analysis to see, if the data in the MoMA records is coherent to one of the authority files (e.g., the *Nationality*) or if MoMA is, additionally to linking their records to external sources, adding metadata on the persons in their own system.

The *Artworks dataset*

The *Artworks dataset* contains 30 features: *Title*, *Artist*, *Constituent ID*, *Artist Bio*, *Nationality*, *Begin Date*, *End Date*, *Gender*, Date, *Medium*, *Dimensions*, *Credit Line*, *Accession Number*, *Classification*, *Department*, *Date Acquired*, *Cataloged*, *Object ID*, *URL*, *Thumbnail URL*, *Circumference (cm)*, *Depth (cm)*, *Diameter (cm)*, *Height (cm)*, *Length (cm)*, *Weight (kg)*, *Width (cm)*, *Seat Height (cm)*, *Duration (sec.)*. Each artwork record represents one row in the dataset. The features, *Constituent ID*, *Artist Bio*, *Nationality*, *Begin Date*, *End Date* and *Gender* are originally part of the

[168] As explained here: https://www.wikidata.org/wiki/Q43649390 (accessed on 26.06.2023 11:29). See also https://www.getty.edu/research/tools/vocabularies/ulan/ (accessed on 26.06.2023 11:26).

Artists dataset. In the *Artworks dataset* those features are merged in, if there are multiple artists linked to an object, the info of the different artists is merged into one field. The two datasets can be linked based on the feature *Constituent ID* that is unique and mentioned in both datasets.
On 10th of April 2023, the artist file was 59.2 MB and consist of 140.848 Rows.

Title

Every record holds data for the *Title* feature, there are multiple terms to represent untitled works. The most frequently used title is “Untitled”, which appears 8526 times. Art works can be untitled because the artist explicitly decided to not title the work, or because the title is unknown. There is no distinction made between these two versions of “untitled” artworks within the available dataset.

Artist data

The data for the artists that are linked to the objects is merged in from the referenced artist record (in a one-to-many relationship). As unique reference key the *Constituent ID* is used. The fields that are filled are “Artist”, which is the *Display Name* in the Artists.csv file, and the features *Artist Bio*, *Nationality*, *Begin Date*, *End Date* and *Gender*. For specific details on those features see the data understanding abstracts on the *Artists dataset* above. Multiple occurrences of linked artists, when more than one person was involved in the creation of the objects, are merged into one entry. The specific values are separated differently for each of the features. Because of this merge of occurrences, the analysis of artists data in the *Artworks dataset* is hindered. But an overview of the features will be provided regardless:

Artist:
Separation of multiple occurrences with a comma and a whitespace. The entry itself is then enclosed with quotation marks.
Example:

"Robert Brownjohn, Ivan Chermayeff, Thomas Geismar"

The most frequent name is “Eugène Atget” who is the sole creator of 5050 artworks. Followed by “Louise Bourgeois” who is linked to 3336 records. The third most frequent term is a substitution: “Unidentified photographer” is linked 2736 times.

Constituent ID:

Separation of multiple occurrences with a comma and a whitespace. The entry itself is then enclosed with quotation marks.

Example:

"30473, 30474"

Artist Bio:

Separation of multiple occurrences with a whitespace, each occurrence is enclosed in parenthesis. The complete entry is enclosed with quotation marks, although there are exceptions where no quotation marks are there.

Example:

"(American, born Scotland. 1821–1882) (American, born Ireland. 1840–1882)"

The most frequent term is a NaN value which is displayed for this feature as two quotation marks. This appears 5823 times and means that none of the linked artists holds data for the *Artist Bio* feature.

Nationality:

Separation of multiple occurrences with a whitespace, each occurrence is enclosed in parenthesis. There are no quotation marks.

Example:

(Italian) (French) (Hungarian)

With a relative frequency of 0.4 % "(American)" is the nationality assigned the most frequently to the artworks (58345 times). Followed by "(French)", "(German)" and "(British)". 5004 records don't hold any data for the *Nationality* feature.

Begin Date and *End Date*:

Separation of multiple occurrences with a whitespace, each occurrence is enclosed in parenthesis. There are no quotation marks. Empty values are replaced with a null in between parathesis.

Example:

(1924) (1926)

(0) (1881) (1881) (0) (1885)

8037 records (0,06 %) don't have any data recorded for the *Begin Date* and 45382 records (0,32%) are missing an *End Date*.

The most frequent *Begin Date* is "(1857)" which is recorded 5111 times, followed by "(2010)" which is recorded 4174 times.

Gender:
Separation of multiple occurrences with a whitespace, each occurrence is enclosed in parenthesis. There are no quotation marks.
Example:

(Female) (Female)

If data is missing the occurrence is still there but empty.
The most frequent assigned *Gender* feature is "(Male)", which appears 105602 times and represents a relative frequency of 0.75%.
Followed by the much less used category "(Female)" which only appears with a relative frequency of 0.13% (18824 times). 7257 records (0.05%) don't hold any data for the *Gender* Feature.

Date

Is a text field and holds dates with details about the production. There are 2099 records without data for the *Date* feature (0.01%). The most frequent entry after that is "1967" which appears 1866 times. There are also entries like "n.d." which is assumed to be an abbreviation of "no Date" (639 times), "(n.d.)" (112 times), or "Unknown (242 times). In order to work with this data, it would need to be cleaned and standardized. The following example entry shows that this would not be a trivial task to do because a lot of semantic that is not easy to automatically extract lies in the entries: "1945-48, published 1948".

Medium

The *Medium* feature holds textual descriptions of the artwork, the entries are not standardized, there are 21643 unique values listed. Some of the entries are enclosed within quotation marks. The most frequent term is "Gelatin silver print" which appears 16638 times which is a relative frequency of 0.12%. Second in the line are the records without no value for *Medium*: 9632 records, 0.07%, are missing an entry.

Dimensions

The *Dimensions* feature holds free text regarding the measurements of the objects. It consists of a combination of the values from the features *Circumference (cm)*, *Depth (cm)*, *Diameter (cm)*, *Height (cm)*, *Length (cm)*, *Width (cm)* and *Duration (sec.)* with additional free text. The cm values (decimal numbers are

rounded to one number) are enclosed in parenthesis that follow the inch values. There are values with additional text, for example “Each: 14 x 18” (35.6 x 45.7 cm)” or “Object stacks into an 18” (45.6cm) cube”.
If there is data for Diameter(cm) available, it is not specified in the Dimension feature as such. It appears in the same way as would the display of the values “height x length”. For example, in this case 17.8 is the diameter but it is not recognizable as such:

7 3/4 x 7" (19.7 x 17.8 cm)

The data from the *Weight (kg)* feature is not included in the *Dimensions* feature. See further below for more detail on all the specific dimension features.

Credit Line

The values are free text entries. Data is available for almost all records, only 1.3% of the artwork records don’t hold data for the *Credit Line* feature.
“The Louis E. Stern Collection“ is, with a relative frequency of 0.8%, the most used term and appears in 11258 records. Followed by the terms “Gift of the artist” which appears 10620 times and “Purchase” which appears 8396 times.

Accession Number

All but 3 records hold a “*Accession Number*”, there are 11 erroneous records were instead of a number “Photography” is entered in the field. All other numbers are unique. The field is a string field, in most (but not all) cases containing a combination of numbers and characters.
This feature is labeled as “Object number” on the MoMA website and it can be assumed that there is some semantic imbedded into it. When compared with the acquisition year that is stored under “*Date Acquired*” a match of the acquisition year and the second number part of the *Accession Number* is recognized for a majority of the records. For example, the record with the *Accession Number* “488.1986” was acquired in the year 1986. Presumably the first part of the number represents the number of acquisitions that were already catalogued in that year.
But there are other records that don’t follow that schema. Some hold additionally to the numbers also character combinations. There are, for example, 61 records with an *Accession Number* starting with “MC” or “ASCM” followed by integer numbers.

Classification

38 different values appear as classification, plus one record without data in the field. The largest group is “Photography” with 34100 records, which is a relative frequency of 0,24%. Followed by “Print”, 32130 records and “Illustrated”, 27491 records. The classifications “Fashion”, “Software”, “Architecture” and “Document” only appear once. There are three faulty entries: “39.8”, “20.5” and “22.86”.

Department

There are 8 different values for the *Department* feature listed. 3 records don’t hold any data, eleven others hold the value “Y”, which seems to be a faulty entry.
The largest group is “Drawings & Prints” with 77771 records, which sums up to 0.55 relative frequency. The second largest group is “Photography” with 32965 records. “Film” and “Architecture & Design – Image Archive” appear at least often.

Date Acquired

Almost all records hold data for the *Date Acquired* feature which is understood to be the data the artwork got acquired by the museum. 6685 records are the exception and don’t hold any data (0.05 %).

Cataloged

All records but one hold data for the field *Cataloged*. 66% hold “Y”, the rest holds “N”. 13 other records hold faulty data in the field (*URL*s or floating-point numbers).

Object ID

All but three records hold a unique *Object ID*, consisting of integers numbers counting up from 1. Three records don’t hold any data.

URL

This feature holds URLs, uniform resource locators, to the web presentation of the artworks on the MoMA website. The path consists of the landing page http://www.moma.org/collection/ followed by “works/” and then an increasing integer number. For example: http://www.moma.org/collection/works/2. Interestingly this number is not the *Object ID*. Both lists of numbers hold missing positions (numbers missing) and diverge.

Thumbnail URL

This feature holds URLs to the visual web presentation of the artworks on the MoMA website. The path consists of the landing page http://www.moma.org/collection/ but instead of "works" for the *URL* feature, "media/" is selected followed by a string of characters that appear to be media file names. Not all values are unique, some records link to the same media. 40% of the records don't hold any value for the *Thumbnail URL* feature.

Circumference (cm)

Only 10 records hold data for the feature *Circumference (cm)*. All the remaining records don't hold any data.

Depth (cm)

0.89% of the records have no data for the *Depth (cm)* feature. The remaining ones hold floating point numbers, with up to 9 decimal numbers.

Diameter (cm)

0.99% of the records don't hold data for the *Diameter (cm)* feature. The remaining ones hold floating point numbers.

Height (cm)

Most of the records hold *Height (cm)* data, only 0.12% have no values. The most frequent *Height (cm)* is "0", which is catalogued for 2512 records.

Length (cm)

Almost all of the records are missing data for the *Length (cm)* feature, 0.99%.

Weight (kg)

Also, this feature is almost always not catalogued. 0.99% don't have any values.

Width (cm)

Most of the records hold *Width (cm)* data, only 0.13% have not values assigned to them. The most frequent *Width (cm)* is "0", which is catalogued for 2364 records.

Seat Height (cm)

There is no data for the feature *Seat Height (cm)*.

Duration (sec.)

0.99 percent of the records don’t hold any value for the *Duration (sec.)* feature. The most frequent *Duration (sec.)* is “120”, which appears 228 times.

Summary

The features in the dataset differ in their quality and relevance. They can be put together into six information groups:

1. fields regarding the creator(s) (*Constituent ID*, *Artist Bio*, nationality, *Begin Date*, *End Date*, *Gender*)
2. information on the object identification (Title, Date, *Medium*, *Classification*)
3. information on object logistics (*Department*, *Object ID*, *Cataloged*),
4. information on physical characteristics (*Dimensions*, *Circumference (cm)*, *Depth (cm)*, *Diameter(cm)*, *Height (cm)*, *Length (cm)*, *Weight (kg)*, *Width (cm)*, *Seat Height (cm)*, *Duration (sec.)*
5. information on acquisition details (*Credit Line*, *Accession Number*, *Date Acquired*)
6. and their publication status (*URL*, *Thumbnail URL*).

All fields of group 1 are retrieved from the linked artist records, it is unclear why the merging of the specific values is not done in a standardized way but is done based on individual rules for each of them.

This raises questions of constituency and overall quality of data, something that, as we have seen in the chapter "Collecting and cataloguing cultural objects" and specifically in the discussion of the complex process of recording cultural objects in collection management systems, is an undertaking that is more difficult than it might seem. The inconsistencies might be the result of different data structures that could not easily be put together in homogeneous ways or of old habits that are ingrained in the workflows of the staff. It might also just be an example of how small decisions of individuals (on how to merge things) that are not consistent affect the quality of data on a larger scale and are hard to undo.

Group 2 consists of features that hold unstructured data, with the exception of *Classification* that provides us with a list of unique values. The title feature holds data for every record, but there are multiple placeholders for objects without a title, like “untitled”. The *Date* feature would need to be cleaned in order to work with it properly, originally it holds a mix of string and numerical values. The *Medium* feature is not standardized but there are values that appear repeatedly. It will be interesting to see what kind of values that are, and which objects they are linked to.

Group 3 holds the *Object ID*, which is unique throughout the dataset, although surprisingly there are three records without it, those can probably, after closer examination, be discarded for the analysis. The *Department* feature is structured to hold only 8 different values. There is a mismatch in numbers between the little appearance of the "Film" *Department* compared to the by MoMA stated "more than 30,000 films and 1.5 million film stills"[169] in its possession. After the first analysis it can be assumed that those are not part of the available datasets, but it might be that there are placeholder records or set/collection datasets that represent more than one object. This needs to be further investigated.
It is unclear what the Catalogued feature represents. Maybe the values "N" and "Y" represent a form of cataloguing status, "Y" meaning something like "yes, this object record is complete" and "N" meaning "no, this object record is not finished yet". But this is only an assumption, and it needs to be further investigated. Maybe there can be a correlation found between the completeness of the records and the *Cataloged* feature.

Group 4 consists of the *Dimension* feature that is a kind of summary of all other measurement fields, although the rule about how the values are combined is not obvious on the first glance. The values are floating point numbers with multiple decimal numbers. This is probably due to initially recording of the measurements in inch and an automatic transformation to centimeters. This assumption is supported by the fact that the online presentation of the artworks shows the inch and (rounded) cm values, and the inch values are mostly integers numbers.[170] Most records only hold data for the *Width (cm)* and *Length (cm)* feature. The other features can be

[169] https://www.moma.org/about/curatorial-departments/film (accessioned 26.06.2023 13:08)

[170] See for example Kara Walkers "African Boy Attendant Curio with Molasses and Brown Sugar, from "The Marvelous Sugar Baby" Installation at the old Domino Sugar Factory Warehouse. (Rear Basket) from 2024. The following dimension data is displayed: 59 x 25 x 33" (149.9 x 63.5 x 83.8 cm). https://www.moma.org/collection/works/190540?classifications=10&date_begin=Pre-1850&date_end=2023&q=&utf8=✓&with_images=1 (accessed on 26.06.2023 13:41)

ignored because they hold to little data to work with them, *Seat Heigth (cm)* is completely empty.

Group 5 provides information on the acquisition of the object. The *Credit Line* is an unstructured field holding textual data. The second most often appearing entry is “Gift of the artist”, which could, together with the recorded artist name, be further analyzed. The *Accession Number* seem to be more than just an identifier but to hold semantic meaning. It might be possible to find out some of it, but to fully understand the meaning insider knowledge is probably needed. *Date Acquired* holds date data (YYYY-MM-DD).

Group 6 provides information on the object’s publication status. If they hold data in the field *URL*, the object record is accessible in the web as part of MoMA’s online collection. If there is also a *Thumbnail URL* available there is also an image online. The links work, a hand full of random examples were successfully tested on 05.06.2023. It will be interesting to see what kind of objects are presented online with their catalogued metadata but without images and what the reasons for this might be. With the provided *URL*s the museums website could be crawled to get the additional information, like links to publications or exhibitions. But this is something for future projects since it would exceed the scope of this thesis.

3.3.2 What is not in the datasets?

The data understanding process of the *Artists* and *Artworks datasets* showed what kind of data is made available by MoMA. This section will focus on what is not included in the data, what is missing and where the gaps are that make deep analysis difficult.

At first, we will compare the available data with what the “readme” file has promised: although introduced, there is no evident tag or note or anything similar that could be used as a hint to indicate that the specific record is not “curatorial approved”. Also, against what has been promised by MoMA, the datasets do not represent the complete collections of MoMA. As has been shown above at least the film and film still collections are not represented entirely. Additionally, the constituencies records, that were described by Padilla, and which should be recognizable on specific postfixes in the *Accession Number* feature were excluded from the datasets. While it might be the case that the film and film stills are not yet catalogued or stored in a different database (which might be plausible because time-based media records often come in a different structure), the exclusion of the constituencies records must have been an active decision by the museum during the publication process of the datasets.

As has been stated, MoMA uses TMS to manage their collection. This software holds a large variety of fields and different modules and is structured as a relational database. Every .csv or .json file that is created to store exports from this database is bound to limitations, because those files can not represent the same information in the same structure as a relational database. Due to this, and also to the need of the museum to make sure not to publish sensitive information, it was necessary to handpick the features to be published. Data can be sensitive out of various reasons. For example because it holds internal knowledge that is not meant to be shared with the public like insurance values, location data, condition, or hazard information of artworks. It can also be considered sensitive because private data would be disclosed like the names of staff, the addresses of artist, donors, restaurateurs etc. Or sensitive because it might be ethically or politically problematic, like descriptions or label texts that use words that are outdated today. Other fields might have been excluded because the aim was to provide datasets in reasonable size and the specific features were regarded not so important by the decision makers. It might also be that fields were excluded because the data quality was declared to be too bad for publication. It is a pity that all those things are not addressed by MoMA and that we therefore are limited to assumptions on what has happened.

In order to see what features might be available in the database but were not included in the datasets, we will at first compare the available features with the *Object ID* standard of Interpol. Secondly, the fields will be compared with those that MoMA publishes on their online collection. As we will see, they publish more data there than in the datasets, which is a regretful fact and limitation on possible research on the collection.

If we compare the available features with the *Object ID* norm of ICOM and INTERPOL introduced above in the chapter “Categorization and Classification”, we find that not all of those basic information groups is included in the datasets made available by MoMA. The information categories that are necessary to fulfil the norm are as follows: type of object, materials and techniques, measurement, inscriptions and markings, distinguishing features, title, subject, date or period and maker. The type of the object can be gained from combining the *Medium* and the *Classification* feature. Those two features might also provide assumptions on what material and technique were used, but only with expertise knowledge of the artworks in questions and not with certainty. For example, it is an easy guess, that a record with “Gelatine Silver Print” as *Medium* with “Photography” as *Classification* describes an object that consists of paper and printing ink and was created by applying photographic techniques. But this guessing turns to high uncertainty quickly. When we have “Model” as *Medium* and “Architecture” as *Classification* we can only know what material and technique was used, when we know the specific artwork. The measurements are provided, although in the vast majority of the artworks only hold data for width and length. Information on inscriptions and markings are not part of the available dataset, the same goes for any “distinguishing features”. Those are considered features that make the object unique, a stain, for example, or a scratch, or a specific mark that could help to identify the object if every other identification is missing. Title, date, and the maker are part in the datasets. But information regarding the subject of the artworks is missing. The dataset does not include any subject related field. This might be fields were meaning or interpretation of objects are recorded, were iconographical or iconological information is catalogued or where descriptive texts are stored. Overall, the available datasets would not meet the standard of INTERPOL, but we can assume that all the relevant data is collected and stored by MoMA internally.

If we now compare the features in the datasets with the data that is published on MoMA’s online presentation, see page 50, we find that the museum publishes more data online than in the datasets on GitHub. As we have seen, there are additional fields based on the *Department* artworks are assigned to. For example, the publisher of records of the Prints & Drawings *Department*. The online collection also shows for some of their records accompanying texts and their source. Neither the text nor the reference to when the text was written is part of the datasets. The described feature of showing if objects are part of a larger set, like drawings or prints in a portfolio, is also only available in the online collection but missing in the datasets. Furthermore, the current exhibition status as well as links to further resources like, videos, audio files, texts, publications and so on are only available in the web presentation of the collection but not in the dataset. The “uncatalogued object” filter, which shows records with incomplete content is also not available in the datasets. To summarize, the datasets do hold more records as are represented in the online collection, but the information available online is more diverse and richer.

What the read me file does not clarify, but what is very important for the work with the dataset, is that the record count does not represent the count of artworks. Based on my own experience in the field, I can say that, in all collections, I know there are object records that represent multiple artworks at once and others that represent a collection, set or group of objects, although each of the parts are already represented with a record. The record count alone does not allow us to retrieve the count of artworks within a collection. Some museums have fields in their databases that allow the entry of part numbers and to specify the type of records (records of single objects, sets, collections, parts etc.). But even if this data is available the extraction of an exact number of artworks stays a complex task. The available *Artworks dataset* does not provide either of that. But in the *Accession Number* feature we can observe that also in MoMA the way of how objects are catalogued is not consistent. The *Accession Number* “179.1962.1-28”, for example, represented a “Portfolio of twenty-eight lithographs” by Anatoli Lvovich Kaplan[171], each of those lithographs is represented with individual artworks records with the *Accession Numbers* 179.1926.1, 179.1926.2 and so on. This means that the 28 lithographs are represented with 29 records. On the other hand, there are *Accession Numbers* like “251.1934.1-2”, “Serving Trays” by Walter Von Nessen, where one record represents multiple, in this case two, art objects and there are no additional records representing each one of them. Here we have one record

[171] https://www.moma.org/collection/works/portfolios/12764 (accessed on 20.07.2023 15:02)

representing two artworks.[172] Unfortunately, we can’t solve this issue with the information provided in the dataset. Therefore, it is important to remember throughout the analysis that the artwork record count does not necessarily equal the artwork count. This might also affect all calculations that are done, like gender distribution of acquisition numbers of specific directors.

What is also truly missed is of more organizational matter: a detailed description of the features and how their data was created. As already shown above in the chapter “What are ”, detailed documentation of how data was recorded, structured, and used is essential for any further analysis and use. There are some fields where we are left to assumptions on what it might represent. This is a missed opportunity that could be done better. Other museums, which did similar publication projects as MoMA, include those extensive lists of descriptions.[173]

In summary, the greatest shortcoming of the datasets is their lack of metadata description and documentation. A concise description by the museum of the specific features and their significance and why certain features were selected for publication while others were not, would greatly enhance the usefulness of the dataset. Currently, it is difficult, virtually impossible without insider knowledge, to determine the exact interpretation of specific features. What also would allow for more profound and farer reaching research would be including the fields that are published online to the datasets. There should not be any sensitivity concerns here, since the features are published already, but not in such an easy to access way as in the datasets.

Overall, the available datasets provide enough data for primary research, but important information is missed.

[172] See for a third example the Accession Number 255.2014.1-20, representing Mladan Stilinović’s “Exploitation of the Dead”. In this case there are no additional records even though the artwork consists of multiple smaller pieces. See: https://www.moma.org/collection/works/181109 (accessed on 20.7.2023 15:34).

[173] The example is from the Carniege Museum of Art and its publicly accessible collection datasets. See: *Padilla, Allen, Frost, Potvin, et al.*, Always Already Computational: Collections as Data, 29.

3.3.3 Data Analysis

Prior to presenting the analysis steps and its outcomes, the software *Tableau*, which was used for this analysis will be introduced. Followed by a section on insights that are expected to be retrievable from the available dataset.

Tableau

Tableau is an interactive software for data visualization.[174] It is based on Edward Tufte's definition of "visual display of quantitative information".[175] The basic principles of visualizations, introduced by Tamara Munzner, are implemented and give users an easy but still elaborate tool to work with data on a visual level.[176] Originally started as a research project at the Stanford University, the successful product was purchased by *Salesforce* in 2019. *Salesforce* develops and sells cloud-based software to companies; *Tableau* is one product out of their large product palette.[177]
The interface of *Tableau* is based on drag and drop functionality and shows the results of edits and changes in the visualization pane immediately. Visualizations can be grouped to so called dashboards and even added to "data stories" that allow to narrate the visualizations and guide users in their use of available interactive filters, highlighters, and selections. Many features that the software offers, most prominently all interactivity, cannot be transported with images, as they will be presented in this thesis. This is unfortunate but the images should be sufficient in transporting the findings and might invite readers to explore the dataset on their own in *Tableau* (a free trial version is available).[178]

Tableau provides the possibility to link multiple data source. To do this, a unique identifier has to be specified that appears in the datasets that are to be linked. In our

[174] *Chabot*, *Stolte*, *Beers*, *Hanrahan*, Tableau.

[175] *Drucker*, Graphical Approaches to the Digital Humanities, 239.

[176] *Munzner*, Visualization Analysis and Design.

[177] https://www.salesforce.com/de/products/what-is-salesforce/ (accessed on 08.07.2023 09:02)

[178] For detailed instructions and guidance on how to use Tableau see: Alexander *Loth*, Datenvisualisierung mit Tableau, 1. Auflage (Frechen 2018); Daniel G. *Murray*, Tableau Your Data! Fast and Easy Visual Analysis with Tableau Software®, Second edition (Indianapolis, Indiana 2016).

case, this is the *Constituent ID* of the artists records that is repeated in the artwork dataset. As already explained above, if there are multiple artists linked to one artwork record their IDs are combined.[179] For linking the two datasets together in *Tableau*, these values had to be split. Only the first occurrence of the *Constituent ID* was used because *Tableau* does not allow the use of multiple fields as an identifier. Unfortunately, this means that we lose data additional linked artists on all objects made by multiple artists. In every analysis regarding objects and the data of their creators, only the attributes of the first listed artist are recognized. However, for larger-scale analysis, this does not significantly affect the results and can therefore be accepted as the following analysis shows: From the complete dataset 7.741 object records have more than one artist linked to them. 4.687 of those (more than 60%) are linked to multiple male artists or at least one male artist and others who's gender is not categorized. They will appear in *Tableau* as created by a male artist and therefore will be clustered correctly regarding the *Gender* feature.

2.871 records are linked to at least one male and at least one female artist. 1.224 records list the male artist as first, which means that they appear in the split-up version in *Tableau* as created by male artists.

1.122 records list one female artist as first. This evens out almost completely and will not affect our analysis of gender representation considerably.

284 records are linked to multiple female artists or one female artist and one artist who's gender is not categorized; their records will be assigned to the correct *Gender* group in the analysis.

525 records list an artist in the first position who does not have a gender attribute assigned. In those instances, we lose information, but regarding the small number of the records (0.37%), this is an acceptable loss.

183 records are linked to multiple artists whose genders are not specified. They will be assigned to the correct gender category (no gender). This limitation should be remembered throughout the analysis and also that, as has already been explained, the record count does not equal the artwork records.

[179] This difficulty is described also by Padilla and others in their case study of the Carnegie Museum of Art. See: *Padilla, Allen, Frost, Potvin, et al.*, Always Already Computational: Collections as Data, 27–28.

Beforehand Expected Insights and selection of research questions

As the scope of this thesis is limited, it is therefore crucial to limit the questions that will be further analyzed. The following questions will examine in more detail in the following sections:

To initiate the analysis, the *Artists dataset* will be examined, with a focus on their *Nationality* and *Gender*. The primary objective is to extract insights into the gender distribution of artists associated with the artwork records and their respective nationalities. Particular attention will be given to terms that may diverge from the conventional assignment of citizenship or hold historical meaning. These terms, that have been introduced already above, will be investigated in detail. One brief paragraph section will also explore the birth and death dates of the artists, aiming to identify patterns between their age and their appearance in the datasets. Following the initial section, a subsequent analysis will focus on the artists yet again, but this time in combination with the artworks records they are linked to. This should allow a gain in deeper understanding of the representation of and potential gender-related trends artists in the museum's collection. The primary questions to be addressed are as follows: What is the gender distribution of artworks creators? Are there more artworks attributed to male artists compared to female artists? When were the first artworks of non-binary artists acquired, and is their number increasing? On average, how many artworks are attributed to each artist in the MoMA collection? Does this number differ based on the gender categorization of the artists?

The next part of the analysis will investigate the artwork records, based on the features *Department, Classification* and *Medium*, with a goal to find insights on how the MoMA collection is put together and how the different objects are organized within the database.

The primary focus of the analysis revolves around the acquisition history of MoMA. Investigating the feature *Date Acquired* to gain insights on how, when and what artworks the museum acquired. The following key questions will be explored: Is it possible to discern discernible patterns within the acquisition history of the collection? To what extent can the datasets enable differentiation between the acquisition policies employed by various directors? Did the directors exhibit divergent collection practices concerning the gender of the artists? Or did their focus shift between the different *Departments* of the museum?

The section will be expanded with the search for specific moments or periods in MoMA's history and the question if they are retraceable in the datasets. The first one is the Second World War; the assumption is that before and during the war artists fled Europe in order to escape the murderous NS-regime. It will be examined if this, maybe in an increase of European artists in this time, is findable in the dataset. Secondly, the Cold War and its potential impact on the museum politics will be examined. The assumption here is that the museum hesitated during this period to acquire artworks from Russian artists. An increase of Russian artists and artworks might be traceable after the conflict ended. After these global events two events which directly affected MoMA will be investigated. The one is the acquisition of the Lilli P. Bliss collection in the early years of the museums. This is, in the narration of MoMA, a key moment within its development and on its way to today's recognition. Secondly a look at an important protest the W.A.V.E from 1984 will be examined. Did the acquisition policy change after these protests? Did it actually effect the acquisition of female artists and artworks positively? To conclude the analysis the feature *Cataloged* will be investigated, due to the uncertainty of what this feature represents the curiosity is large to find correlation to other features and possible explanations.

It is important to emphasize the recognition of serendipity during this process, accompanied by curiosity and openness to unexpected insights, questions, and directions that have not been predefined. These qualities are valued and encouraged throughout the analysis process, as they hold the potential for valuable and unforeseen discoveries. If stumbled upon interesting findings, they will also be briefly included.

In Depth Analysis

MoMA's artists

The artist dataset provides a feature named "*Nationality*" and as we have seen in the Data Understanding part of this thesis, there are some terms recorded that cannot directly be translated to Nation states, although most of the records do without any difficulties. In order to use geo references in visualization in *Tableau,* it was necessary to assign each *Nationality* term to a nation state. "French" was, to name one example, assigned to the state France. For the terms that did not fit that logic decisions had to be made on how to deal with them. Before this is explained in more detail, a detailed examination of the deviating terms will be provided:

The terms in question are “Candian Inuit”, “Catalan”, “Coptic”, “Czechoslovakian”, “Korea”, “Native American”, “Persian”, “Scottish”, “Welsh” and “Yugoslav”. There are 76 records that hold those categories.

- Interestingly there is the name “Unidentified designer” that appears three times with three different unclear nationalities: Coptic, Czechoslovakian and Persian.[180]
- Three persons are categorized as “Canadian Inuit”.[181]
- Only one person is marked as “Catalan”.[182]
- There are 5 records with “Czechoslovakian” as *Nationality*.[183]
- Ten persons are marked as “Native American”.
- 20 person records are listed as “Scottish”.
- Three as “Welsh”.[184]
- One person is recorded as “Yugoslav”.[185]
- There are 31 records of “Korean” artists. While some of them can be identified of being South Korean through online research, there are some where no further information can be found or like in the example of “Ung-no Lee” (*Constituent ID* 3457) or “Chung Chang-Sup” (*Constituent ID* 68033) the artists were born before the separation of Korea into North- and South Korea.

Remarkably there are discrepancies between the values in the *Artist Bio* and the *Nationality*. The discrepancies can be read as attempt to add additional information to the limited possibilities of the *Nationality* field. For instance, if we examine the

[180] *Constituent IDs*: 74896, 48691 and 74897. All records of the “Coptic” and the “Persian” artists are both linked to artworks that were acquired from the Lillie P. Bliss Collection in 1934.

[181] *Constituent ID* 2852: Iyola; *Constituent ID* 4505: Parr, *Constituent ID* 5964: Akesuk Tudlik.

[182] *Constituent ID* 48421: Luis Claramung.

[183] *Constituent ID* 9460: Unidentified Czechoslovakian Manufacturer, *Constituent ID* 48691: Unidentified designer, *Constituent ID* 49100: FZ, *Constituent ID* 49101: M. Palasek, *Constituent ID* 49102: Tvar Schlosser.

[184] *Constituent ID* 40462: John Cale, *Constituent ID* 67735: Lucy Jones, *Constituent ID* 133249: Philip Jones Griffiths.

[185] *Constituent ID* 6212: Wagula.

records with “Native American” as *Nationality*: we can see that the museum stuff attempted to add additional information of the tribe or community the persons belong to in the *Artist Bio* field. For example, Cara Romera’s *Artist Bio* reads “Native American (Chemehuevi), born 1977” [186]. For Dyani White Hawk the *Artist Bio* reads “Sičáŋǧu Lakota and American, born 1976” [187]. For James Luna, the only one of the ten “Native Americans” with Ulan and *Wiki QID*, the *Artist Bio* reads “Payómkawichum/Luiseño, 1950–2018” [188]. These examples show two things: First, as assumed, the *Artist Bio* is unstructured and allows the museum stuff to add text without any restrictions. Secondly, the field is used for the attempt to diversify the information they provide about their artists, but they do not seem to find a unified way that fits to all of them. There is awareness of the limitation of the *Nationality* field and how it might not do justice to the individuals that are to be represented.

The three persons with “Canadian Inuit” as *Nationality* also show discrepancies in their *Artist Bio* data. Akesuk Tudlik’s reads only “Canadian, 1890-1966” [189] while the other two read: “Canadian Inuit” plus their birth and death dates. “Parr” is the only of the three with a *Wiki QID.*[190] In Wikidata he is listed as “Canadian” without a reference to his Inuit descendent. His record is an exemplar to show that the list of *Ulan* links is not complete in the MoMA datasets. For his person there would also be an *Ulan* entry available.[191]

Examining the “Scottish” records shows again that there are no rules embedded in how the *Artist Bio* field should be filled. Out of the 20 records 16 read “Scottish” and the birth year of the person, the other four read “British, born 1961”.[192] The three “Welsh” records all also hold “Welsh” in the *Artist Bio* field.[193]

[186] *Constituent ID* 132372.

[187] *Constituent ID* 132278.

[188] *Constituent ID* 34698.

[189] *Constituent ID* 5964. And *Constituent ID* 2852, *Constituent ID* 4505.

[190] https://www.wikidata.org/wiki/Q7139716 (accessed on 08.07.2023 12:16)

[191] http://vocab.getty.edu/page/ulan/500127305 (accessed on 02.07.2023 10:16). To make the point with another example, also Elsa Stansfield does not hold an *Ulan* refence although a record would be listed in Getty’s databases: see *Constituent ID* 35564. See http://vocab.getty.edu/page/ulan/500350533 (accessed on 02.07.2023 10:23)

[192] See for example Sally Osborn, *Constituent ID* 28746.

[193] See for example John Cale, *Constituent ID* 40462.

Besides the view examples of Korea, Yugoslavia, and Czechoslovakia, no historical names of states can be found in the data set. Historical states that would appear, if they were recorded would be for example the DDR[194], the Soviet Union[195] or Austria-Hungary[196], because artists represented in the collection were born or lived while those state forms were still in existence. But since those terms don't appear, we can state that MoMA is (besides a minimal number of exceptions) assigning the nationality according to contemporary boarders and states.

In order to geo-reference the *Nationality* of the artists, the following decisions were made: "Canadian Inuit" was assigned to the Nation "Canada". "Catalan" was assigned to the Nation "Spain". "Korean" was assigned to the Nation "South Korea", although some of the artists were born before the separation. "Native American" was assigned to "United States of America". "Scottish" and "Welsh" was assigned to the "United Kingdom". The records marked as "Czechoslovakian", "Persian", Coptic" and "Yugoslav" were discarded because no additional information on the represented persons was found and therefore the assignment to specific nations not possible.
I am aware that the assignment of the records that hold data like "Native American" to nation states like the US is problematic. But not assigning them would mean that their records do not appear in the visualization and the following analysis at all. This is another example for the tradeoff between standardization and individual information and it shows that the available structures of recording and displaying data that are so deeply entrenched in how we think, actually create biases and misrepresentations. The tribe structures of Native Americans are a perfect example for that, their way of grouping is not represented in the current structure of databases. Additionally, each Nation was grouped based on the seven-continent-system to continents.

The category "Native American" also spurred the search for records that represent Native American's but aren't categorized as such. To find examples, a targeted search for specific names of ethnic groups and tribes within the *Artist Bio* field was

[194] See for example Carlfriedrich Claus, born 1930, *Constituent ID* 1137, is categorized as „German".

[195] See for example Varvara Rodchenko, born 1925, *Constituent ID* 8476, categorized as "Russian".

[196] See for example Erika Giovanni Klien, born 1900, *Constituent ID* 3146, is categorized as Austrian.

conducted, but besides that the search for records like this is based on chance, since there is no mark or highlighter that would allow to query the records easily. The first two example records have the *Nationality* "Native American "assigned to them. Hock E Aye Vi Edgar Heap of Birds' Bio reads "Cheyenne and Arapaho Nations, born 1954",[197] Ishmael (Angaluuk) Hope's *Artist Bio* reads "Tlingit and Iñupiaq, born 1981".[198] The third record belongs to Sheroanawe Hakihiiwe, who has "Venezuelan" as *Nationality* but his *Artist Bio* reads: "Venezuelan and Yanmami, born 1971".[199] Three examples can be used to illustrate exemplarily that it does occur in the MoMA dataset that artists are assigned to common *Nationalities* although that might not do justice to the reality of the artists. This also shows again the attempt of the museum catalogers to record that important information, when not in the *Nationality* feature at least in the free text field *Artist Bio.*

After the terms were assigned visualizing the results was possible. The first one in Figure 4 shows the top 15 nationalities of the artists represented in the *Artists dataset* as assigned by MoMA.

[197] *Constituent ID* 35836.

[198] *Constituent ID* 134447.

[199] *Constituent ID* 134447.

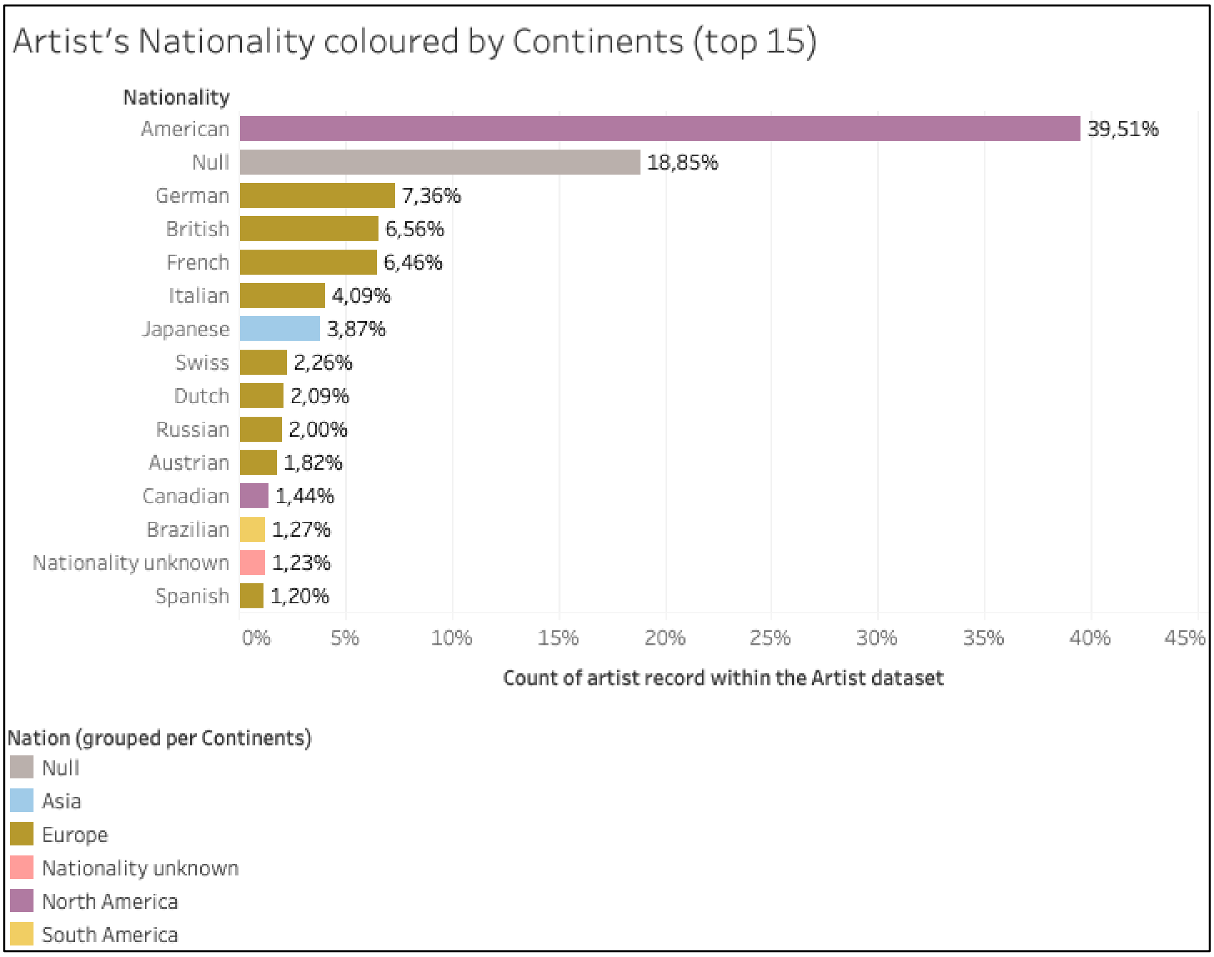

Figure 4 Artist's Nationality colored by Continents (top 15). The top 15 Nationalities that are assigned to the artists in the Artists dataset (y-axis). The x-axis shows the percentage of the records in regard to the complete dataset. Sorted from most to least frequent. Color encoded are the continents the Nationalities were assigned to. Some terms had to be reassigned because they did not represent "Nations". For example, "Native American" or "Canadian Inuit" were added to the Nationalities that are considered to be closest (America and Canada). 4 values could not be assigned and are therefore missing: Coptic, Czechoslovakian, Persian, and Yugoslav. The image was exported from Tableau.

The largest group of artists is marked as "American" followed by almost 19% of records without any data on *Nationality*. This should not be mixed up with the records where "*Nationality* unknown" was recorded. This entry shows that the attempt to select a *Nationality* was there, but due to unknown reasons was not unknown to the museum. The color encoding shows that under the 15 top values 9 represent European countries. Only North America and Japan are representants off other continents that make it in the ranking.

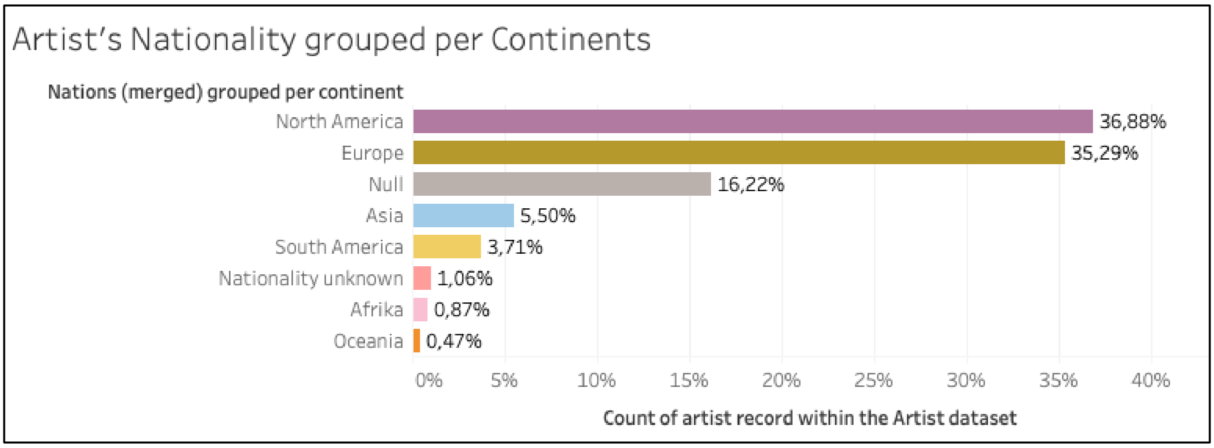

Figure 5 Artist's Nationality grouped per Continents. The continents that are represented through the Nationality feature of the Artists dataset on the y-axis. The x-axis shows the percentage of the records within the complete dataset. Sorted from most to least frequent. 4 values could not be assigned and are therefore missing: Coptic, Czechoslovakian, Persian, and Yugoslav. The image was exported from Tableau.

When the Nations are grouped by continents as in Figure 5, we can see that 72% of all artists in the MoMA dataset are either from North America or from Europe. The third largest group of artists records does not have a nationality assigned to them, 16%, plotted in gray. The observation was made that most of records that represent companies, studios, bands or schools and universities fall into those groups.[200] Given the nature of these entities, it seems reasonable to not assign values for *Nationality*. Since nationality is a concept that is commonly based on individual citizenship or identification, it may not apply for those records. This observation serves as an example that the predefined structure of how to enter records, and in this case what to enter, assumes uniformity and homogeneity.

There is a remaining part of records (around 10%) (when we exclude the records we grouped as institutions, groups, and companies) that would have data in the *Artist Bio* field, a handful of records even holds links to *Ulan* and or *Wiki QID*, those records could be enriched manually, and a nationality could be assigned to them.

[200] Result of a wildcard filter on the *Artist Bio* field for all records that contain "est.", as it was noticed in the data understanding part that this is repeated way of how MoMA catalogues those kinds of records. 186 of the 221 records that read "est." in the *Artist Bio* field, don't have a *Nationality* assigned to them. See for example: Yamaha Corporation Hamamatsu Japan, *Constituent ID* 9633; yo2 Architects Ltd., *Constituent ID* 37375; The Beatles, *Constituent ID* 34671 or the New York University, *Constituent ID* 39496)

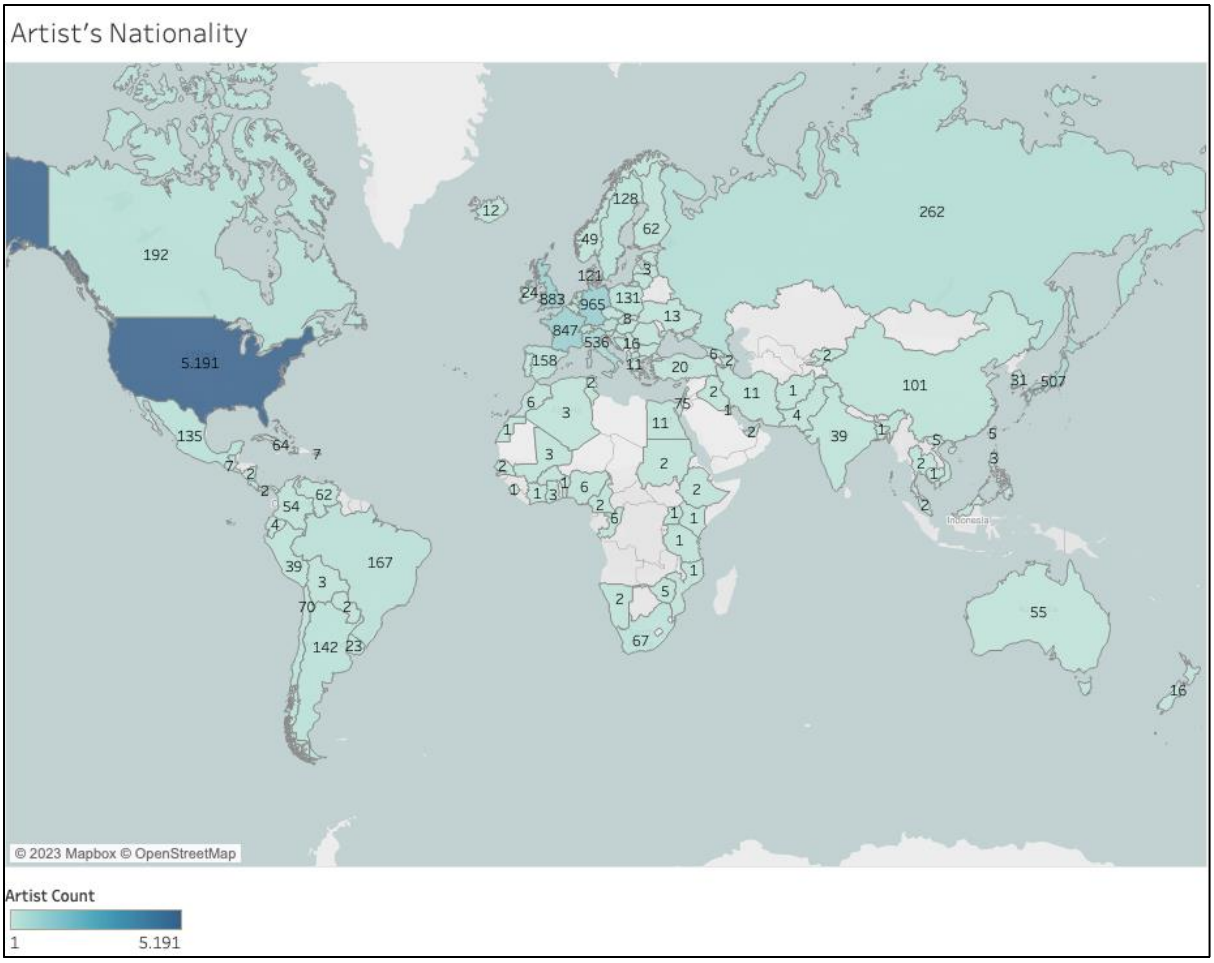

Figure 6 Artist's Nationality on a World Map. Map visualization depicting the countries the artists are referenced to (based on Nationality feature in Dataset). The specific count of records is added as label to the country (were readable). The amount is additionally encoded in the color, ranging from light to dark blue. 6 values could not be assigned and are therefore missing: Coptic, Czechoslovakian, Persian, Yugoslav, Nationality unknown und records without data for the Nationality feature. The image was exported from Tableau.

The superfluity of artists identifying as American is clearly visible in the map visualization of Figure 6. Slightly noticeable is the darker shade of some European countries, again repeating the already made observation that American and European artists are best represented within MoMA's collection. The map also shows absences, the states, and regions that are not represented in the MoMA dataset. Most noticeably African countries, the Middle East and Central- and Southeast Asia.

As seen **Error! Reference source not found.**, the six available *Gender* values had to be unified in spelling in order to represent the three categories used by MoMA.

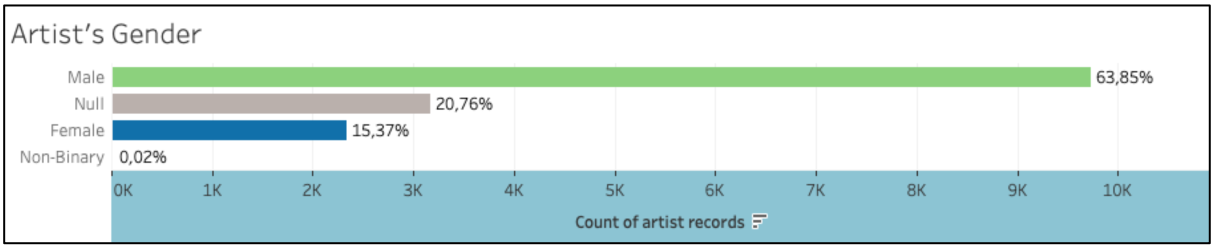

Figure 7 Artist's Gender distribution. Each of the three categories and the records without gender category on the y-axis and their occurrence count within the Artists dataset on the x-axis. The image was exported from Tableau.

Figure 7 shows a bar plot of the three categories and their percentage of appearance within the complete MoMA *Artists dataset*. Only 0,02% list Non-Binary as *Gender* category. 15% of the artists in the dataset are categorized as female. 20% are without gender category, and the vast majority of artists represented in the collection are male, 64%. This shows a huge imbalance of gender representation within the MoMA collection. If the records without *Gender* values are excluded, the percentage of male artists rises to 81%, compared to 19 % of female artists (non-binary stays at 0.02%). Combining the *Nationality* and the *Gender* feature shows that this unbalance is similar for all continents. As depicted in Figure 8, male artists build, for all represented continents, the majority between 69 and 76 percent (here the Null values are included). Africa is the continent where the female artists form percentual the largest group compared to the other continents, 23%. It is important to keep in mind, that the representation of the specific continents differs enormously within the dataset (compare with Figure 5). The 23% of female African artists are formed by 31 records, the 21% of female North American artists by 1193 records.

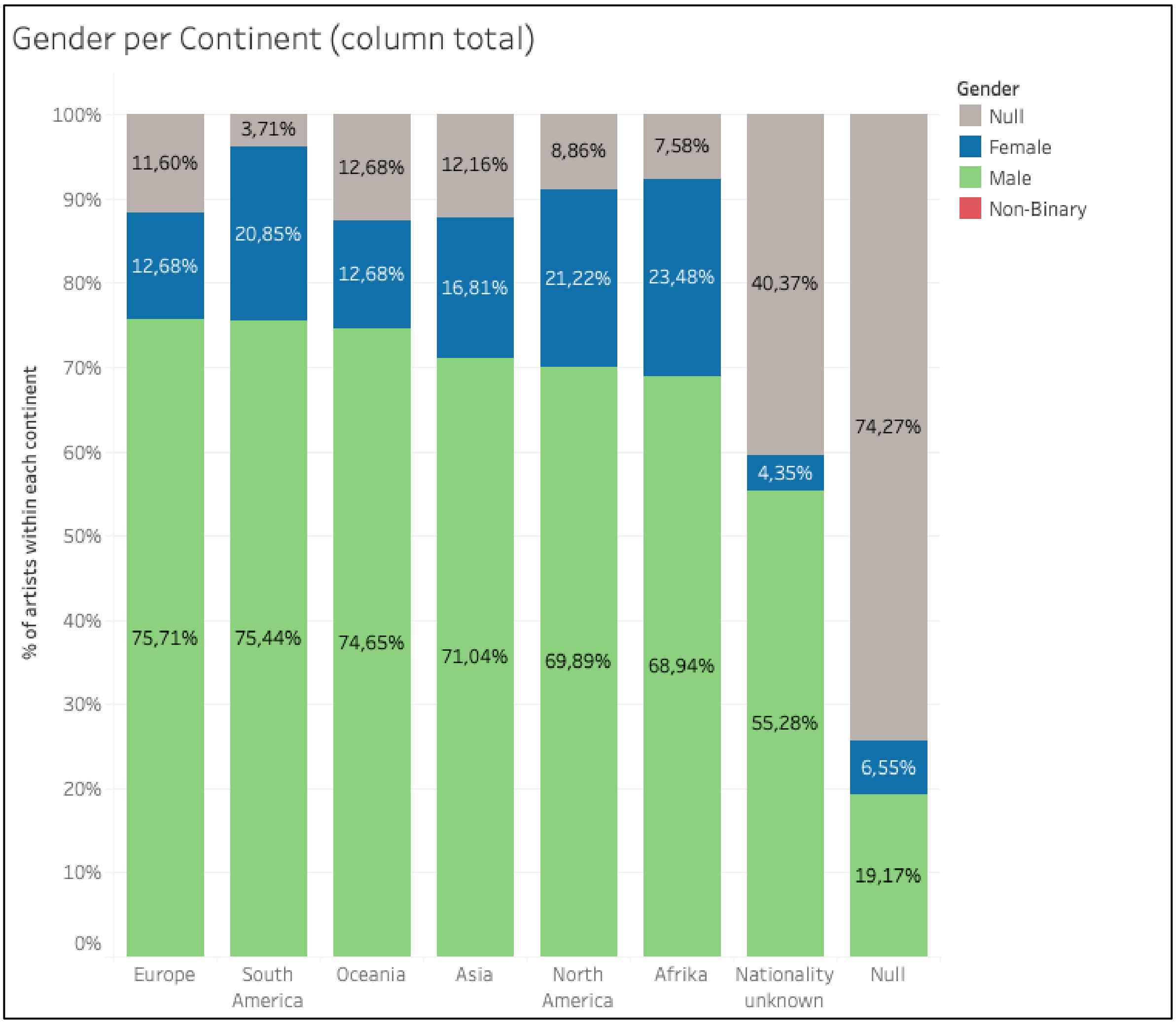

Figure 8 Gender categories per Continent. The % of records in the Artists dataset within each continent on the y-axis. The continents, the “Nationality unknown” and the records without data for the Nationality field (on which the assignment to continents is based on), on the x-axis. Gender category is encoded in the color, as explained in the color legend. The image was exported from Tableau.

If we compare the gender distribution of artists based of the top 10 nations (see Figure 9 and additionally Figure 4 for the ranking of the nations), an interesting observation can be made: The imbalance in gender representation is more pronounced in foreign countries compared to the two North American countries in the selection, namely the United States of America and Canada. Italy and France have the lowest representation of female artists, with only 8% and 8.5% respectively.

Compared to 22% and 21% percent of female artists of USA and Canada.

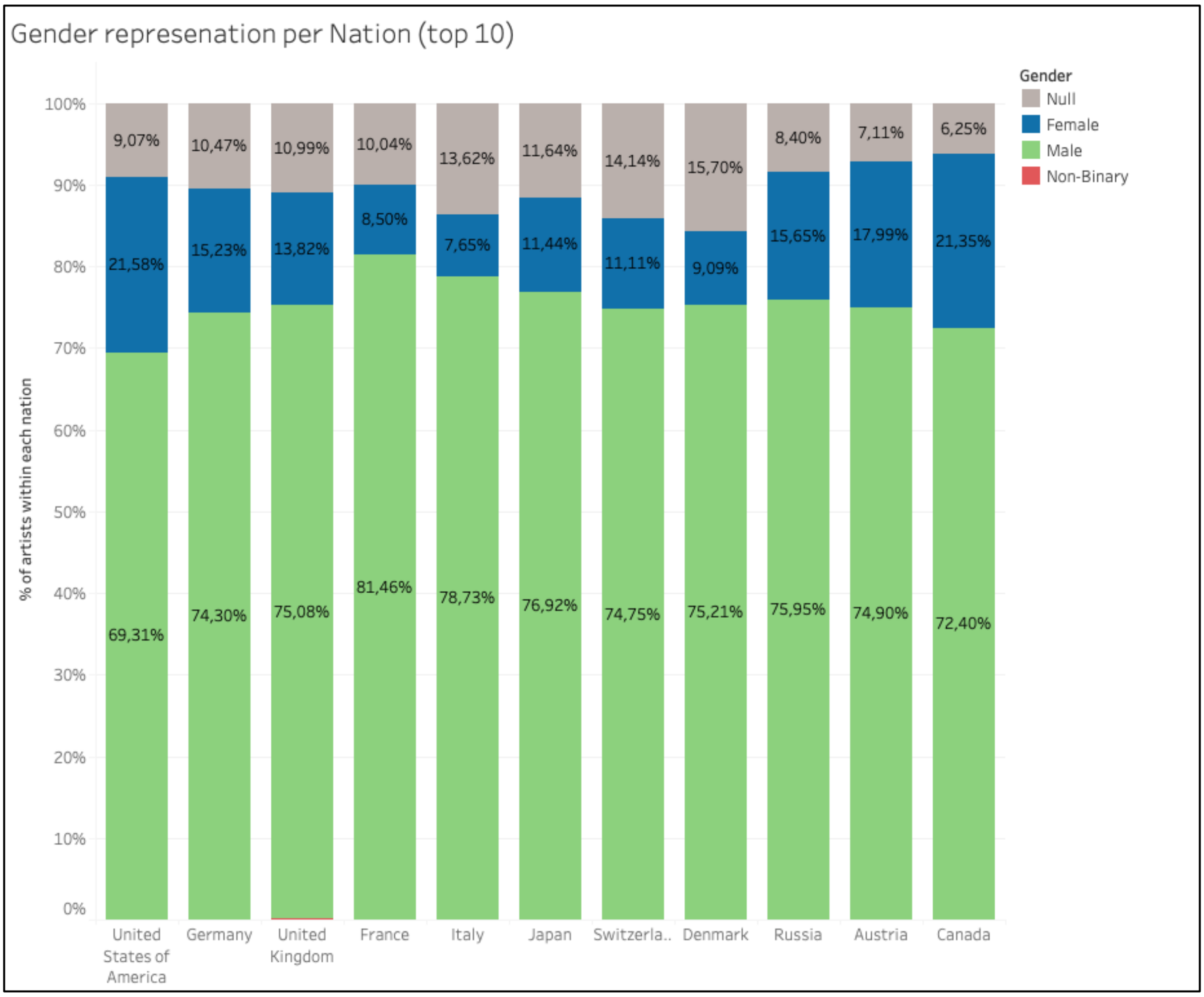

Figure 9 Gender categories per Nation (top 10). The % of records in the Artists dataset within each nation (based on Nationality feature) on the y-axis. The group of records without data for Nationality are excluded from the list (would be number 3 in ranking – see Figure 4) The other top 10 most often appearing Nations are listed on the x-axis, sorted by occurrence count. The gender category is encoded in the color, the legend on the side explains which color represents which gender category. The image was exported from Tableau.

Until now we did not consider the fact that artists can be represented with multiple objects within the collection of MoMA, in the following section the nationality and gender representation of the artists based on the objects they are linked to are examined. As the first step of analysis, it was confirmed that every available *Constituent ID* from the *Artists dataset* actually appears within the *Artworks dataset*.[201] It is important to repeat that we are missing data for those artists that are linked to objects with more than one creator. This will not affect the results

[201] This was done with a so called „Calculated field“ in Tableau and a simple calculation that checked if the ID in the Artworks table also appears as ID in the Artists table. The calculation returned Boolean values, in our case "TRUE" was returned for all cases, showing that all IDs are present (CONTAINS([*Constituent ID* (Artworks.csv)], [Constituent ID]) = TRUE).

majorly, but could be the reason for slight shifts and some missing data points. This needs to be in kept in mind.

Using the *Nationality* feature of the artists records as the source for a Map visualization and the record count of objects from the *Artworks dataset* for the color encoding, we see a very similar image as in Figure 6: darkest (the most objects) is the United States of America, some countries in Europe show slightly darker shade than the majority of all other countries that appear in the dataset. More insightful is a closer look at one continent in particular and its nations object count. Important to note is, that there is no creation location of objects recorded in the *Artworks dataset* of MoMA, using the *Nationality* of the linked artists does not necessarily give insights on the context of the artwork or where it was created.

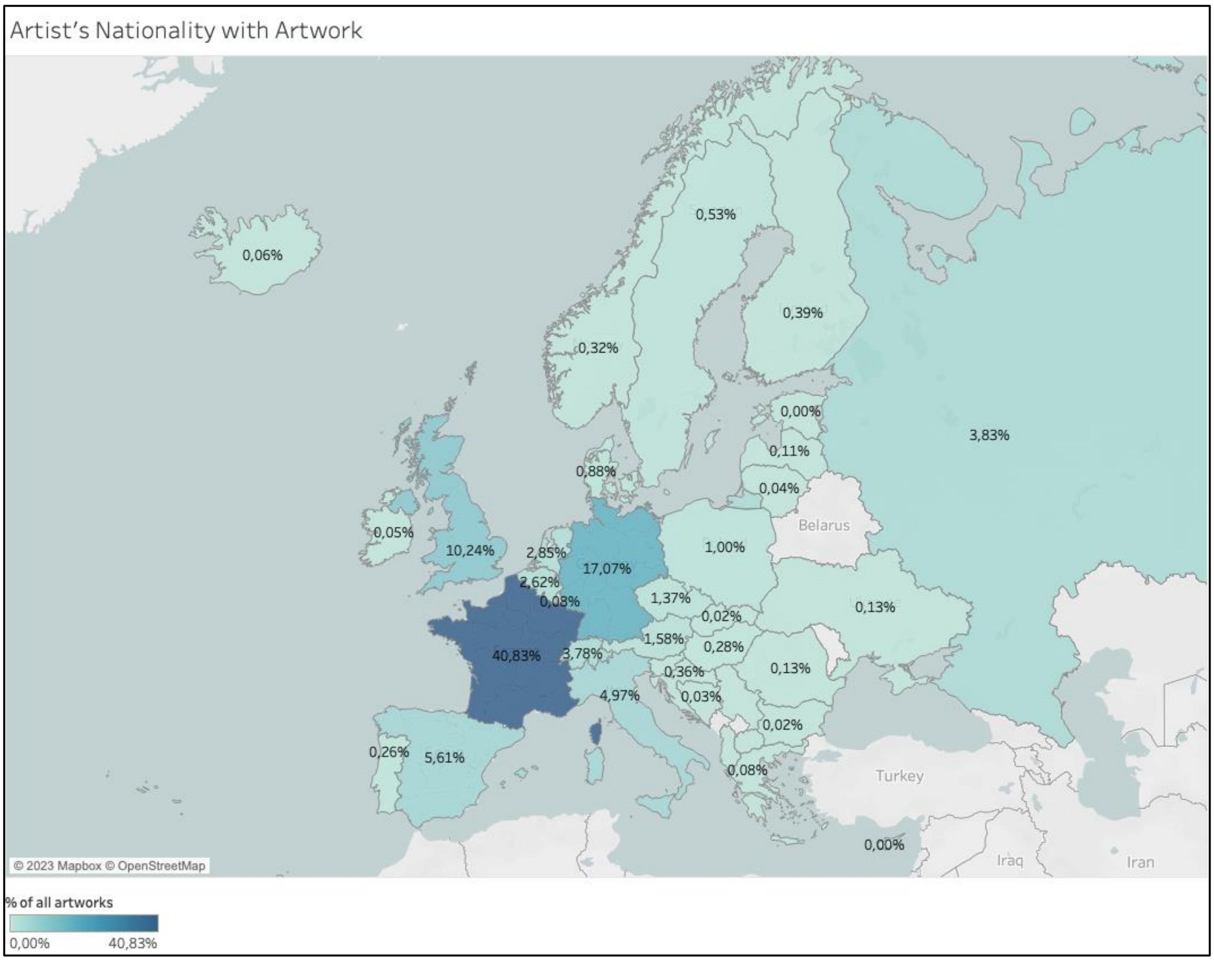

Figure 10 Artist's Nationality on a World Map, detail of Europe. Map visualization depicting the countries the artists are referenced to (based on Nationality feature in Artist's dataset). The specific count of object records that are linked to the artists is added as label to the country (were readable). The amount is additionally encoded in the color, ranging from light to dark blue. 2 values that might be added to one of the countries of Europe are not assigned: "Czechoslovakian" and "Yugoslav", also records without data for the Nationality feature of the linked objects are not represented within the plot. The image was exported from Tableau.

As clearly visible in Figure 10, France is the country with a significant lead over the most artworks within the MoMA dataset are related to through the *Nationality* of the artists. 40% (almost 23 thousand object records) of the total collection were created by French artists, an enormously large amount considering that MoMA is an American modern art Museum. Germany, the United Kingdom, and Spain are the three other European countries that represent each more than 5% of the collection. It would be an interesting inspection to calculate the percentage in regard to the population of the countries. The ranking would probably shift towards favoring smaller countries, like Switzerland, Denmark, and Austria that are already represented with considerable high numbers even though their small territorial size.

As we have seen in Figure 5, most of the artists have a North American nationality assigned to them (37% of total), followed by European artists (35%) and records without *Nationality* (16%). When we consider the frequency of artists nationality based on the count of artwork records, we can see that the percentage of North American and European representation is larger (43% and 40% of all object records), see Figure 11. We can also observe that respectively South American artists and Asian artists are less present compared to the percentage of artist records. Additionally, we can see that South American artists overtake Asian artists in the ranking. The takeaway point here is that those North American and European artists that build the largest group within the represented artists are also present with per average more artworks than the artists from other continents.

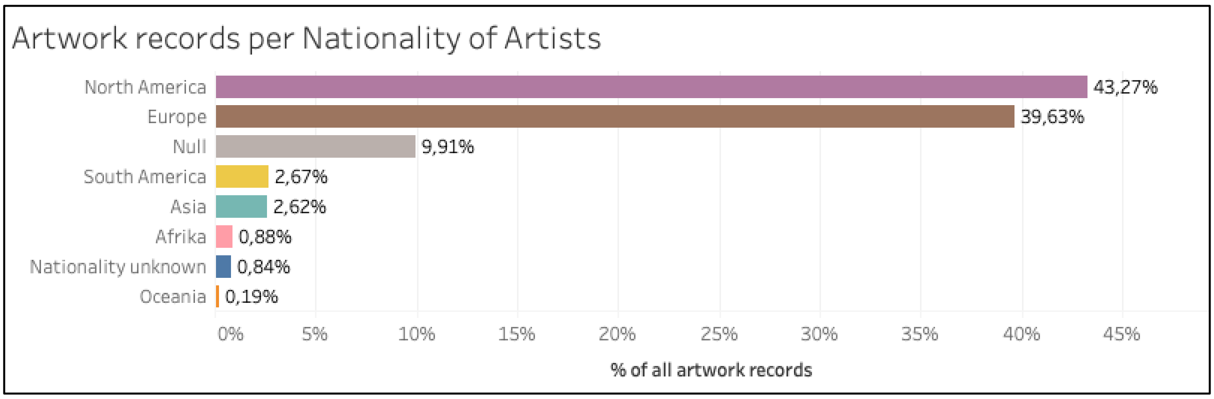

Figure 11 Artist's Nationality grouped per continents for each artwork record. The continents that are represented through the Nationality feature of the Artists dataset on the y-axis. The x-axis shows the percentage of the occurrence within the complete Artworks dataset. Sorted from most to least frequent. The image was exported from Tableau.

Figure 12 shows that the gender representation based on the object count is even more favoring male artists then when the distribution is made based on the artist record counts. 75% of all objects in MoMA's collection were made by men, only 13% by female artists, and 0.01% percent by non-binary artists (compare with Figure 7). This means that there are not only significantly more male artists present in the collection, but that each of them is also represented with a higher quantity of artworks. In average each male artist is represented with 10 artworks, female artists on the other hand only with 8 artworks.

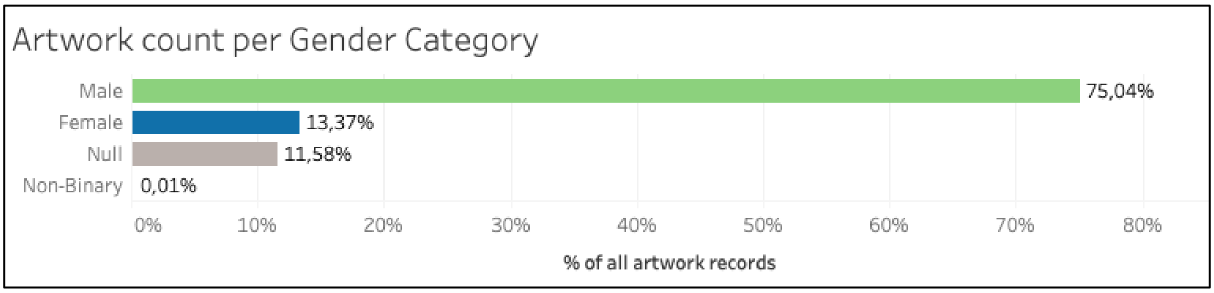

Figure 12 Artwork count per Gender Category. The three available gender categories with the group of objects who's artists were not assigned one of them on the y-axis. The percentage of the appearance within the entire Artworks dataset of MoMA on the x-axis. The image was exported from Tableau.

This section has shown that the gender distribution of the MoMA collection favors male artists over female artists, in both the artists represented but also when looking at the artworks and their creators. Non-Binary artists are represented in so little that they are almost not distinguishable in the plots. The inspection of the nationality feature showed that most of the artists are North American, followed by European artists and their artwork. This shows the focus of the Museum of Western Art and absences of other regions.

Before the focus of the analysis will shift to the objects record only a brief look will be taken on the lifetime data available in the *Artists dataset.* The following bar plot in Figure 13 shows the average age of the artists based on their gender category.

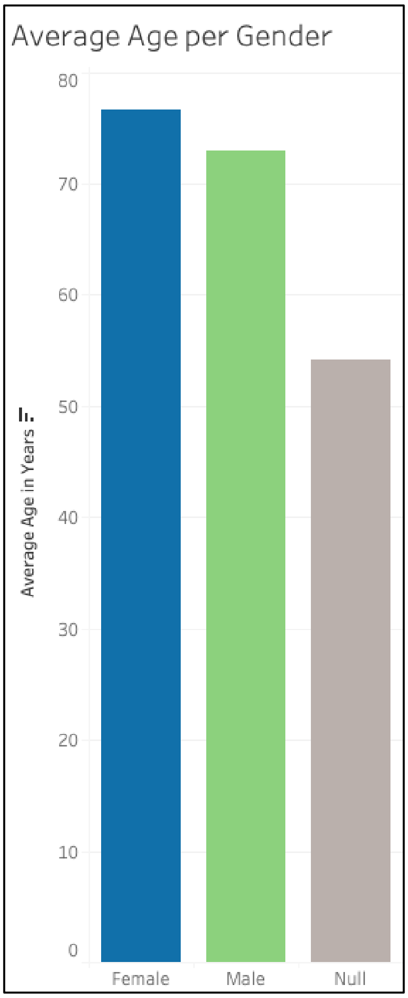

Figure 13 Average Age per Gender. Bar plot shows the average age in years of each artist record in MoMA's dataset. Records that hold either only one or no date (Begin Date or End Date) are excluded. The bars are sorted descending by average age. The plot shows that the female artists are of older age than the male artists. The average age for the records without any gender category is significantly lower. The image was exported from Tableau.

The records that hold either only one or none of the dates (*Begin Date* and *End Date*) are excluded from the graph, which is also why no Non-Binary artists are not represented. Female artists represented in the MoMA collections were by average older than the male artists.

The same insight can also be gained when we look at the same data in a scatter plot, see Figure 14.

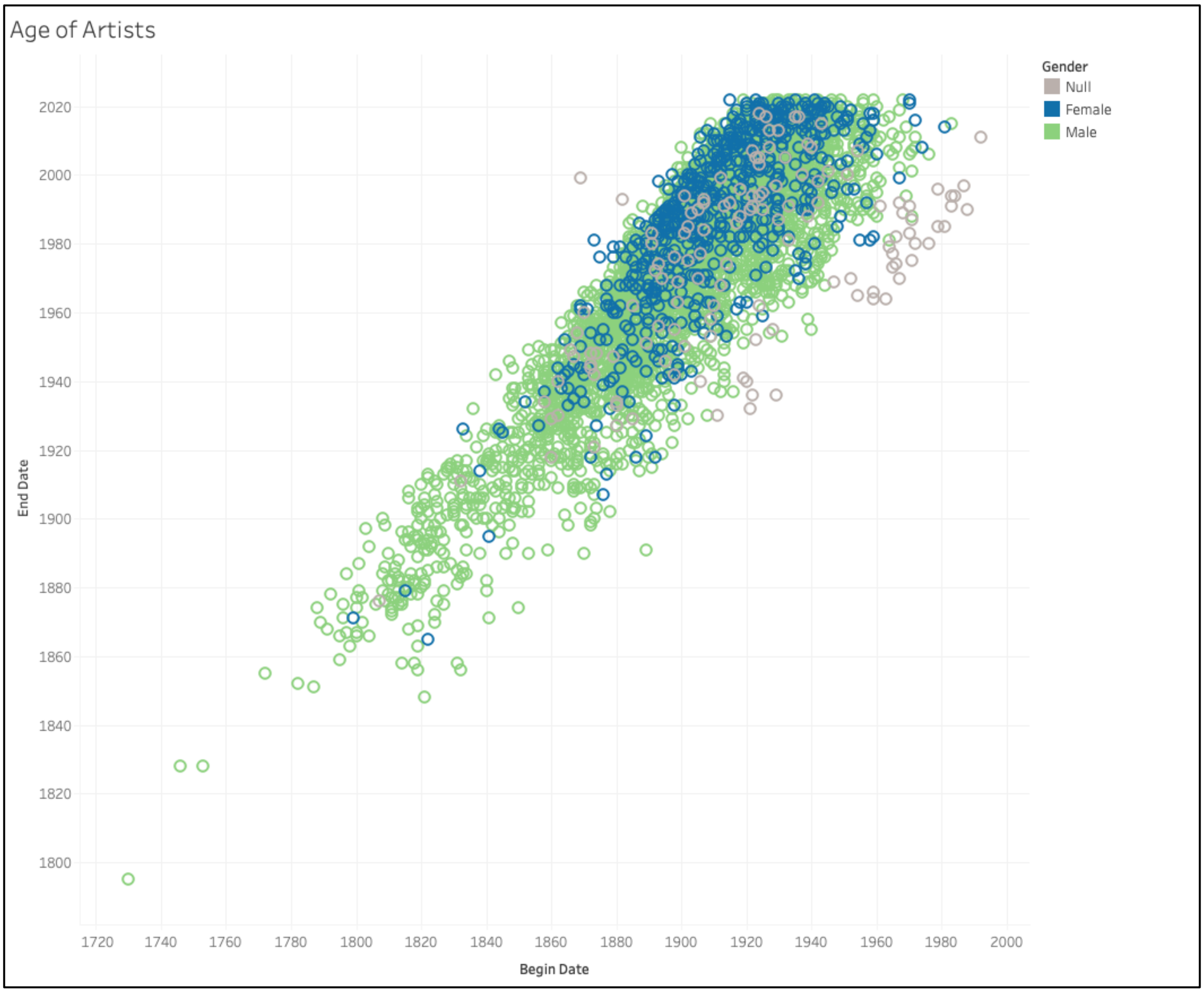

Figure 14 Age of Artists displayed in a scatter plot. The Begin Date is set in correlation with the End Date. Color shows details about the gender categories. Records that hold either only one or no date (Begin Date or End Date) are excluded. The image was exported from Tableau.

In this plot, the *Begin Date* is regarded as the birth date, and the *End Date* is assumed to be the death year are set in correlation to each other. As expected, there is a linear correlation between the two dates. What we can also see here, what is not visualized in the bar plot, is the greater number of male artist records (green) in comparison to the female artist records (blue). The gender category, that is encoded in the color of the circles, allows for additional insights. We can see that the male and female artists, regardless of their unequal count, are not equally distributed. While the male artist records span from the mid 19th century until today, the female artists, besides some individual examples, become recognizable in considerable number only as late as the middle of the last century. What is more interesting still, is the distribution of the records without any gender representation (gray). It is important to keep in mind is that those record hold begin and *End Date*s (all others were excluded), a fact which speaks for a quite high completeness of the record. The gray records are spread broader then the female or male datapoints, reaching farer away from the middle line. The (imagined and not plotted) middle line is the average age, everything above is older, everything beneath is younger. In Figure 15, the records with a lifetime above 100 years (squares) and under 20 years (crosses) are displayed. While the 37 records on the top range are representing almost exclusively people that actually got very old, the bottom group consists mostly of records for artists groups, organizations, or collectives[202]. There are records left that have no gender attribute, even though they represent people. Why this information is missing is not stated. If a tag, marker, or feature would be available that somehow categorizes the artists records as representing an induvial artist, an artist group or a company, analysis of the dataset could be made more focused and without to many distractions.

[202] Three examples of very old artists: Grete Lihotzky, 103 years (Constituent ID 36721); Leni Matthaei, 108 years (Constituent ID 3844) or Carmen Herrera, 107 years (Constituent ID 30075). Three examples of young artist records that actually represent groups: Gruppo N, active for 5 years (Constituent ID 2377), “a.r.” group, active for 7 years (Constituent ID 48911); Nice Style The World’s First Pose Band, active for 4 years (Constituent ID 32944)

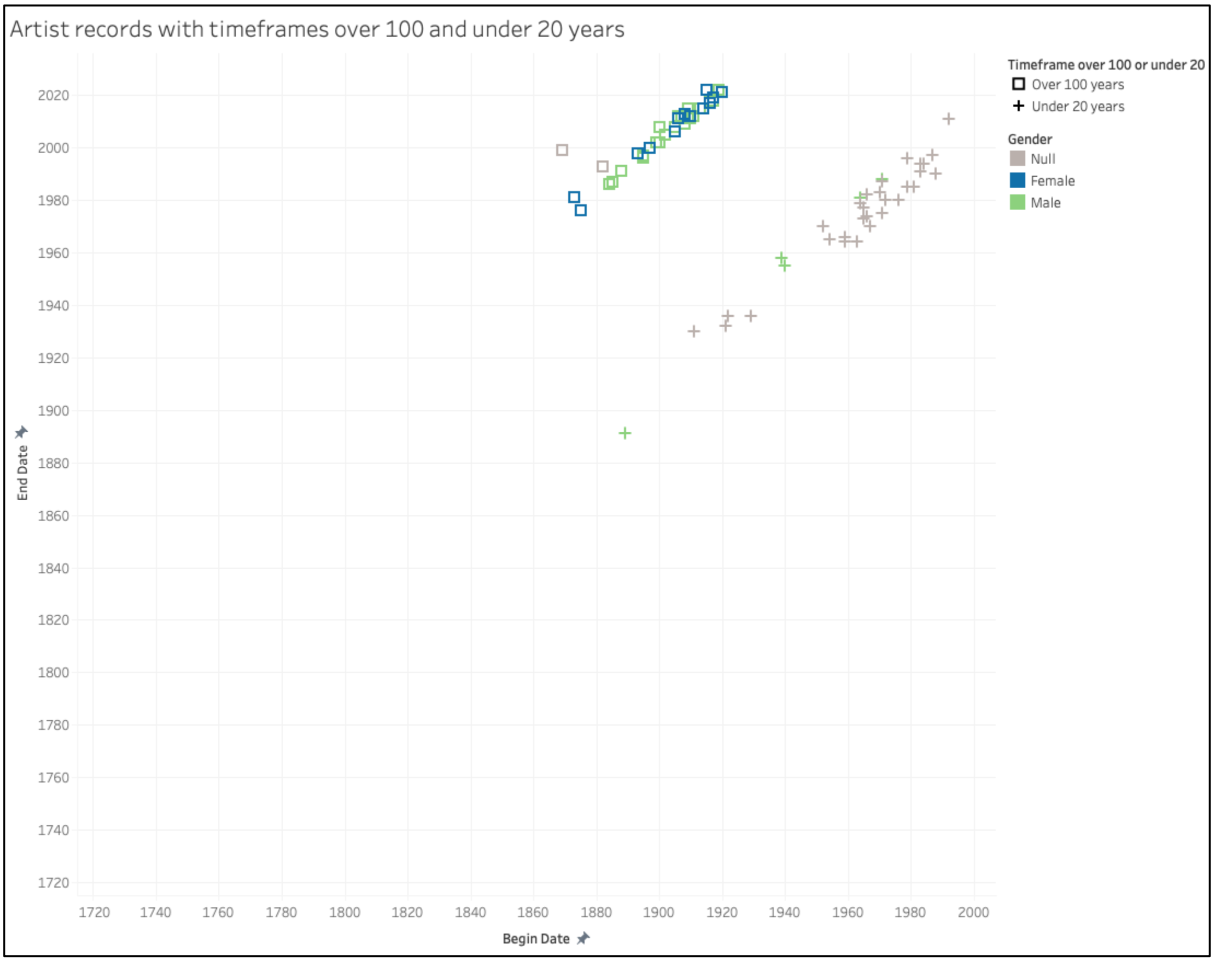

Figure 15 Artist records with timeframes over 100 and under 20 years (>100 and <20). The two categories are encoded in the shape of the datapoints. Records that hold either only one or no date (Begin Date or End Date) are excluded. Also excluded are all records with a timeframe that falls between the two extremes. Color shows details about gender. The image was exported from Tableau.

This short section showed that the average female artists become older than the average male artists in the MoMA collection. It was possible to show that the female artists appear later within the complete dataset, the early decades of modern art are represented almost exclusively by male artists. The need for a categorization of artists records was stated, to keep apart records that represent artists, groups, or companies.

The subsequent section will redirect the focus of analysis towards the artworks records and examine what insights we can retrieve about the objects in MoMA's collection.

Artwork types

The following part will briefly examine what kind of objects are part of MoMA's collections. As mentioned, above in chapter 3.1, the museum is structured by departments that have specific responsibilities based on object category. There are 8 departments listed in the *Artworks dataset*, only 6 of them are the so called "collecting departments".[203] The two departments that are not included into this selection of "collecting departments" are the "Fluxus Collection" and the "Architecture & Design – Image Archive".

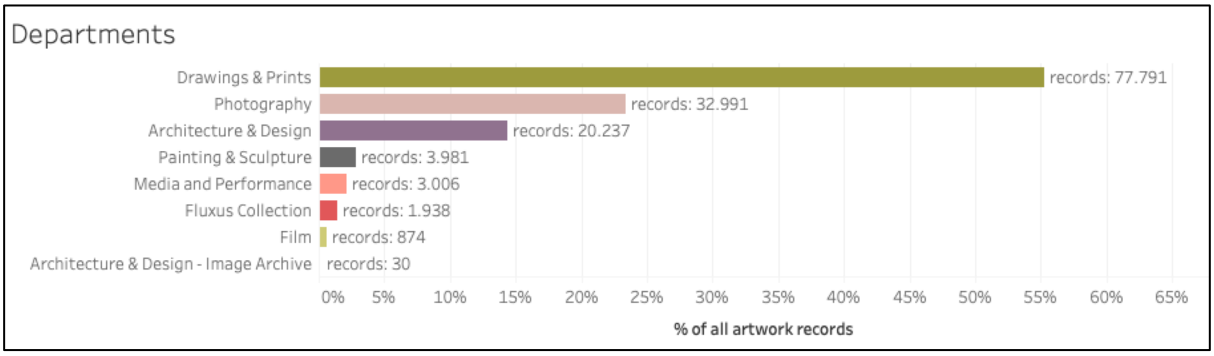

Figure 16 Departments with artwork record counts. Bar plot displays the Departments on the y-axis, sorted from largest to smallest, the percentage of the total count of records for each department on the x-axis. The bars are labeled with the distinct count of records. The image was exported from Tableau.

Figure 16 shows the *Department*s in the *Artworks dataset*, sorted by the count of records assigned to them from largest to smallest. The largest *Department* is "Drawings & Prints", making up 55% of the entire dataset, followed by "Photography", 23%, and "Architecture & Design", 14%. After a significant gap the *Departments* "Painting & Sculpture", 3%, and "Media and Performance", making up 2%, follow. The three smallest groups are the "Fluxus Collection", "Film" and the "Architecture & Design Image Archive". It was already stated that the film collection of MoMA cannot possible listed completely in the available dataset, its size would be of similar size as the "Photography" *Department* (see chapter 3.3.2 on what is not in the dataset above).

Next to the *Department* feature the dataset provides two more features that can be used to gain insights about the type of the artwork: *Classification* and *Medium*.

The dataset holds 36 *Classification* values. 17 of them appear in less than 0.1% of the records, only the 19 values that appear more frequent are shown in the following plot of Figure 17.[204] The bar plot visualizes the *Classification* values on the y-axis

[203] *Lowry*, Introduction, 18.

[204] The following Classification terms appear less then 0.1% times within the complete dataset (in descending order): Poster, media, textile, Performance, notebook, Collage, graphic Design, Photography Research/Reference, Film (object), Publication, Furniture, and Interiors, Digital, Software, Fashion, Document, Architectural model. One record has no classification assigned.

and their record count on the x-axis. Sorted from the most frequent value ("Photograph", appears more than 34 thousand times, which is 24% of the entire records) to the least frequent value ("Film", 159 records, 0.1%). The color encoding shows the already in Figure 16 introduced departments. Within this selection of 19 values the "Mies van der Rohe Archive" and the "Frank Lloyd Wright Archive" do not describe object types but rather organizational grouping of records within specific sub collections.

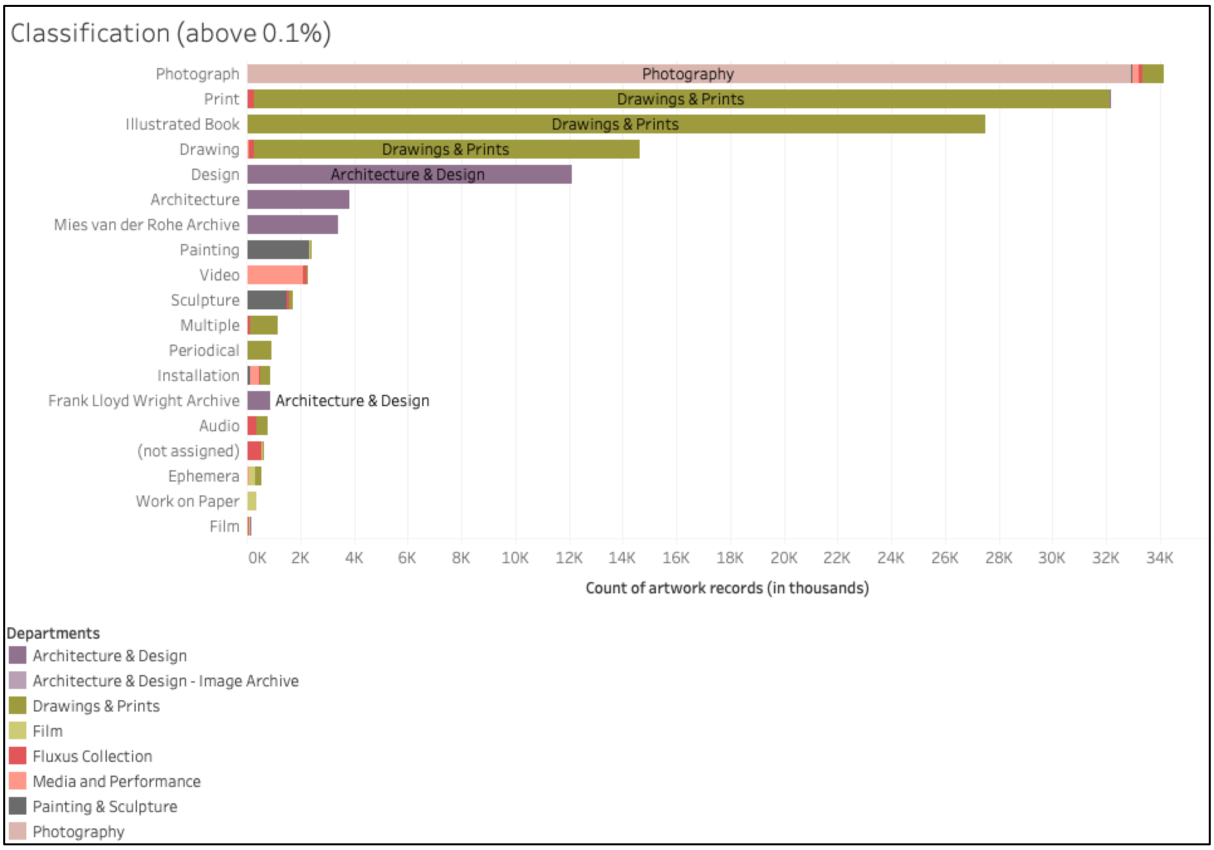

Figure 17 Record count for each Classification. Bar plot displays the Classification values on the y-axis, sorted most to least frequent. The values that appear less then 0.1 times are excluded from the list (18 from 36). The x-axis shows the count of artwork records in thousands. Additionally, the Departments are encoded in the color, according to the color legend on the right side. The image was exported from Tableau.

The color encoding makes visible that the *Classification* values can be assigned to multiple *Departments*. When we examine the first bar of the plot, the term "Photography", closer, we can see that the majority of the records are also assigned to the "Photography" *Department*, but others are assigned to the" Media and Performance", the "Fluxus Collection" and the "Drawing & Prints" *Department*. While this may be plausible content wise, because one could image photographs of performances or drawings or prints, there are other examples where the combination of C*lassification* and *Department* appear surprising. The *Classification* value "Sculpture" is assigned to the largest part to the *Department* "Painting & Sculpture", which seems logic. But the same term is also combined with the “Media and Performance”, the “Fluxus Collection” as well as the “Drawing & Prints” *Department*. As a consequent the plot also shows that each *Department* consists of different *Classification* values, “Drawing & Prints” for example has large numbers in the *Classification* values “Print”, “Illustrated Book” and “Drawing”.

The composition of *Classification* terms in relation to each *Department* can also be examined in detail, as has been done in Figure 18. The Plot shows us the *Classification* values that are used in combination with the *Department* value “Drawings & Prints”. The values are sorted based on the count of their appearance, from the most frequent term “Print” to the values that only appear once: “Film (object)” and “Fashion and Document”. As we have already seen in Figure 17, the most frequent combinations with the Drawings & Prints *Department* are “Print” (41%), “Illustrated Book” (35%) and “Drawing” (18%).

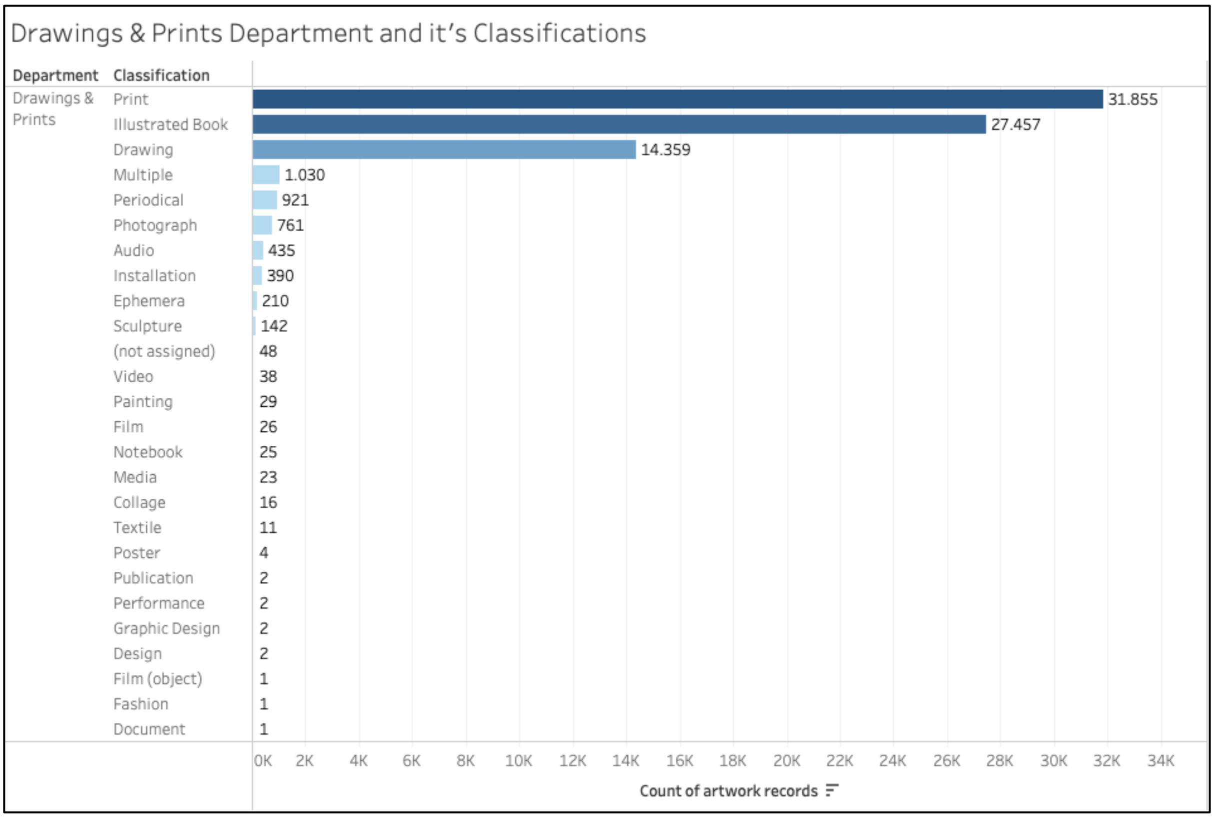

Figure 18 Drawings & Prints Department and it’s Classification values. The Classification values are listed on the y-axis, the specific count of artwork records that hold the value is shown on the x-axis in thousands. The values are sorted from the most frequent to the least frequent ones. The color encoding emphasizes the record count from bright to dark. The image was exported from Tableau.

The *Medium* feature is in contrast to the *Department* and *Classification* feature not structured. For the 140 thousand records we can count more than 20 thousand *Medium* values. Each of them appears in average 6.5 times. There are some rare terms that appear repeatedly, the maximum “Gelatin silver print” which appears 16638 times, while 16290 other values only appear once within the complete *Artworks dataset*. The bubble plot in Figure 19 gives an impression on the large amount of different values that are used to describe the medium of the artworks. It also shows the terms that appear repeatedly, next to already mentioned “Gelatin silver print” (11%), “Lithograph” (6%), and “Albumen silver print” (3%) are the most frequent ones. Although it should be noted that on second place in this order are actually the records without a *Medium* value, they yield 7%.

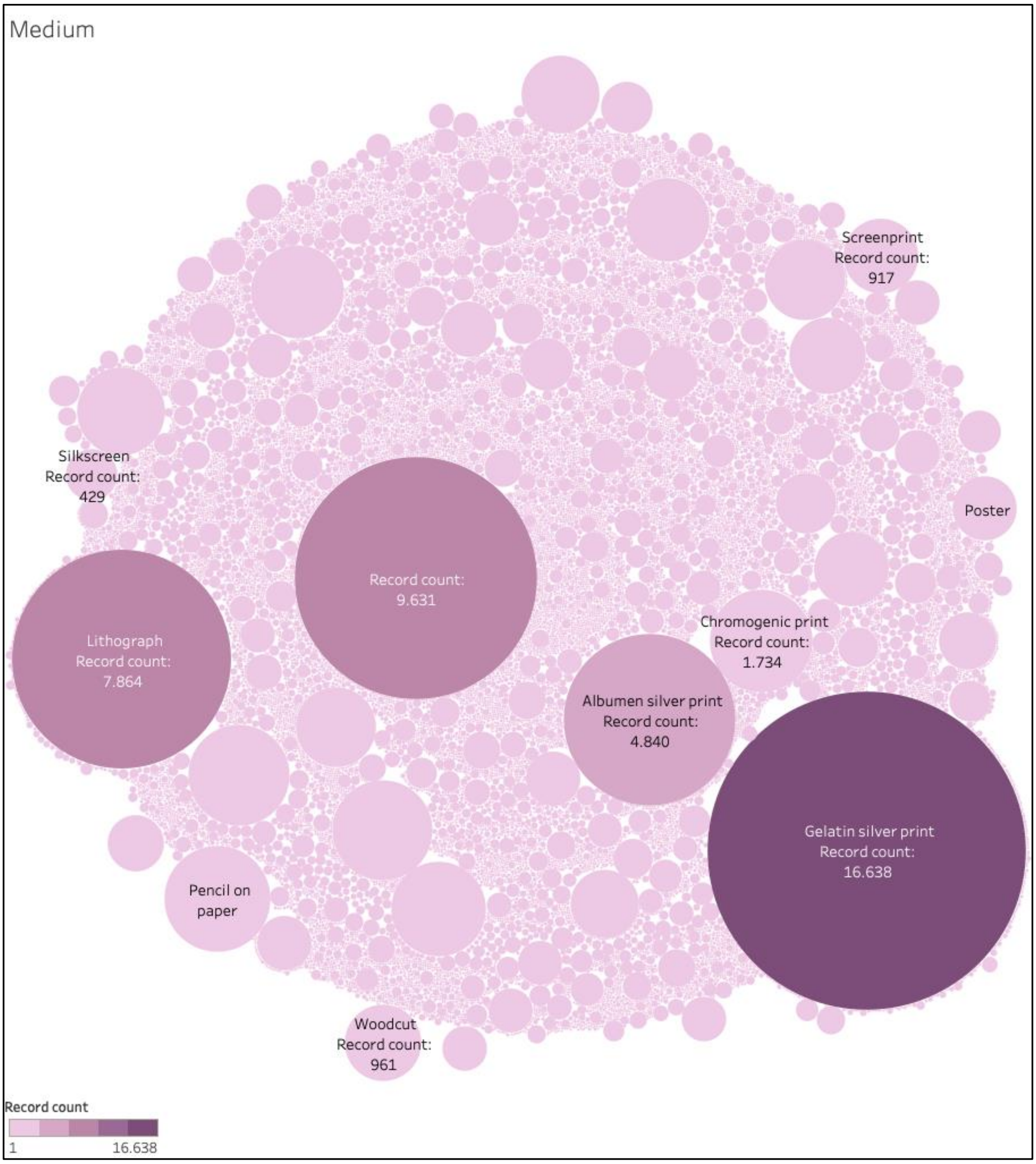

Figure 19 Medium values and their appearance in the dataset. The packed bubble plot shows the Medium values and their frequency within the MoMA dataset. The record count is encoded in the size and the color of the

bubbles. The color is stepped in 5 parts from 1 occurrence (lightest) to 16638 occurrences (darkest). The image was exported from Tableau.

Aside from the dimension fields that are included in the *Artworks dataset,* there is no additional feature that would allow us to analyze the types of objects. And since the *Medium* feature holds only unstructured data, further computational steps would be necessary that would exceed the scope of this thesis. With natural language processing, to name one method, it might be possible to gain further insights in the cataloguing and describing practice of the museums and the composition of its collections.

When we look at the *Department*s of the museum and combine it with the *Gender* feature of the linked artists, we can see a difference in the distribution, see Figure 20. The percentage of artworks by female artists varies from 2% in the records assigned to the "Film" *Department*, to 22% in the "Media and Performance" *Department*. This is also the only *Department* were artworks from non-binary artists are assigned to (the percentage is too small to appear in the plot). The percentage of male artists ranges from 29% in the "Film" *Department*, here the largest part is assigned to artists with no *Gender* attribute – which is an interesting observation that should be investigated further in future analyses, up to 85 % in the "Painting & Sculpture" *Department*.

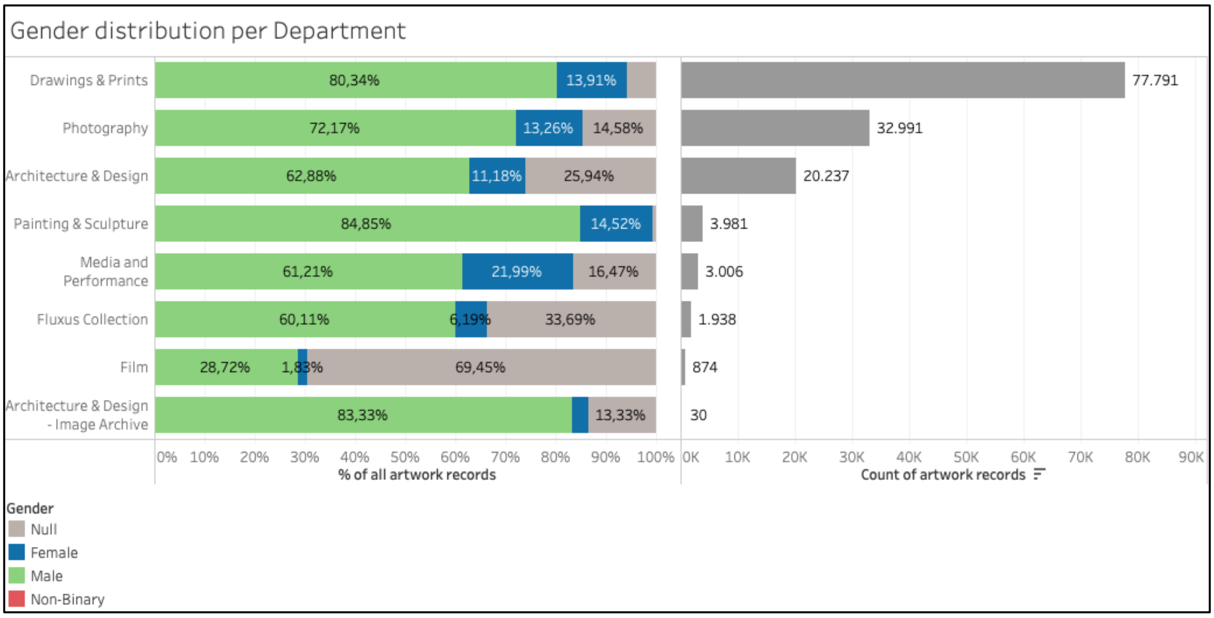

Figure 20 Gender distribution per Department. The Departments that are present in the MoMA dataset on the y-axis, sorted from the one with the highest artwork count to the one with the least. The first x-axis displays the Gender feature of the linked artists in percentage of the artworks within each Department. The second x-axis shows, for comparison the total count of artworks for each department. The color legend shows the Gender categories. The records linked to Non-Binary records are too small in number to be displayed, they are part of the Media and Performance Department. The image was exported from Tableau.

The examination of the features *Department, Classification* and *Medium* has provided the insight that while the first two are structured and their values limited, the third one is unstructured. It was possible to find common combinations of values of the *Department* and *Classification* features, even though not all of them appear logical from a distant view. The *Medium* feature is a free text field that holds very specific information regarding the single object, it is individualized and does not follow any rules. Only a couple of terms appear regularly, those are all descriptions of either photographic or printed artworks. The distribution of male and female artists within each department differs, but male artists are in all departments better represented, in some cases with enormous lead.

The following part of the analysis will focus on the acquisition history of MoMA and in particular of each of the directors. Insights on how the individuals and their decisions effected the distribution of departments and the representation of female and male and non-binary artists are expected. An analysis of the *Credit Line* feature will be expected to bear insights in the donation and acquisition history of the museum and each director.

MoMA's acquisition history

The following investigation on the acquisition history of MoMA will be based on the *Date Acquired* feature. At the beginning, we will take a look at the count of acquisitions over time followed by an analysis of the different directors and what and from whom they acquired artworks from during their time at the museum. At the end of this section, a closer look at the doners and sellers of artworks, based on the *Credit Line* feature will be provided.

Figure 21 shows the count of artworks acquired in each year from the foundation of MoMA in 1929 until (November) 2022, the year the datasets were published by the museum. During all the years, there was no time when no artwork was acquired. Additionally we can distinguish multiple peaks. The first one can be located at the year 1964, almost 13.000 artworks (9% of total collection) were acquired and recorded this year, the second largest peak is in the year 2008 where more than seven thousand artworks (5% of total collection) were required, closely followed by the peak in year 1968 with almost seven thousand acquisitions. The fourth highest peak appears in 2001 with little more than 4000 new acquisitions, followed by a steep decline down to 547 acquisitions in the year 2003.

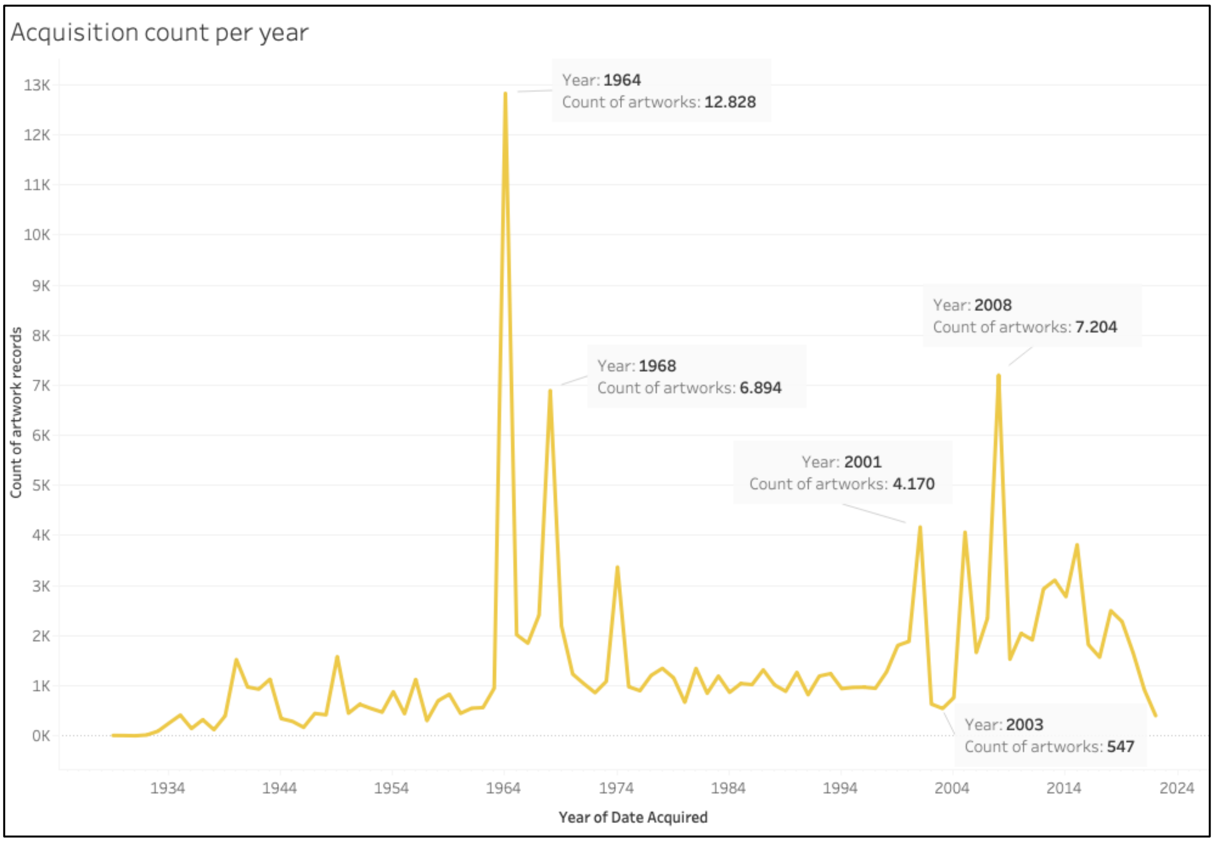

Figure 21 Acquisition count per year. Line plot that shows the count of artwork records on the y-axis and the years from the Date Acquired feature on the x-axis. Four peaks and one drop point are selected and annotated (displaying the specific year and the exact number of artworks that were acquired that year). The image was exported from Tableau.

The first peak is part of the directorship of Rene d'Harnoncour and will be examined more closely", see Figure 22. The records acquired in 1964 consist of 91% of illustrated books from various artists, as the assigned *Classification* feature tells us. 5% of the records acquired in 1964 are classified as "Photographs" followed by "Drawings. 96% of the acquired illustrated books were part of the "Louis E. Stern Collection". The peak in 1964 is therefore explainable by the acquisition of the Louis E. Stern Collection by Rene d'Harnoncourt.

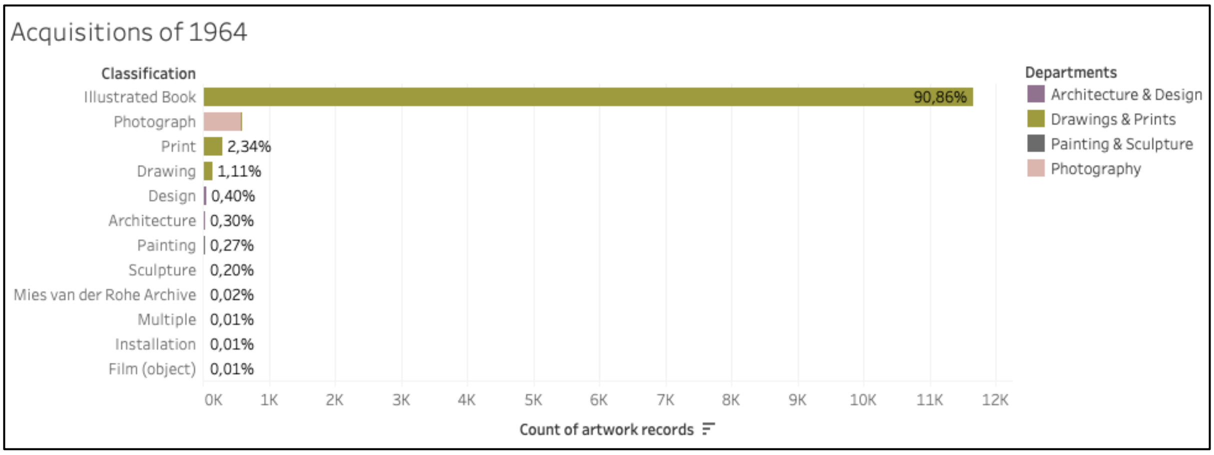

Figure 22 Acquisitions of 1964. Bar plot displaying the Classification feature on the y-axis with the count of artwork records assigned to them on the x-axis. A filter on the Date Acquired feature excluded all records except those that were acquired in 1964. The Department feature is encoded in color, explained by the legend on the side. The image was exported from Tableau.

One other peak will be examined closer: the one in 2008 is more diverse in regard to which *Classification* and *Department* the artwork records are assigned to, but again the peak can be explained by one very large acquisition source. In 2008, it was the acquisition of the "Gilbert and Lila Silverman Fluxus Collection", making up 65% of this year's total acquisitions.[205]

The acquisition history of MoMA can also be explored based on the *Department*s and their acquisition numbers over the years as visualized in the plot below in for each department over time. The visualization allows us to retrieve information on the development of each department. We can see that the *Department*s were not established at the same time, but that the "Drawing & Prints" *Department* was the first one that acquired artworks (1929) and that the "Media and Performance" *Department* has its first acquisition of 1975 assigned to. We can also see that the Fluxus Collection, has only one point of acquisition in 2008 (compare also to the

[205] There are seven different *Credit Line* values used to describe "The Gilbert and Lila Silverman Fluxus Collection". They were combined for this step for this analysis, 99,38% of the records hold "The Gilbert and Lila Silverman Fluxus Collection Gift" as Credit Line, the others are used only in a hand full of records. The other Credit Lines are as follows: "The Museum of Modern Art, New York. The Gilbert and Lila Silverman Fluxus Collection Gift", "The Gilbert and Lila Silverman Fluxus Collection Gift, 2008", "The Gilbert and Lila Silverman Fluxus Collection Gift", The Gilbert and Lila Silverman Fluxus Collection Archive, The Museum of Modern Art, New York, NY.", "The Gilbert and Lila Silverman Fluxus Collection Archive, The Museum of Modern Art Archives, New York, NY.", "The Gilbert and Lila Silverman Fluxus Collection Archive, The Museum of Modern Art Archive", "Gift of Gilbert and Lila Silverman".

peak in acquisition numbers in Figure 21). The 30 records that are linked to the “Architecture& Design – Image Archive” are not displayed because they are missing values for the *Date Acquired* feature.

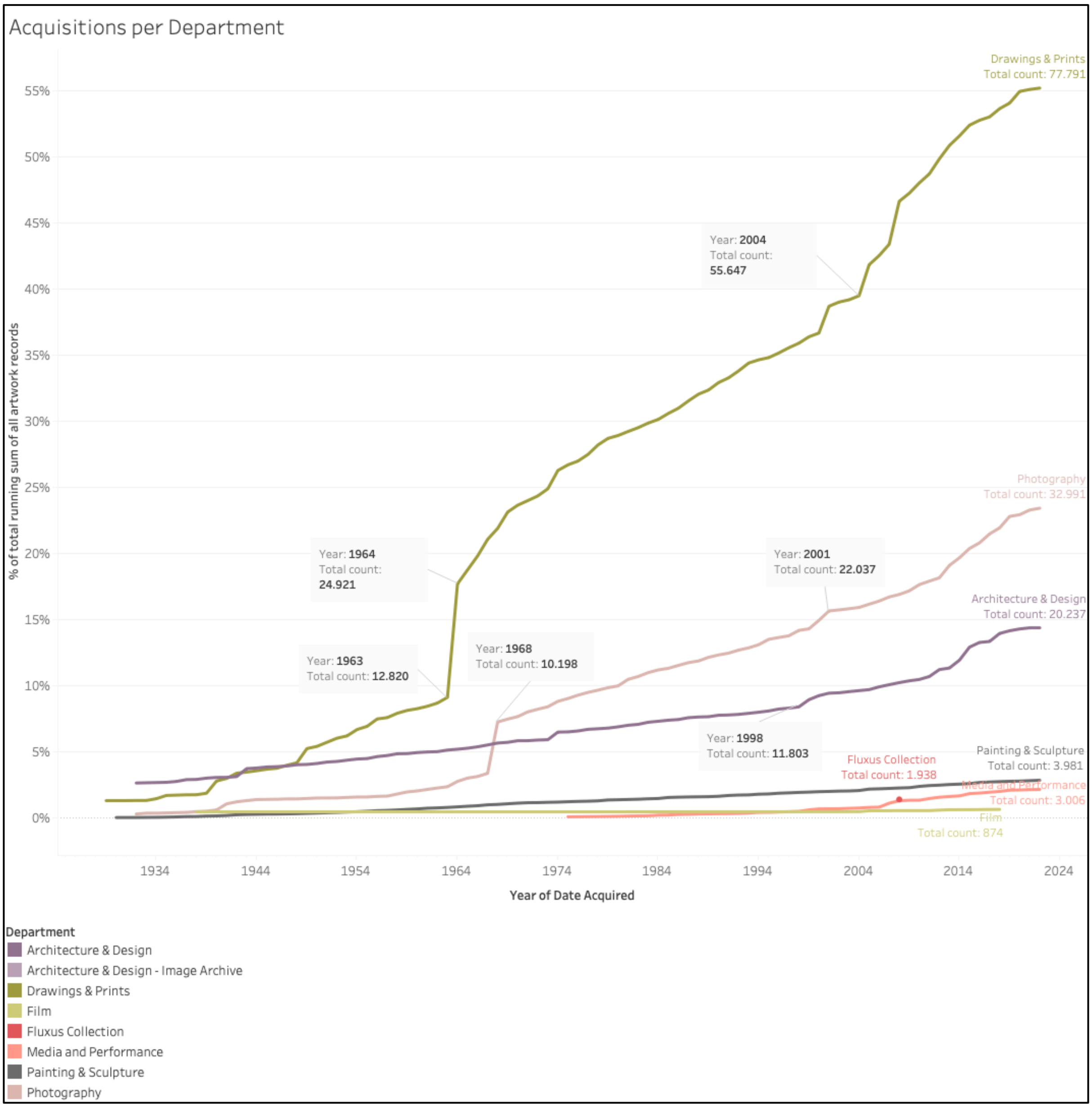

Figure 23 Acquisitions per Department over time. Line plot displays the percentage of the total running sum of the artwork records that were acquired for each Department on the y-axis. The years from the foundation of the museum in 1929 until 2022 on the x-axis. The departments are encoded in color, the end of each line is annotated with the total running sum of artwork records assigned to the department. The image was exported from Tableau.

When we take a look at the course of the lines, we see that the four *Department*s “Architecture & Design”, “Painting & Sculpture”, “Media and Performance” and “Film” are more or less linear with a constant growth rate. Compared to that the *Department*s “Drawings & Prints” and “Photography” both show a time span with steep inclines in artwork count. For “Drawings & Prints” this increase begins in 1963 and holds on until in recent years, having the steepest trend line of all the

Departments. The step of increase in numbers for the “Photography” *Department* takes place in 1968 and 1969, followed by a flatter growing rate, compared to the “Drawing & Prints” department. It should be noted that the data that was used was provided by MoMA in November 2022, the numbers for the year of 2022 might therefore not me complete. Figure 24 below allows to see a pattern on when MoMA acquires most of its new artworks within the calendar year:

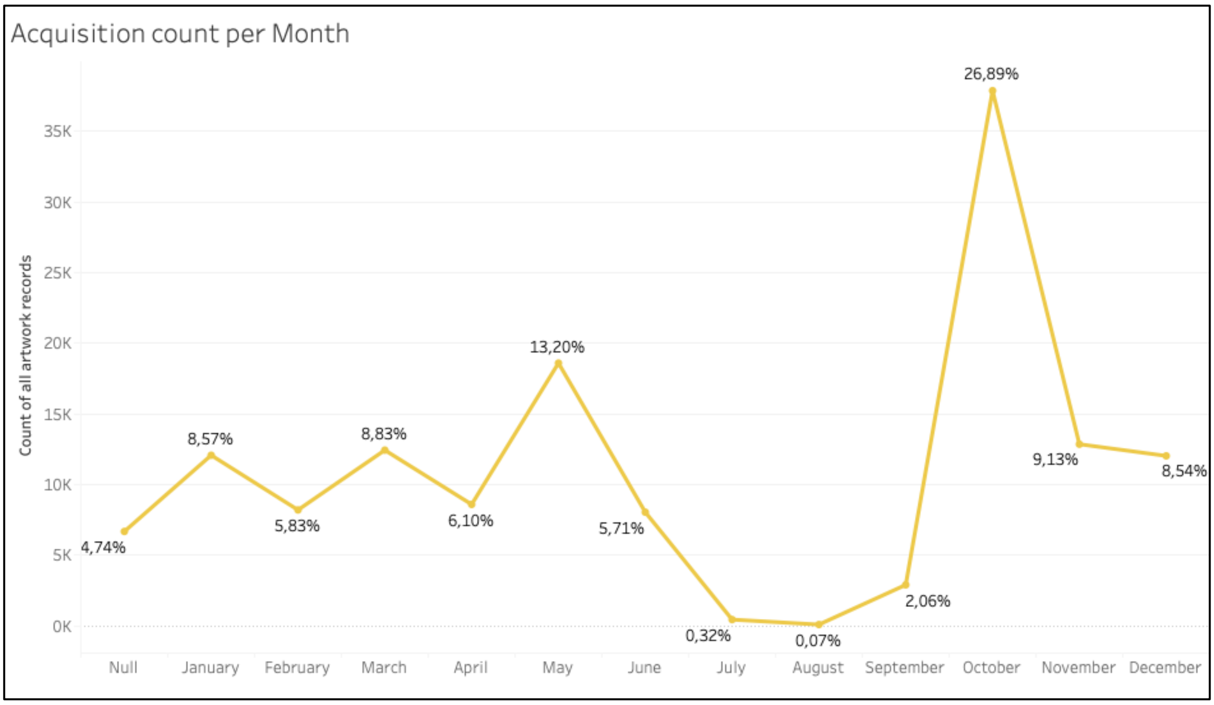

Figure 24 Acquisition count per Month. The trend of all artwork records and the Month in which they were acquired. The count of the objects on the y-axis, the Month on the x-axis. The datapoints are labeled with the percentage of the total artwork record count. The image was exported from Tableau.

. The plot shows the percentage of the total running sum of artwork records for each department over time. The visualization allows us to retrieve information on the development of each department. We can see that the *Department*s were not established at the same time, but that the “Drawing & Prints” *Department* was the first one that acquired artworks (1929) and that the “Media and Performance” *Department* has its first acquisition of 1975 assigned to. We can also see that the Fluxus Collection, has only one point of acquisition in 2008 (compare also to the peak in acquisition numbers in Figure 21). The 30 records that are linked to the “Architecture& Design – Image Archive” are not displayed because they are missing values for the *Date Acquired* feature.

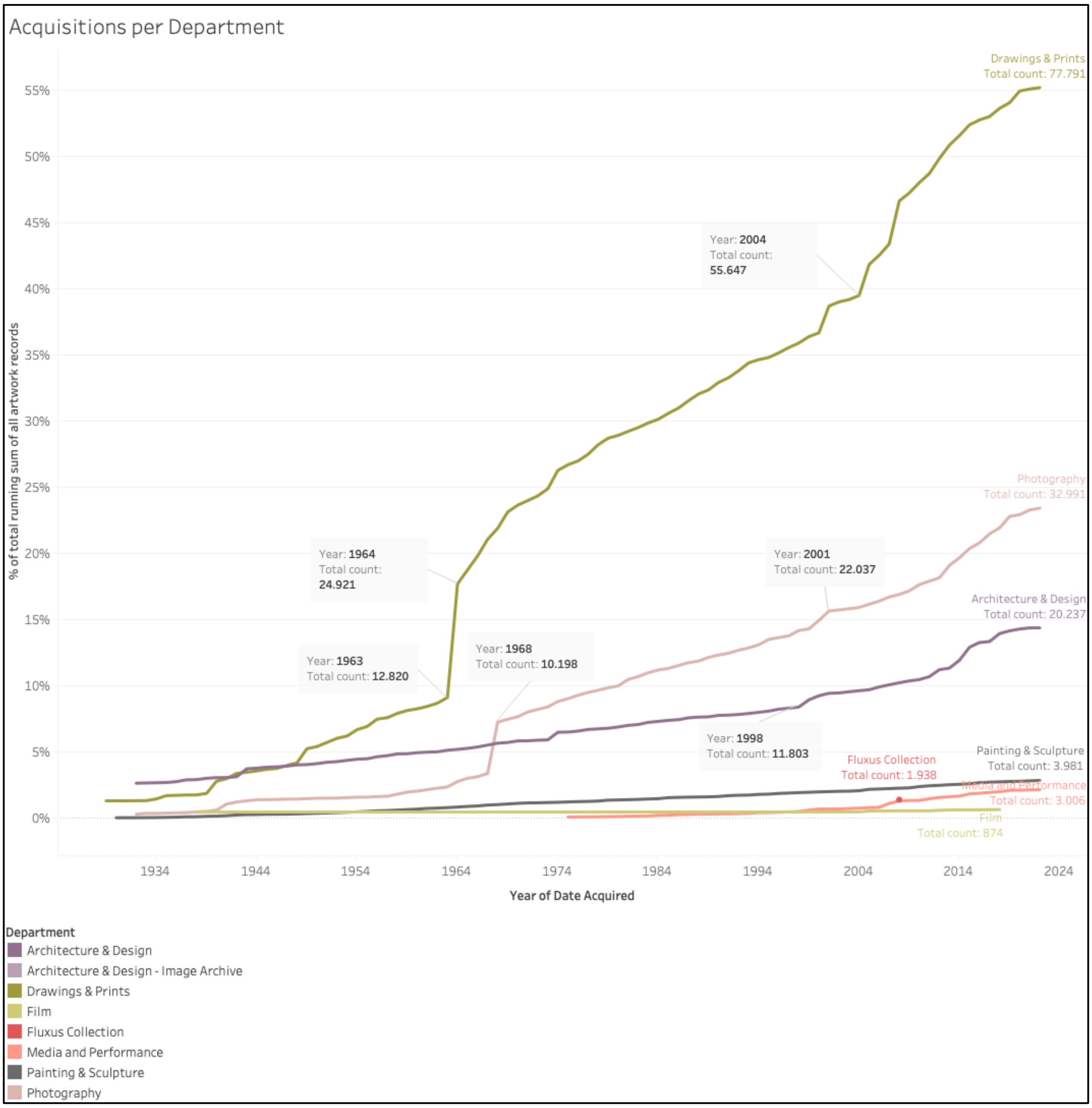

Figure 23 Acquisitions per Department over time. Line plot displays the percentage of the total running sum of the artwork records that were acquired for each Department on the y-axis. The years from the foundation of the museum in 1929 until 2022 on the x-axis. The departments are encoded in color, the end of each line is annotated with the total running sum of artwork records assigned to the department. The image was exported from Tableau.

When we take a look at the course of the lines, we see that the four *Department*s “Architecture & Design”, “Painting & Sculpture”, “Media and Performance” and “Film” are more or less linear with a constant growth rate. Compared to that the *Department*s “Drawings & Prints” and “Photography” both show a time span with steep inclines in artwork count. For “Drawings & Prints” this increase begins in 1963 and holds on until in recent years, having the steepest trend line of all the *Departments*. The step of increase in numbers for the “Photography” *Department* takes place in 1968 and 1969, followed by a flatter growing rate, compared to the “Drawing & Prints” department. It should be noted that the data that was used was provided by MoMA in November 2022, the numbers for the year of 2022 might therefore not me complete. Figure 24 below allows to see a pattern on when MoMA acquires most of its new artworks within the calendar year:

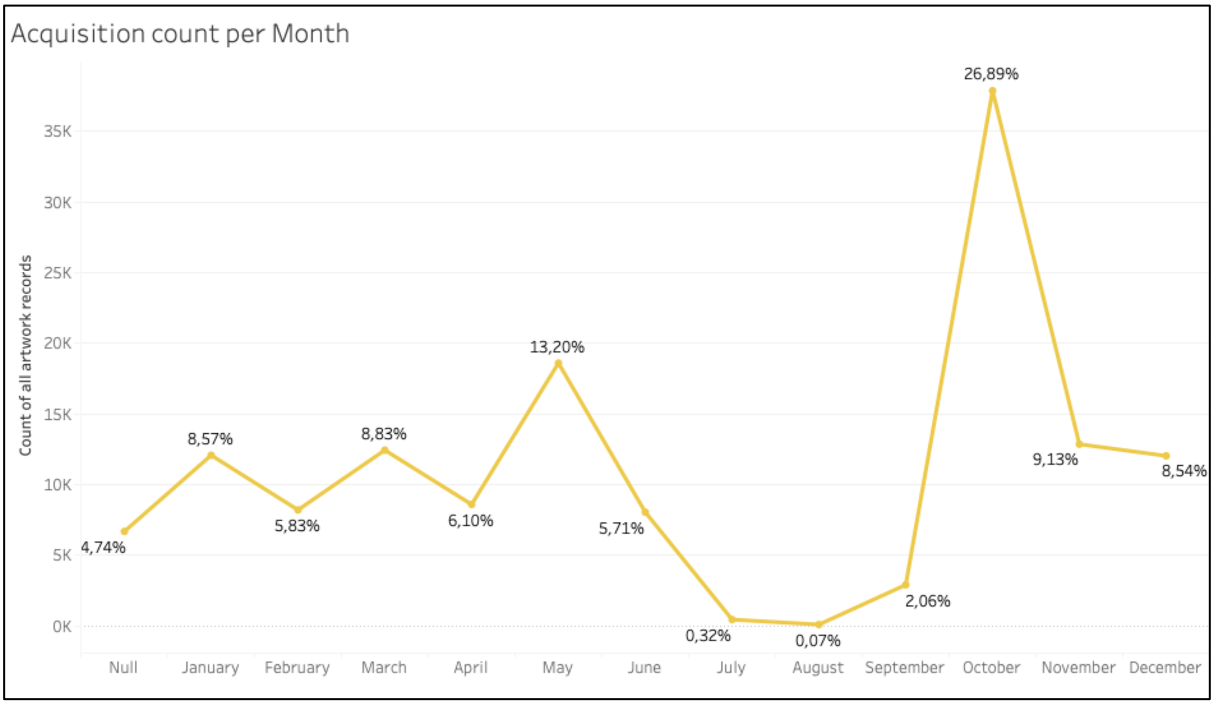

Figure 24 Acquisition count per Month. The trend of all artwork records and the Month in which they were acquired. The count of the objects on the y-axis, the Month on the x-axis. The datapoints are labeled with the percentage of the total artwork record count. The image was exported from Tableau.

A considerable large number of records, 5%, have no data for the *Date Acquired* feature, which was used to compute this visualization. The Null values are displayed on the left side of the plot. In the months January, February, March, and April between 6 and 9 percent of the yearly acquisitions are done, followed by a small incline in May up to 13%. A considerable decrease to June down to 6%, and further down to almost no new acquisitions in July and August (0,32% and 0.07%) might be explained by a summer break of the museum. This break is followed by the peak of the yearly's acquisition numbers in October, were 27% of all year's acquisitions are done. November and December are again on a similar level as the first month of the year.

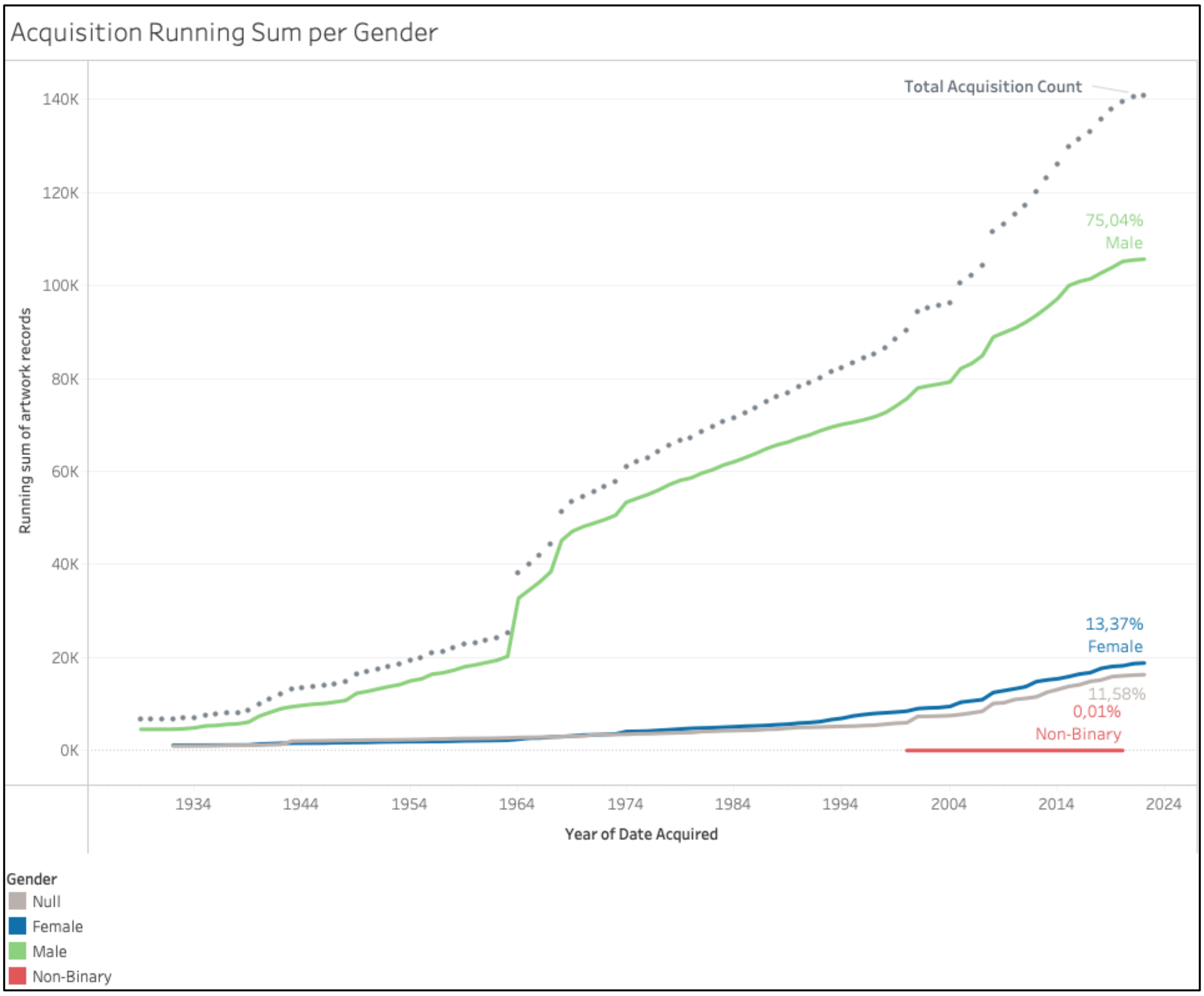

Figure 25 Acquisition Running Sum per Gender. Line plot displays the running sum of artworks for each year. The count is displayed on the y-axis, while the time is on the x-axis. The acquisitions are separated based on the Gender of the linked artists. The Gender category is encoded in color. Additionally, as dotted line, the total acquisition count of artwork records is displayed in gray. The lines are annotated with the percentage of running sum of the total artwork count. The image was exported from Tableau.

Figure 25 shows, as inFigure 23, the total running sum of the artwork records over time, but here the lines were split up by the *Gender* category the artists are assigned to. The visualization shows the already multiple times found imbalance of artworks by male and female artists, and it is possible to see that this trend has not changed significantly over the last decades. Although a slight increase in the artwork numbers of female artists can be found since the late 1990s. Artworks from Non-Binary artists were collected between 2000 and 2020 but in latest years no new acquisitions were recorded. Per average 206.9 records linked to female artists are catalogued each year, compared to 1.112,5 records of male artists. Further down, when specific moments in MoMA's history are examined, we will again look at the *Gender* distribution to see if the protests of the 1980s changed the acquisition policy.

As described in the chapter on the history of MoMA above, the museum had multiple directors that impacted the development of the collections. Four of them are of significance because they hold the role for multiple years: Alfred Barr, Rene d'Harnoncourt, Richard Oldenburg und Glenn. D. Lowry. Determining the periods during which the directors ran the museum and were responsible for its acquisitions is not a trivial task. This is not only because there were sometimes transitional periods that could last several years, such as between Barr and d'Harnoncourt. But also, because the acquisition of works of art can be a process that can take months or even years. One director may initiate the process, but the actual acquisition in the collection might take place in his successor's tenure.

The available data gives no indication on who authorized the acquisitions. What the data does allow, however, is the attribution of artworks to directors based on the *Date Acquired* feature and the knowledge of the director's tenures.[206] Figure 26 displays the running sum of artwork records for each director in a line plot. The two short time directors, John Brantley Hightower, and Bates Lowry together with the period of interim directors between 1944 to 1948 have the smallest impact on the total artwork count. The four named significant directors also appear here with a higher count of artwork records. Alfred Barr, even though he is so significant in the museum's narration of its foundation and history, has the smallest record count (little more than six thousand) compared to d'Harnoncourt (almost 30 thousand), Oldenburg (more than 26 thousand) und Lowry (almost 60 thousand).

[206] To do this it was necessary to assign specific calendar years to the individuals, which lead, in some cases to limitations. The aim was to not create overlapping periods, which would have let to incorrect object counts and results. The time frames were specified as follows: Alfred Barr 1929-1943, Interim directors 1944-1948, Rene d'Harnoncourt 1949-1967, Batey Lowry 1968-1969 (although he did not have the post for a complete year), John Brantley Hightower 1970-1971, Richard Oldenburg 1972-1994, Glenn D. Lowry 1995-2022.

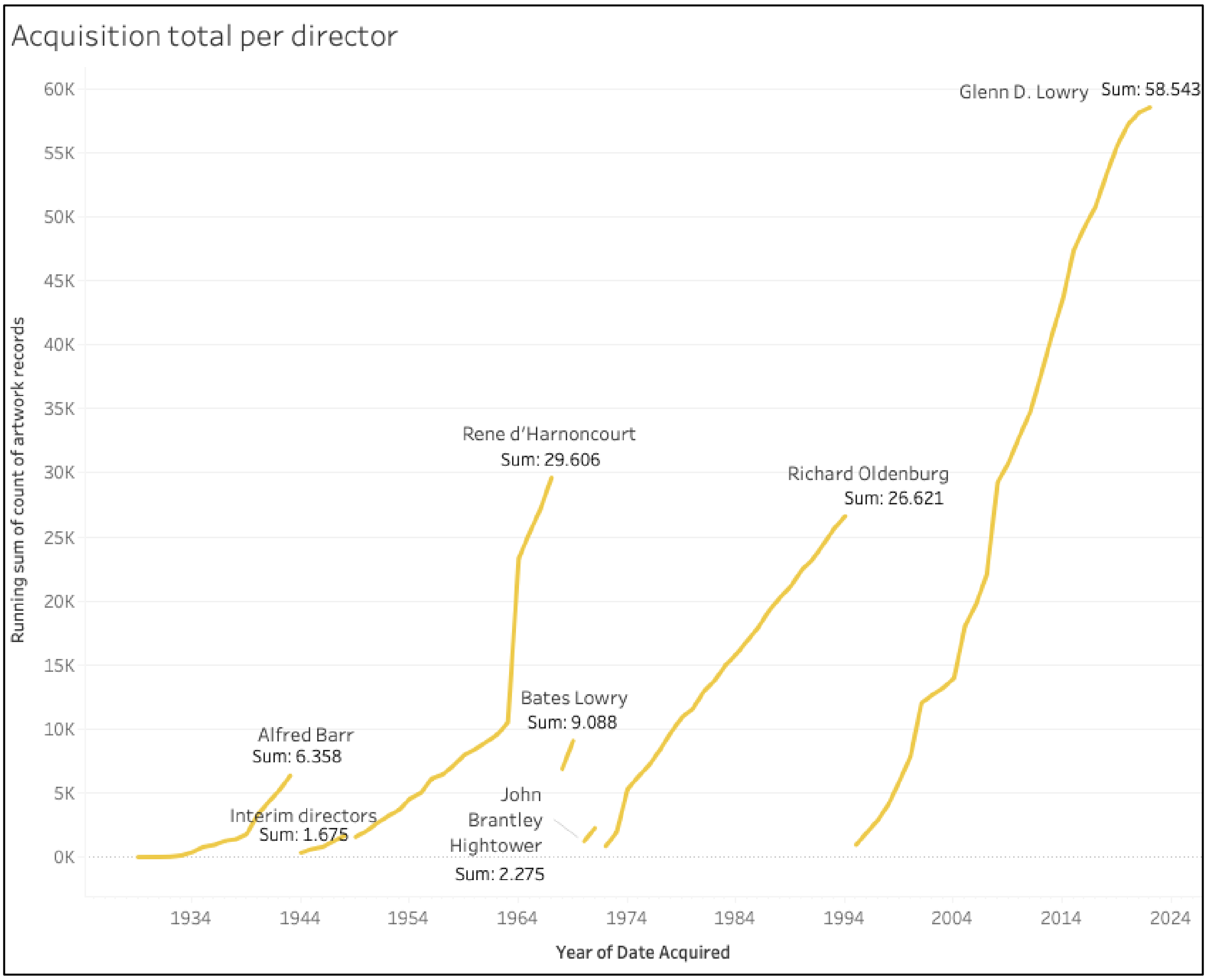

Figure 26 Acquisition total per director. Line plot displays the running sum of artwork records on the y axis and the time in years on the x-axis. The data is separated based on the tenure of the directors of MoMA. The ends of the lines are annotated with the running sum of the director's acquisition count and his name. The image was exported from Tableau.

Lowry has the highest average acquisition count of 2090 artworks per year, followed by d'Harnoncourt who's average yearly count was 1558 during his tenure. Richard Oldenburg has the yearly average of 1157 new records and Alfred Barr only 423 new records per year. To summarize, during all tenures, the collection grew in size, the speed and average acquisition count per years differs for the individual directors.

The following visualizations aim to show differences and commonalities of the four main directors of MoMA. At first the acquisitions will be examined based on the departments the directors assigned the artworks to, followed, again, by an examination of the gender distribution within each acquisition set of the four directors.

Figure 27 shows the distribution of artworks within the departments for each of the four main directors. The x-axes, showing the record count, is synchronized for all four plots which allows again to show the difference in artwork count between them. Alfred Barr's acquisition policy can be summarized as to be focused on three *Departments*, while "Drawing & Prints" made up almost half of all his acquisitions. "Architecture & Design" and "Photography" are close on second and third place. 5% of his acquisitions was assigned to the "Painting & Sculpture" *Department*. Barr also already collected for the "Film" *Department*, albeit in very small number.
His successor, Rene d'Harnoncourt, mainly focused on "Drawing & Prints", 80% of all his acquisitions were assigned to this *Department*. He did not access artwork for the "Film" *Department*. Richard Oldenburg was the first director who collected artworks that were assigned to the "Media and Performance" *Department*, even though it made up only 1.85% of his entire acquisitions. As well as the previous directors, he again had acquired mostly "Drawing & Prints" (65%), followed by "Photography" (27%). 0.02% of his acquisitions were acquired for the "Film" *Department*. Glenn D. Lowry has acquired most diverse in regard to *Departments* from all four of key directors. Again, the largest part, 50% in his case, of the acquired artworks are assigned to the "Drawing & Prints" *Department*. But for the first time the "Media and Performance" *Department* exceeds the 2 percent mark with 4.19%. The one-time acquisition of the "Fluxus Collection" makes up 3% of all of Lowry's acquisitions.

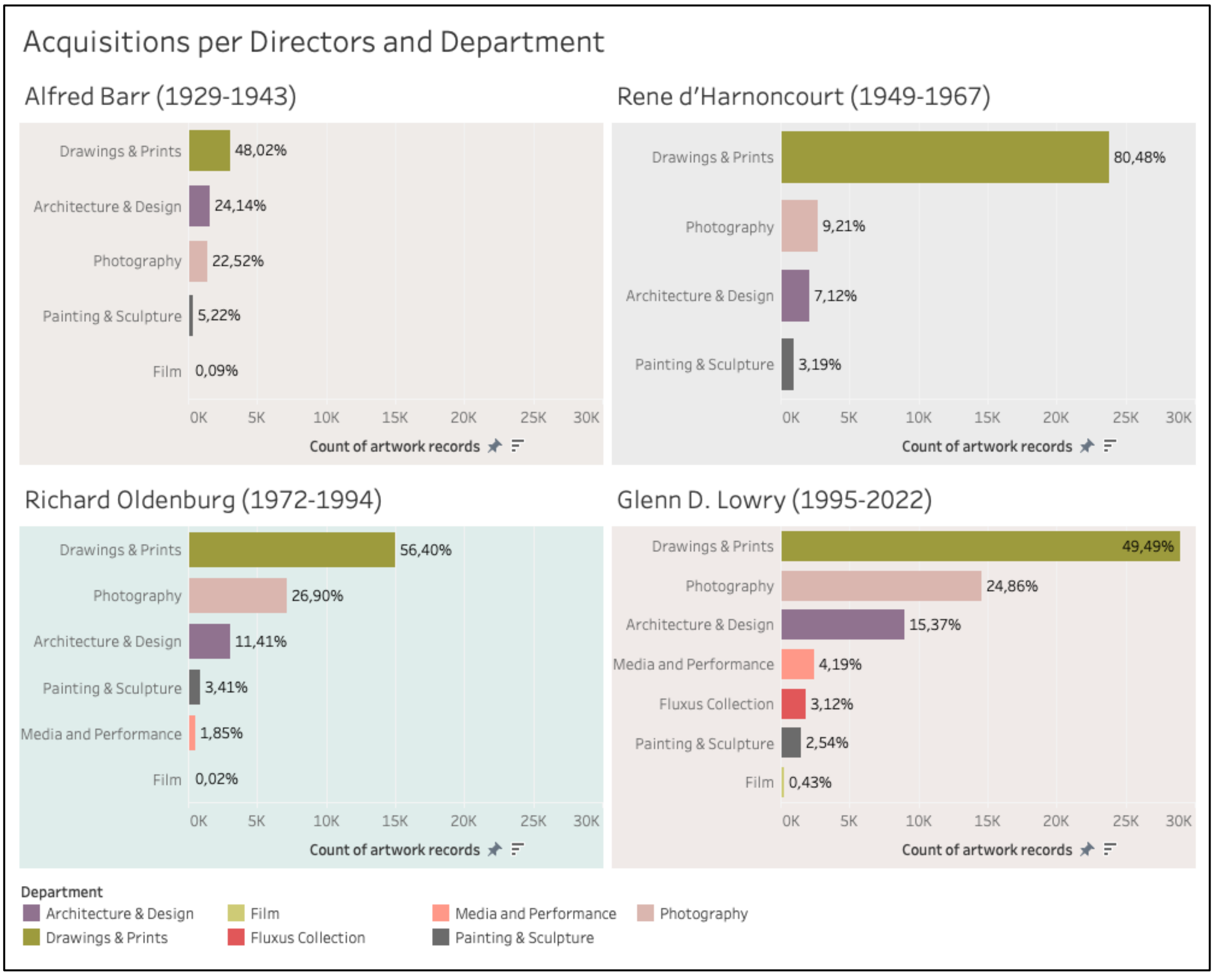

Figure 27 Acquisitions per Director and Department. Facetted view of four visualizations. Each of them dedicated to the acquisitions of one of the directors of MoMA. The departments the acquired artworks for on the y-axes, the count of artwork records per department on the x-axes. The axes are synchronized, all ranging from 0 to 30 thousand. The image was exported from Tableau.

The four directors share that the largest part of new acquisitions is assigned to the "Drawing & Prints" *Department*. Also, "Photography" and "Architecture & Design" are always place two and three, Barr is the only one where "Architecture & Design" ranks before "Photography". As for the small number of records assigned to the "Film" *Department*, it has already been stated above that the records within the available *Artworks dataset* cannot be a complete representation of the film collection of MoMA. It might be a topic for future analysis to further investigate the available film records and how and where the remaining film objects are managed.

The *Gender* feature has already been used in multiple visualizations and it will now be used to examine the distribution of artworks by female, male and non-binary artists within the acquisitions of the four main directors.

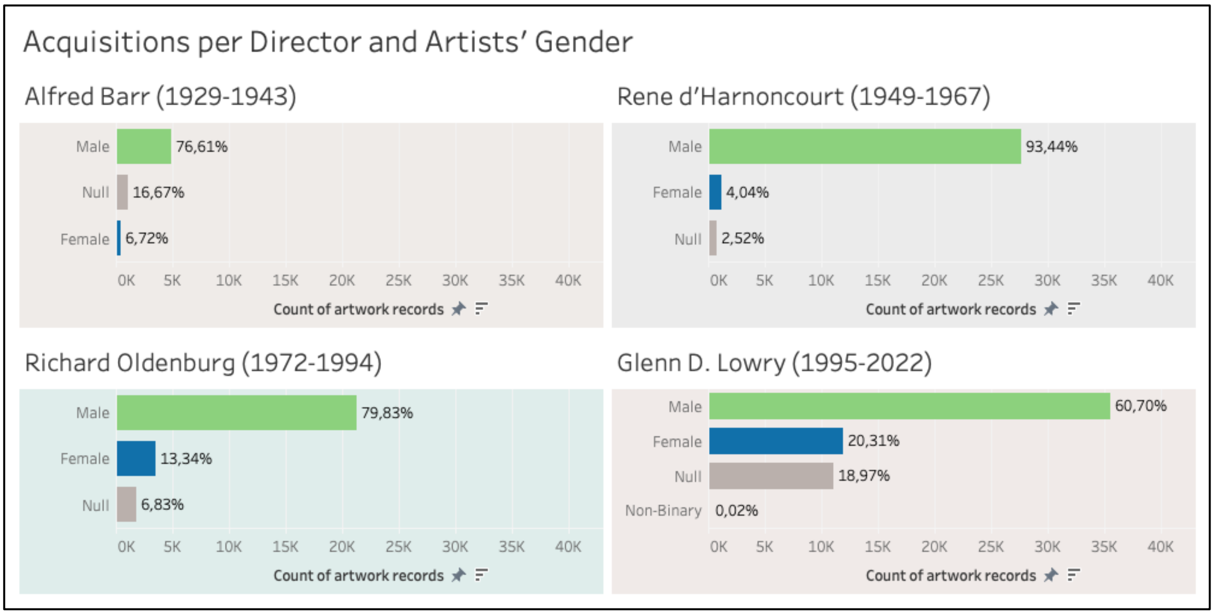

Figure 28 Acquisitions per Director and Artists' Gender. Facetted view of four visualizations. Each of them dedicated to the acquisitions of one of the directors of MoMA. The Gender categories, assigned to the artists linked to the artwork records, on the y-axes and the count of artwork records on the x-axes. The axes are synchronized, all ranging from 0 to 43 thousand. The image was exported from Tableau.

Figure 28 shows a facetted view of four visualizations, each of them representing the artwork record count of one of the four main directors and the *Gender* distribution of the linked artists. The synchronized axes allow again to see the huge difference in number between the director with the highest acquisition number, Glenn D. Lowry, and the one with the lowest, Alfred Barr. The male artists are during all tenures in favor compared to female artists. As already repeatedly shown appear non-binary artist only in a very small number, and exclusively in Lowry's tenure. All directors cataloged artworks with artists without a *Gender* category assigned to them. The percentage differs from surprising 17% for Barr and 19% for Lowry. Especially for Lowry's time this is interesting, because one might assume that the knowledge about the person who's artwork is acquired exists. Especially because those artworks were acquired recently and there might be direct contact to the artist or his/her/their representatives. Missing information might therefore not be the reason for so many artworks without a *Gender* attribute. An assumption is that the awareness of how difficult and problematic it might be to categorize people and their gender holds the museum stuff back to enter anything data they might not be sure about. One other reason might be that more groups or studios, who's records don't hold *Gender* attributes, are responsible for the artwork creation.

The lowest percentage of female artists are recorded for Rene d'Harnoncourt, only 4 % of the artworks were created by female artists, 93% by male artists. During

Oldenburg's time as director the percentage of artworks per female artists grew up to 13%, followed by 20% in Lowry's time.

Summarizing we can state that there is a large difference in the representation of male and female artists. All directors acquired more artworks by men than by women. Only Lowry acquired artworks by artists that are assigned to non-binary as *Gender.* We can distinguish a slight upward trend of acquiring artworks by female artists since Oldenburg's time at MoMA.

The subsequent section will examine the *Credit Line* feature in more detail. Less than 2000 records don't hold data for this feature, it can therefore be regarded as almost complete. But as stated above already during the data understanding phase, the feature holds unstructured, free text data. Nevertheless, there are some values that appear repeatedly.

Credit Line	Count of artwork records
The Louis E. Stern Collection	11.258
Gift of the artist	10.620
Purchase	8.399
The Gilbert and Lila Silverman Fluxus Collection Gift	5.438
Abbott-Levy Collection. Partial gift of Shirley C. Burden	4.929
The Judith Rothschild Foundation Contemporary Drawings Collection Gift	2.472
Gift of Kleiner, Bell & Co.	2.383
Gift of Abby Aldrich Rockefeller	1.889
NULL	1.863
Gift of The Judith Rothschild Foundation	1.686
Fund for the Twenty-First Century	1.615
Anonymous gift	1.456
Mies van der Rohe Archive, gift of the architect	1.429
Given anonymously	1.406
John B. Turner Fund	1.401
Gift of Peter J. Cohen	1.340
Gift of the designer	1.315
Gift of the manufacturer	1.239
Mies van der Rohe Archive, gift of the architect	1.171
Monroe Wheeler Fund	1.097
Gift of Jack Shear	1.030

Table 1 The values of the Credit Line feature that appear the most often and more than 1000 times within the complete Artworks dataset.

Table 1 shows the most often appearing *Credit Line* within the complete *Artworks dataset* is “The Louis E. Stern Collection”. Followed by “Gift of the artist” and “Purchase”. On rank four we find the “Gilbert and Lila Silverman Fluxus Collection Gift”. Both, the “Louis E. Stern Collection” and the “Fluxus Collection” were already examined as peaks in the acquisition numbers over time.
When the *Credit Line* terms are examined for each *Department,* we can see that those four overall highest-ranking terms are also the top four terms of the “Drawings & Prints” department. For the “Architecture & Design” *Department* the most frequent *Credit Line* is “Mies von der Rohe Archive, gift of the architect” (1413 records, acquired 1963 and 1968)[207], followed by “Gift of the designer” (1315 records) and “Gift of the manufacturer” (1239 records). The most frequent *Credit Line* within the “Film” *Department* is “Gift of Chris Lewis”, appearing in 303 records (all acquired in 2017). The “Media and Performance” *Department* lists “Gift of the artists” (433 records) and “Purchase” (348 records) followed by “Committee on Media Funds” (222 records, all acquired 2007 and 2008) as largest sources of acquisitions.
For the “Painting & Sculpture” *Department* we have “Purchase” (207 records) as most frequent *Credit Line* value, followed by “Given anonymously” (109 records) and “Gift of Edward R. Broida” (103 records, all acquired in 2005).
The “Photography” *Department* holds “Abbott-Levy Collection. Partial gift of Shirley C. Burden” as the *Credit Line* with the highest occurrence count (4929 records). Followed by “Purchase” (3934 records, all acquired in 1968), “Gift of the artist” (3421 records) and “Gift of Peter J. Cohen” (1339 records).

This allows for some observations. First, a large part of all records don’t hold specifics on their previous owner, they list “Purchase” as only reference. Second, if a *Credit Line* value appears repeatedly, it is in many cases because of the acquisition of a larger set of objects at one time. For example the acquisition of the “Louis E. Stern Collection”. Third, MoMA specifies the role of the creators in some of their *Credit Line* values when they reference to “Gift of the Artist”, “Gift of the Designer” or “Gift of the Manufacturer” and so on. This information on the role of the artwork creator is nowhere else in the available dataset to be found. And finally, some

[207] There is also the Department „Architecture & Design Image Archive that holds two additional Credit Line values that mention Mies von der Rohe (13 times).

Credit Line values do present information on how the funding for the object was raised but not from where the artwork was acquired from, for example “Fund of the Twenty-First Century” (1615 records) or “Inter-American Fund” (659 records) or the “Latin American and Caribbean Fund” (493 records).

To further investigate the donations by artists the dataset was filtered on all *Credit Line*s that contained “gift of the artist”, “gift of the publisher”, gift of the author”, “gift of the manufacturer”, “gift of the designer”, “gift of the architects”, in all sorts of spellings. 15,870 records were found and the most often appearing *Credit Line* value was “Gift of the artists” (as we have seen) with 10.620 records followed by “Gift of the designer” in 1.315 records and “Gift of the manufacturer” appearing in 1239 artwork records. After examining the combined records, it was found that 65% of the artworks were donated by male artists, while 31% were donated by female artists. A distribution that favors less extreme male artists, compared to the overall gender distribution in the artist and artwork datasets. This is almost entirely because of one female artist who donated a large collection to MoMA, Louise Bourgeois. As we can see in Figure 29, she is, which a huge gap to the next artists, the one person that donated the most artworks (in record count) to MoMA. When her donations are excluded from the results, the gender distribution changes to 85% of donations from male artists and only 9% from female artists. This imbalance is also reflected in the plot, the *Gender* category of the artists are encoded in the color, showing that Bourgeois is actually the only female artists in the top 15 artist donors based on artwork record count.[208]

[208] Beneath the 15 artists only one is listed with a role different of that of the artists. E. McKnight Kauffer is mentioned in 128 Credit Lines of artwork records as artist and as designer.

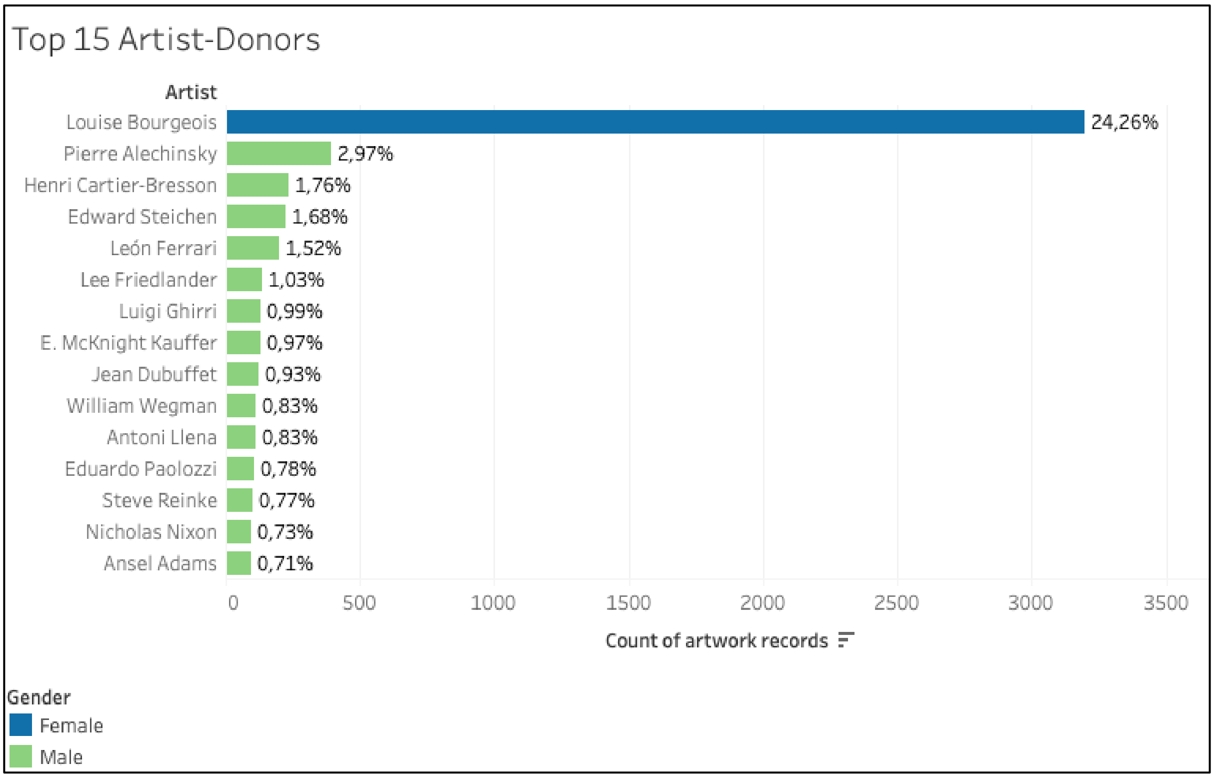

Figure 29 Top 15 Artist-Donors. The artists that are the most frequent donors of their own artworks, based on the mention in the Credit Line feature of the Artworks dataset. The artists on the y-axis, the count of object records in which the specific Credit Line appears on the x-axis. The bars are labeled with the percentage of their record within all artwork records that were donated by their creators. The image was exported from Tableau.

On second rank is Pierre Alechinsky, who donated 391 artworks and followed by Henri Cartier-Bresson with 232 donations. Interestingly, Bourgeois is the only one who's donation span over more than one year, for all others of the top 15 artist-donors the *Date Acquired* value is the same for all their donated artworks.

Louise Bourgeois donated her artworks, in total 3219 records, between 1990 and 2014. Except 3 objects all of them were assigned to the "Drawing & Prints" *Department*. 2837 of them are "Prints", based on their *Classification* feature, 354 are categorized as "Illustrated Books", followed by 32 "Drawings", one "Textile" and one "Multiple". The three donated "Sculptures" are assigned to the "Painting & Sculpture" *Department*. Two of them show in their *Credit Line* the additional information that they were acquired as exchanges.[209]

[209] See Object ID 81981 and 81982 and the Credit Line "gift of the artist (by exchange).

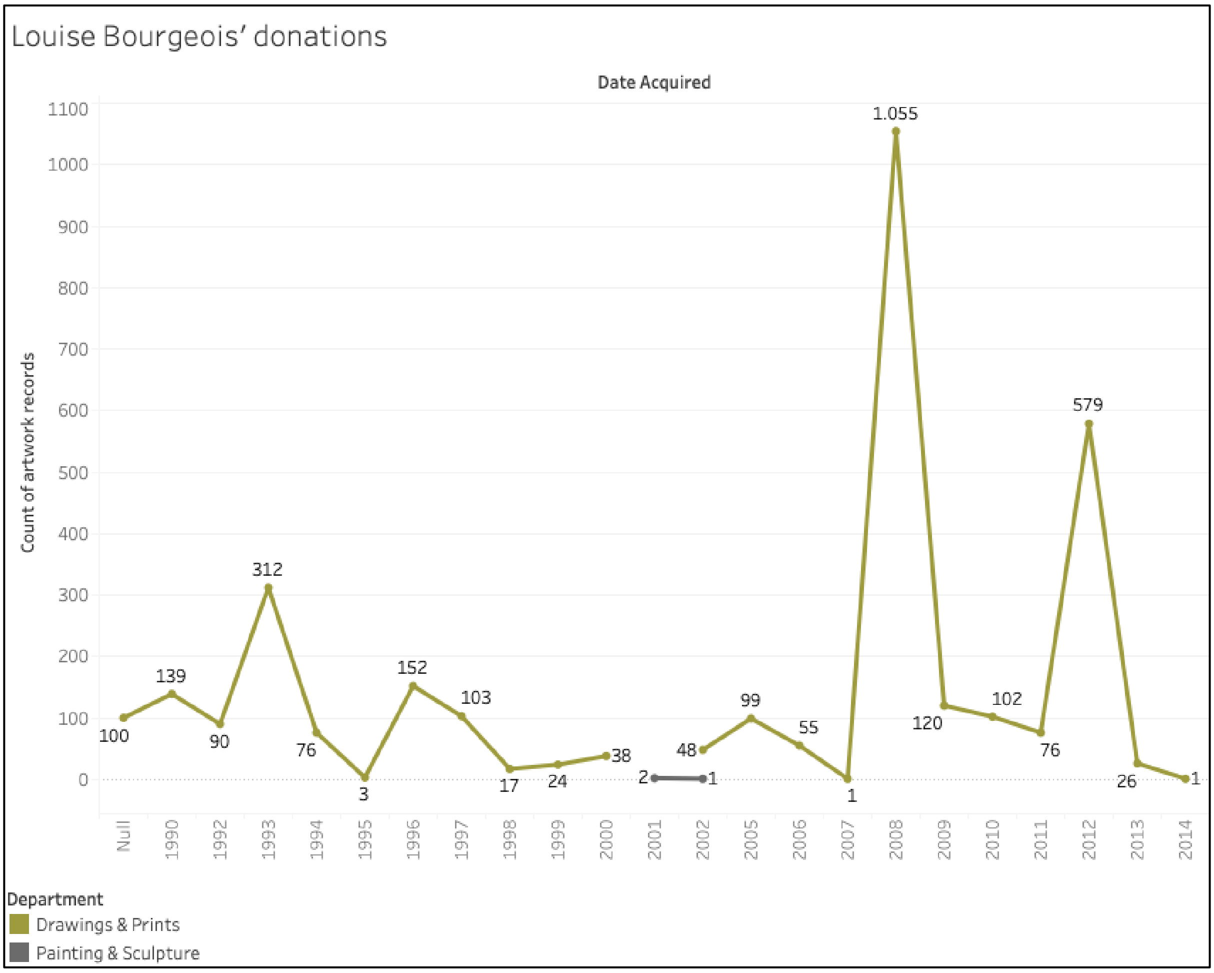

Figure 30 Louise Bourgeoi's donations. The count of artwork records that are linked to Bourgeois and show that she donated the artwork herself in the Credit Line feature on the y-axis. The years of acquisition on the x-axis. The color shows the Departments the artworks were assigned to. With the exception of 3 artworks all are part of the Drawing & Prints department. The datapoints in the plot are labeled with the record count. The image was exported from Tableau.

Figure 30 shows the acquisition numbers of Bourgeois' donations on the timeline. The beginning of her donation history falls into the directorship of Richard Oldenburg, 617 artworks were accepted by the museums as donations during his tenure. The donations got even bigger in number when Glenn D. Lowry took over as director of MoMA in 1995. The largest number of objects, 1055 artworks, were acquired in 2008. Louise Bourgeois passed away in 2010, the 682 artworks that were donated after that death must have been planned beforehand by the artists, otherwise the *Credit Line* would probably not hold the same text as all the ones before.

The examination of the *Credit Line* feature showed that there are some large collections that were acquired together, and which share the same value for *Credit Line*. It was found that not all *Credit Lines* provide details on the previous owner of the artwork. Some only hold the information of the acquisition type, “purchase”, others provide details on the funding but not from where the artworks were acquired from. This hinders the analysis of the provenance of the collection objects. Examining the records that were donated by their creators shows that there is one artist who donated a huge oeuvre to the museum over a large time span: Louise Bourgeois. With her donation she shifts the distribution of female and male artist-donors enormously. Overall, mostly male artists donated smaller numbers of their artworks to the museum, usually once in their lifetime.

One *Credit Line* value that also bridges this section to the following, which looks at specific moments in MoMA’s history, is the one naming Lilli P. Bliss as donor. The collection of the founding member is the narration of the museum the starting point for the story of success of the museum. Some of today’s most iconic artworks of MoMA became part of its collection through her donation. The 66 records represent artworks from Henri Matisse, Pierre-Auguste-Renoir, Paul Gauguin, Pablo Picasso, and Paul Cézanne, to only name the most famous ones. But since this donation was only one of many that took place during Alfred Barr’s donation and because some others are much larger in record count, the donation of Bliss is not recognizable within the dataset as very significant. Additional data which allows to value the worth or significance of artworks would be needed; this could be the exhibition participation, the number of outgoing loans or the insurance values. Figure 31 shows the 15 most frequently occurring values for *Credit Line* in Alfred Barr’s time at the museum. The size of the bubble corresponds to the count of records that hold that specific values.

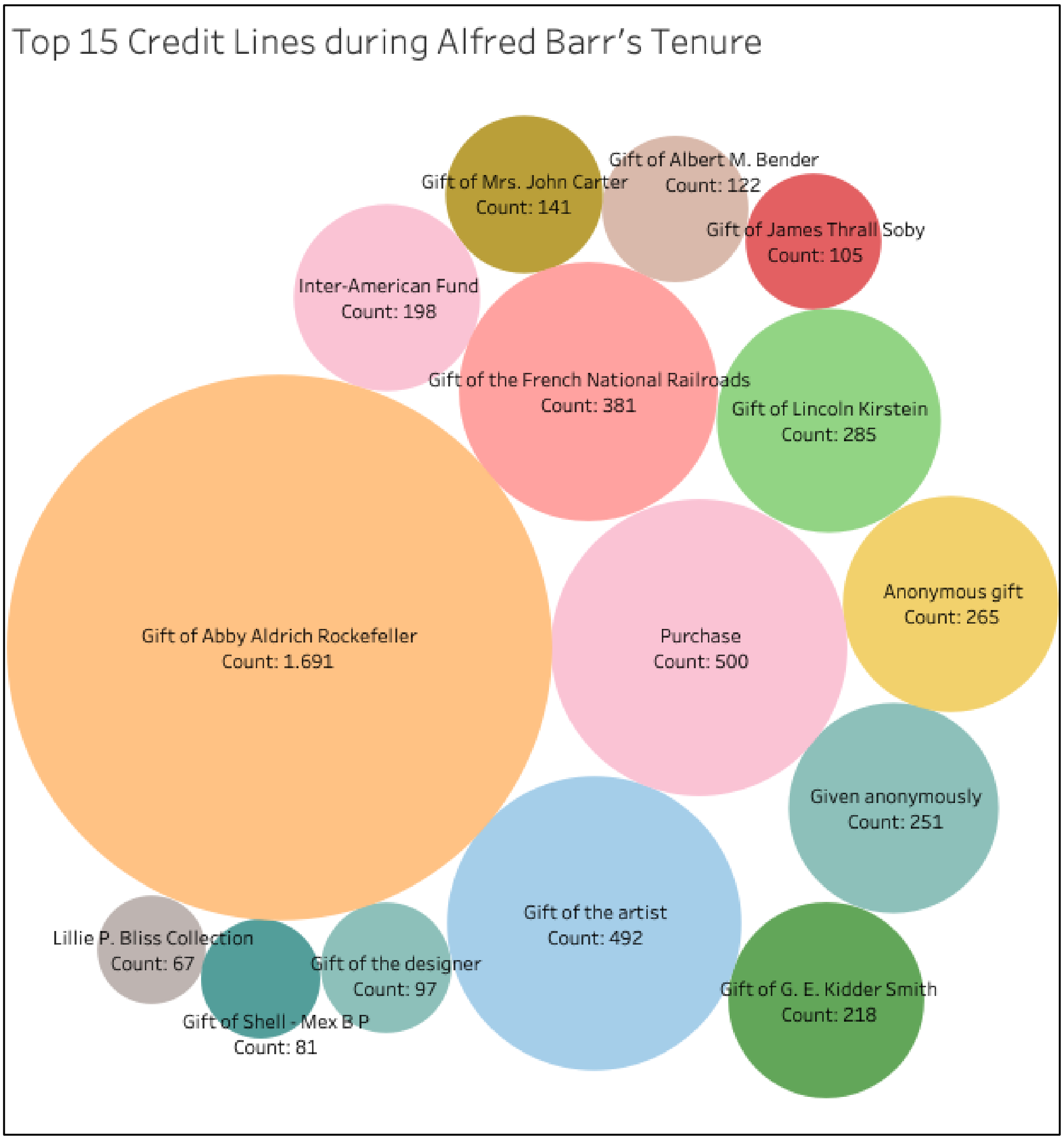

Figure 31 Top 15 Credit lines during Alfred Barr's Tenure. Stacked bubble plot showing the top 15 Credit Line values with their occurrence count during Alfred Barr's time as directors. The size of the bubbles is the count of the artwork records, the label of each bubble reads the Credit Line and the count. The image was exported from Tableau.

This shows vividly that the acquisition of Lillie P. Bliss' collection constitutes only a small portion of Barr's acquisitions. Instead, the majority of artworks he acquired were gifted by another founding member of the Museum, Abby Aldrich Rockefeller. This is interesting to see because Rockefeller's role as driving force in the establishment and development of the museums is emphasized in the narration of the museum, but her substantial donations don't get the same attention (see the chapter "About the Museum of Modern").

One possible reason that Abby Aldrich Rockefeller's significant gifts receive less attention in the museum's narration may be their diversity. While her donations include famous artists such as Matisse and Picasso, they also include lesser-known artists. What also distinguishes her donations from the ones of Bliss, is the department they are assigned to. While Bliss' records are linked to "Painting & Sculpture", Rockefeller donated mostly "Drawings & Prints". These lesser-known artists, representing a genre that is not as iconic as painting and sculpture, were not as easy idealizable to icons of modern art, which may be the reason that the significance of her major donation is somewhat eclipsed in narratives about MoMA's history.[210]

Significant historical moments

At the beginning of this analysis, one goal was to see if it is possible to detect historic moments within the acquisition history; as we have seen, the donation of the Bliss collection does not stick out within the total of Barr's acquisitions. What we have recognized is the donation of the Fluxus collection (see Figure 21) that appears as a peak in time series visualizations of record counts but also as outlier in visualization that show the development of the *Department* feature (see Figure 23). Following now is the attempt to detect global historically significant developments, like the Cold War and the end of the Second World War, within the dataset.

The underlaying assumption regarding the effects of the Cold War on MoMA was that this conflict decreases in the acquisition numbers of artworks by Russian artists. Multiple things made the analysis difficult. First of all, the Cold War prolonged over a large period of time and even when there were multiple periods were the tension amplified, it is difficult to define those periods within the time series analysis of the museum. As has been described above, the process of acquired artworks might span over month or years and we cannot directly translate political events to the acquisition year. While there are some years where no artworks per Russian

[210] Rubio brings forward the argument that objects that are not as sensitive regarding handling and light expose are the ones that are exhibited more often and for longer time. This could be an argument for why paintings and sculptures become easier icons of art, because they are on view more often than sensitive prints or drawings. See: *Domínguez Rubio*, Preserving the Unpreservable.

artists were acquired, the collection on Russian art grew constantly also throughout the years of the conflict, see Figure 32.[211]

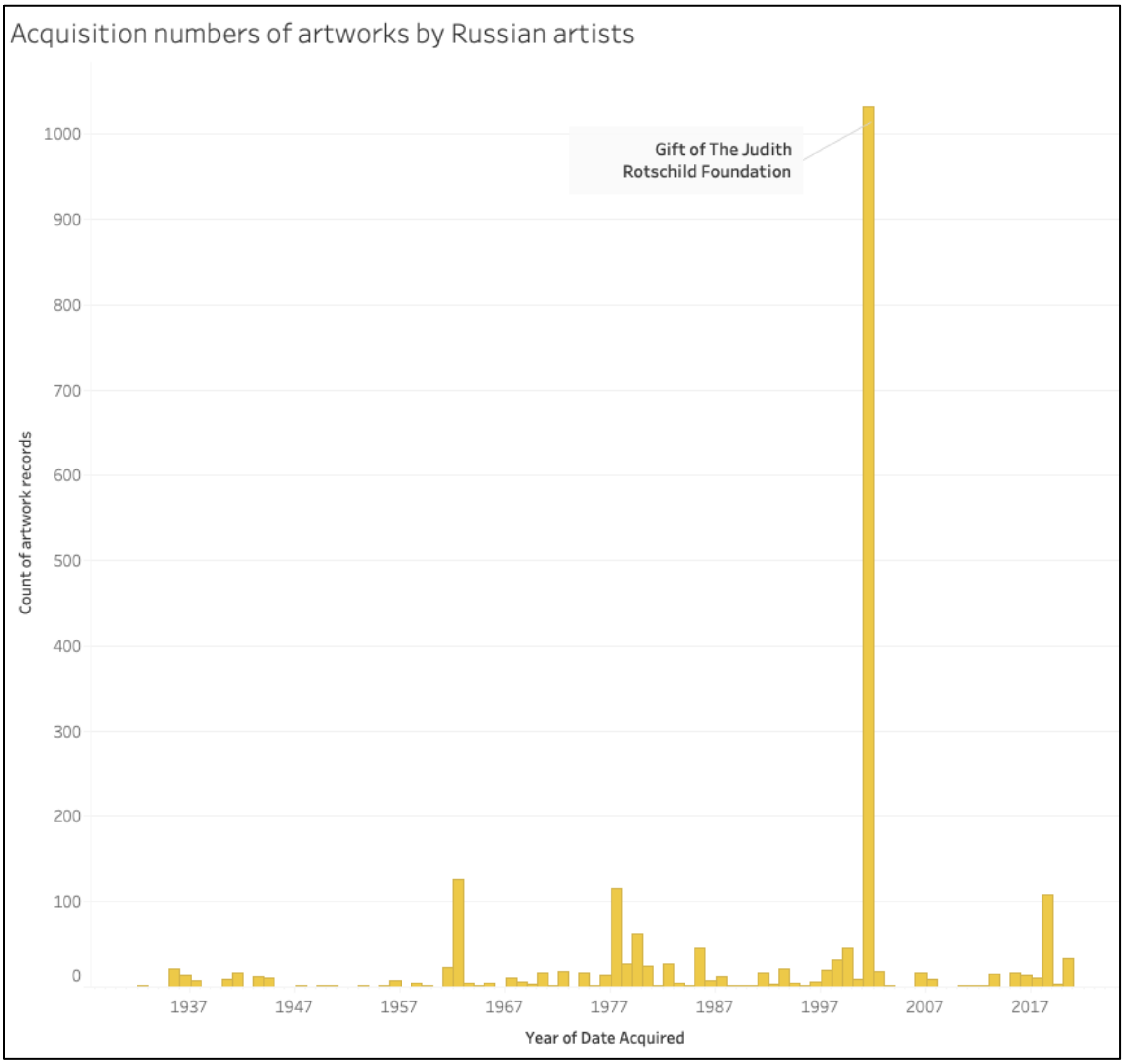

Figure 32 Acquisition numbers of artworks by Russian artists. Bar plot of the count of artwork records for the years in the Date Acquired feature. The data is filtered on Nationality, showing only the records that are linked to artists with the Nationality "Russian". The image was exported from Tableau.

Of course, it is not possible to say if the acquisition numbers would have been higher without the conflict. Sandra Zalman writes in her article about the link between art and politics, how MoMA tried to build on this connection as part of its mission and that the museum had to fight accusations during the Cold War. MoMA had to argue against the claim that modern art would be communistic in general. [212] This proves

[211] No artwork records, linked to artists that have Russian as *Nationality* value assigned to them are available for the years: 1945, 1946, 1948, 1951, 1953, 1957, 1960, 1966, 1973. Considering the period of 1947 until 1991 as the time span of the Cold War.

[212] *Zalman*, Unpacking the MoMA Myth, 291.

that the conflict was topic in the museum and that it probably also affected the museum's decision regarding what to acquire. Nevertheless, the largest acquisition of Russian artworks took place years after the conflict ended. In 2008 "The Judith Rothschild Foundation" donated more than thousand artworks to the MoMA. If that donation took place so late in the history of the museum due to the conflict, cannot be known based on the data in the datasets. But it might be regarded curious even without additional information that the largest part of the museums Russian artworks came to the house so late.

The second significant historic event that might be discoverable in the dataset is the Second World War. The assumption was that there is an increase in European artists and their artworks in the years before and during the war, as many fled Europe in order to escape Nazism. Their appearance in America might increase their visibility in the museums and collections. But also, this examination is difficult, the filter on all nations that are assigned to the continent Europe would include a too large number of countries, some of them not part of the emigration movement we are interested in. After some attempts to select different European countries, the decision was made to only look at Germany and Austria and their artists. The second difficulty lies in the fact that we don't know if the acquisition of German and Austrian artworks during the years of Nazism was caused by this conflict. Artworks of those artists might also be acquired by chance in those years.

When we look at the acquisition numbers of artworks by German and Austrian artists, an increase in the first years of the second world is visible, with a break directly after the war. But when the 78 artists who's artworks fall into this period (1936, to also include the years of prewar emigration, until 1945) are looked at in detail it gets more complicated. We find that twelve of them already passed away before WW2 started[213]. Twelve others died during the years of war, under which circumstances, if in Europe or in America is not clear based on the details in the

[213] Paula Modersohn-Becker, Hermann Bek-Gran, Franz Marc, Paul Adolf Seehaus, Wilhelm Lehmbruck, Lovis Corinth, Johannes Theodor Baargeld (Alfred Emanuel Ferdinand Gruenwald), Paul Leni, Otto Mueller, Max Slevogt, Max Liebermann and Paul Gangolf.

dataset.[214] And also for the remaining once, we are missing data in order to state if there was an increase in acquisitions of artworks by artist fleeing Nazism.
One more thing that makes this even more complex is that it might have happened that artists who fled Europe took up the US citizenship after their emigration. Those artists don't appear in the current search result.
To summarize, it is not a trivial task to find historic developments within the available dataset; additional information would be needed in some cases, in others the complexity of reality stands in the way of finding clear answers, as has been shown in the case of the possible effects on wars on the collection developments.

The last example for the analysis of significant moments within MoMA's history and if they can be detected within the dataset is the W.A.V.E protest in 1984 (for more detail see chapter 3.1 on page 37). The question is, if the protests against the underrepresentation of female artists in the museum's collection and their exhibitions changed the acquisition numbers of artworks by female artists. Figure 33 shows the trend of record count of female artists over the years, based on the *Date Acquired* feature. The timeline starts at 1932, with the first acquisition of an artwork by female artist. The W.A.V.E protest took place in 1984, an increase on the representation of female artists as reaction to this protest could earliest be visible in 1985. But as the line plot shows, only a small increase in 1985 took place and in 1986 the acquisition numbers even dropped down -41% from 124 to 73 artwork records. The 10-year average of the artwork record count of female artists does also not show an increase in the 1980s, but on the contrary, compared to the decade before and after even a decrease in average yearly acquisitions is recognizable.[215] From an average of 146

[214] Otto Schoff, Christian Rohlfs, Ernst Ludwig Kirchner, Ernst Barlach, Paul Klee, Heinrich Nauen, Rudolf Grossmann, Otto Von Wätjen, Alexander Olbricht, Karl Bauer, Oskar Schlemmer and Käthe Kollwitz.

[215] The 10-year periods start from 1932 with the first acquisition until and including 1941. The last period from 2012 onwards is 11 years long, including 2022. The average acquisition per year in the time span from 1962 until 1971 was 123 records (in total 1231 records), growing to a yearly average of 146 in the years from 1972 until 1981 (in total 1463 records). The average drops down to 117 records in the years from 1982 until 1991 (in total 1171 records). The trend increases in the following years to 300 yearly averages, to 470 and 464 in the last 12 years ranging from 2012 until 2022.

artwork records in the years of 1972 until 1982 the numbers go down to 117 average yearly acquisitions by female artists in the years from 1982 to 1991.

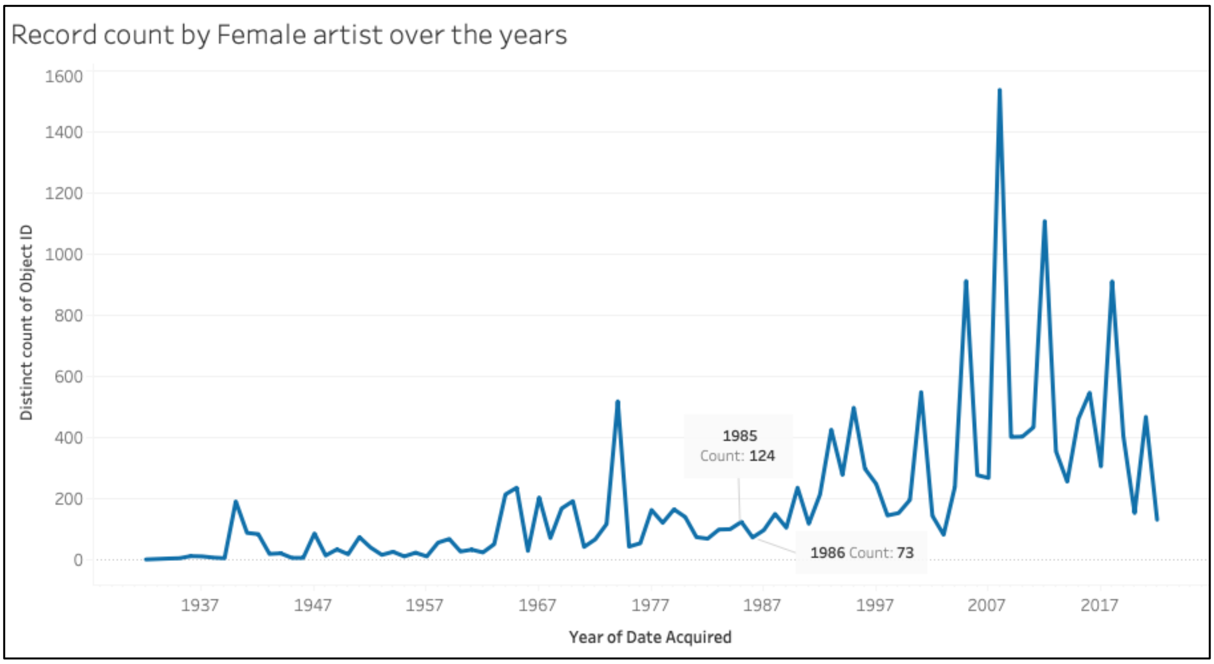

Figure 33 Record count by female artists over the years. The trend line of record counts over the years based on the Date Acquired feature. The data is filtered to only show artworks records that are linked to artists categorized as Female in the Gender feature. Two datapoints are annotated, the year 1985 with 124 records and the year 1986 where only 73 artworks were acquired and cataloged. The image was exported from Tableau.

As we have seen in the chapter on MoMA’s acquisition policy (see page **Error! Bookmark not defined.**), during Richard Oldenburg’s tenure, which lasted from 1972 until 1994, the general yearly acquisition count was lower compared to Rene d’Harnoncourt before him and Glenn D. Lowry after him. The overall percentage of artworks by female compared to male artists rose during his tenure though compared to the tenure of d’Harnoncourt though, as has been shown in Figure 28. The increase is mainly limited to the last years of his tenure, to the 1990s, as can be seen in Figure 34.

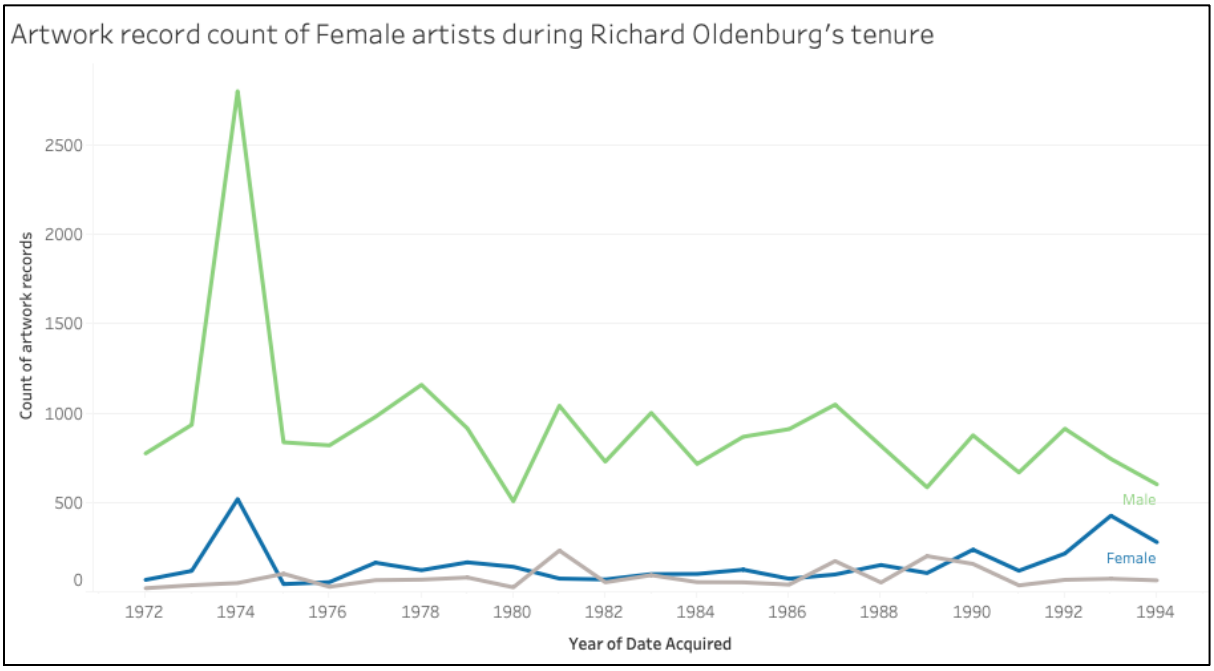

Figure 34 Artwork record count of Female artists during Richard Oldenburg's tenure. Trend lines of artwork record counts split up by assigned Gender category of the linked artists for the years in Date Acquired. The time is reduced to only include the years Richard Oldenburg was director of MoMA. The Gender category is also encoded in color, the lines are labeled accordingly. The image was exported from Tableau.

This might be due to a belated response to the protests, or the general growing awareness of gender-based discrimination in the art world. To state this, additional data or specific knowledge on the decision-making processes during Oldenburg's time would unequivocally be needed. Based on the available data, it is only possible to state that there is no immediate increase in acquisition numbers of female artists after the protests 1984 visible, but the overall it is recognizable that the trend of acquiring more artworks by female artists increases.

Catalogued

As the final step of this analysis the Catalogued feature will be examined. Since its meaning was not explained by MoMA, it was assumed to either represent a kind of publication status of the artwork records or a marker on the content of the records based on curatorial approval. As the "readme" file states there should be a marker like this, but it was not specified in more detail (see chapter 3.2.1 on page 54). As has been shown in the description of the online presentation of the MoMA collection on their website (see chapter 3.1.2 on page 50), there is, in some cases a note that the record may not be complete. There was the idea that this note might correspond to the Catalogued feature. This can be ruled out based on one finding, all records that hold a "N" for Catalogued, don't hold a *URL* nor a *Thumbnail URL*. They

therefore cannot be those that show a note in their web presentation. From the 67% of records that hold "Y" in the feature, all have data in the *URL* field, which means they are represented in the online collection of MoMA. Some of them (11%), don't have a *Thumbnail URL*; those artworks are solely represented by metadata.
The percentage of records that are represented on the online collection differ between the directors, see Figure 35. 78% of the acquisitions of Glenn d. Lowry are part of the online collection, the highest number, on the other side of the ranking are the acquisitions made by Rene d'Harnoncourt, only 51% of the records are online.

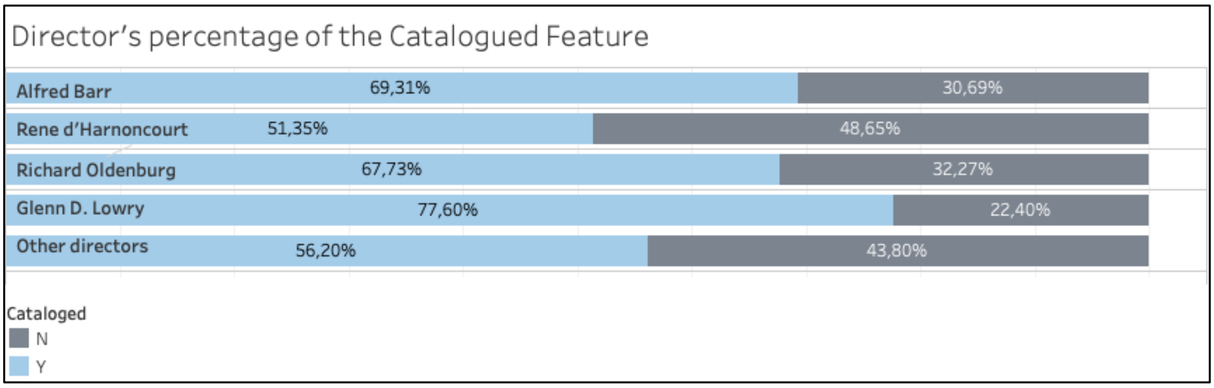

Figure 35 Director's percentage of the Catalogued Feature. Shows the total record count that are assigned to the tenure of the specific directors and their Catalogued feature. The feature Catalogued is encoded in color. The percent are based on each row of the table. The rows are labeled additionally with the names of the directors, all records that are not assigned to one of their tenures are combined visualized in the bar "Other directors". The image was exported from Tableau.

We can summarize, the *Cataloged* feature marks if the record is available in MoMA's online collection. The percentage of records that are available differs for the different directors; this might be an indication of record quality that might be examined further in future analysis.

3.3.4 Analysis summary

The investigations of the *Nationality* feature showed that the idea of assigning one nationality to each record does not fit the reality. Firstly, for those persons whose nationality or belonging does not go along with the nation states and current boarders; for example, Catalan artists but also Native American or Canadian Inuit. The structure of the feature obscures realities of minorities, a fact that the museum seems to be aware of, as has been seen in their attempt to enrich their information on the persons biography in the *Artist Bio* field. The structure also does not work for records that do not represent single individuals. Groups of artists, but also companies or organizations cannot easily be assigned a *Nationality.* The analysis of the *Gender* feature showed a large discrepancy between male and female artists. It also shows that only recently records with non-binary artists appeared, but besides that no further categories are used in MoMA's dataset.

The analysis of the artists records showed that MoMA actually keeps their own records on the artists, even when they are linked to external sources, and they might create diverging data by doing so. It has also been shown that the collection of links to those external sources are not complete, there are records were records in those authority databases would be available, but they are missing from MoMA's records.

Regarding personal information of persons, like gender or nationality, the general question, if that is information that should be recorded by a museum can be raised. How insightful is a feature like *Nationality* when it is so limited in its use and obscures everything that does not fit the Western understanding of citizenship? Cataloguing artists as "Native Americans" might be a distinction welcome to the person the records represent, a sign of visibility and acknowledgement. At the same time, it might be used discriminatorily. And alone the fact that those records are distinguishable from the majority of records can be problematic.

The *Gender* feature can be critically evaluated as well, we saw that in recent years the numbers of records without a value for this feature increases, this might be a hint to the fact that the structure of the field does not fit the needs of the catalogers or the persons that shall be represented by the records. Even if MoMA has a handful of records that are assigned to non-binary as *Gender* category, it is a valid question to ask if those persons are the only ones that wouldn't assign them self to the binary system of gender classifications. The analysis of this thesis used both the *Gender* as well as the *Nationality* feature extensively and valuable insights were drawn from it. But it might be worth considering if the available structure of those fields could be improved or boarded and if the data that is kept in them right now actually can be trusted even though we have seen how limiting they can be.

The analysis of object types in the MoMA collection was short and not very deep since the available features did not allow extensive research. We have seen that the gender representation differs for each of the *Departments*, the primary form of organization within the museum and that there are surprisingly high numbers of incomplete artists records linked to the “Film” department. More fruitful was the investigation of MoMA’s acquisition history. The visualizations allow insights in the acquisition policies of the different directors and the museum in total. And they visualized a shift in focus based on different *Department*s and of the gender distribution for each of the directors.

The search for clues for historic moments that are reflected in the datasets proved to be much harder and complex than expected. It showed that complex developments like the Cold War or Second World War are also complex to investigate. In these cases, additional knowledge and resources are needed to make more informed statements about how they effected the museums development. For example, declined acquisition plans or orders to not buy art from artists from communist states in the case of the Cold War. Without them, and using only the available data, it is unclear how those wars affected the acquisitions of MoMA. The example on the important acquisition of the Lilli P. Bliss collection in the early years of the museum shows that what is deemed to be significant in the narration of MoMA is not always detectable as such within the available data. It also showed how much is left out of the narration, many large acquisitions were conducted in Alfred Barr’s time that could be investigated in more detail. It could also be of interest to investigate why those other acquisitions did not make it in the canon of MoMA’s story. Only the investigation of the effect of the W.A.V.E. protests and their possible effects on the acquisition numbers was to be summarized in a clear answer: there was no immediate effect of the protests on the acquisition policy.

To close this chapter, it should be stated that there is no feature for biases in records. Biases that can hardly be queried; they can be stumbled upon by chance, or, and more importantly so, they manifest in absences. A lot of it lies in the missing artworks, in the missing artists, in the missing genres and the missing possibilities to describe things or relations that go beyond the Western understanding of things. The most obvious bias in MoMA’s collection is, that only a small part of the artworks was made by women, the vast majority was made by male artists. And also, most recent acquisition decisions follow this trend, even though when small changes are noticeable.

A future investigation could ask the hypothetical question of how the acquisitions of the museum will look like in the future. Creating a trend model based on a range of recent years and calculating how long it would hypothetically take to even out the gender imbalance in the MoMA collection when the prevailing trend is kept up. It might also be possible to suggest how to adapt the policy in order to get to an equal representation sooner.

This inequality of gender representation could assumably also be stated regarding the race of artists with the assumption that non-white artists are underrepresented in the collection. But since we have no feature that would allow to make assumption on rase this stays a conjecture.

The datasets MoMA provided were carefully curated and they played it safe in their selection of features. Especially the exclusion of any subject related fields, that I assume might hold problematic entries or texts that do not live up to contemporary standards, shields the museum and their staff for critique. Nevertheless, it was a courageous and significant decision by the museum to publish the records. As far as I am aware, they were the first major museum to take such a step. By doing so, they provided the public with the valuable opportunity to observe the museum’s inner workings, and allowed to find potential weaknesses, limitations, or absences in their catalogues.

4 Conclusion

The central focus of this thesis was the investigation of two datasets and how data analysis could be used to extract meaningful insights into the collection and the museums which created it. The investigation was successful; it was possible to show that new insights are to be found, new directions of research open up when taking a distant reading approach, and that with this distance, biases inherent in the collection and its records appear more clearly.

The main goal of the thesis was to showcase that data analysis is a useful method to retrieve insights on collections, which was achieved. It was possible to demonstrate that common questions regarding the collection can be answered. Even though the aforementioned expected insights in the datasets were answered to different extents. While the questions about the gender distribution and the acquisition policy of the museum were extensively investigated, and meaningful results were retrieved, the search for traces of historical events was more difficult. For global events like the Cold War, it was challenging to determine what actually resulted from those conflicts and how they affected the development of the collections. In these cases, additional knowledge and resources are needed to make more informed statements. For example, declined acquisition plans or orders to not buy art from artists from communist states in the case of the Cold War. Without them, the results are uncertain based on the available data, and it is unclear how those wars affected the acquisitions of MoMA. The historical events that are more specific to MoMA were easier to analyze, and even if, again, there are no traces of the impact of the W.A.V.E protests, for example, this forms part of the answer. The protests did not directly result in a change in the acquisition policy of the museum.

Some limitations of the data analysis should be acknowledged and recapped here. First of all, the linking of the two datasets made it necessary to reduce the links of artworks with more than one creator to the first one. This limits the insights on all additional creators of those specific records and affects the results that were drawn. The assignment of specific nationality terms that appeared in the dataset to nation-states, like Canada for 'Canadian Inuit' (to name one example), was done with the best intention not to lose those records in all analyses where the nationality of the artists was used as a feature. However, it was done with the awareness that this assignment might not be correct and that it might not reflect what the individuals represented by the records would choose for themselves. Every decision on how to group records, what to exclude, or where to investigate further can be challenged and critiqued. It was the aim of this thesis to transparently document all decisions that have been made during the analysis. This allows future researchers to repeat this analysis but also shows where different approaches could be initiated and where other decisions might affect the outcome.

Bringing attention to possible biases, constraints and to aspects that should be critically examined was also accomplished, as the example of Nationality can show. Not only based directly on the MoMA data, but also more broadly, it was possible to demonstrate how deeply Western ideas of how to refer to objects and their creators and how to order and categorize them are engrained in collection management systems.

From the presented results new directions of investigation into collections, such as the reasons behind the exclusion of certain collection parts from the common canon or the effects of global events on the museum's politics, were introduced. The hope that data analysis might shift the focus of the disciplines into new directions was effective.

It has been acknowledged that there are missing elements in the used datasets and that they have not been updated for months. Some fields are excluded because they were identified as too sensitive, such as storage locations, monetary values, or personal information. Other data, like subject descriptions, might be withheld because the museum is aware of possible problematic content. Publishing data is always a risk; it comes with the potential harm to people or communities, and caution and sensitivity are necessary. Nevertheless, it would allow for further investigation, research, and insights if more data were to be published. The public might also be asked to assist with the edition and improvement of the data.
It would therefore be highly appreciated if MoMA adds more data to their datasets in future. At least the information that is already published on their online web collection should be added, here security or ethical concerns were obviously already ruled out

already by the museum. This additional data would already allow for deeper investigations. Additionally, and this is even more pressing and important, enriched documentation on the datasets should be provided by MoMA, as this is a crucial part of all responsible work with data. Nevertheless, MoMA's decision to publish their datasets should be celebrated. With this step, they invite the public virtually into their storage rooms and their filing system. They make transparent what others still like to keep private. And by publishing their datasets, the museum has taken on a role model function, which some museums have already followed and hopefully many others will do too.

The insights found might be obvious to the museum and its staff; they are experts who might have intuitions about for example the gender distributions and how the different directors collected. However, with the help of data analysis, it is now also possible for outsiders to retrieve reliable numbers for those intuitions and assumptions based on the museum records. This forms also possible practical implications of this thesis, it might motivate MoMA and other museums to revisit their way of cataloguing, to acquire specific missing data, and to address some of the limitations that were mentioned. It might also stimulate movement in the general discussion on how cultural heritage objects, particularly art objects, could be described without superimposing the Western view and way of thinking over all other approaches. The purpose of this thesis which sought to contribute knowledge on how to work with datasets of cultural heritage collections and where caution is necessary, was fulfilled. It was possible to demonstrate that there is software that can be used and that there is interesting information in collection datasets that should be excavated to broaden the scope of research questions.

In conclusion, the results of this thesis shed vivid light on the power of data analysis methods in deriving in-depth insights from cultural heritage collection datasets. The study shows the great potential that lies in the distant reading approach by successfully introducing new possible research directions or ideas. It demonstrates where biases are to be expected and where Western-oriented understandings and structures should be challenged and critically investigated. The thesis invites future researchers and scholars but also museum stuff to explore datasets and encourages open minded approaches to explore alternative avenues and directions when analyzing collections. This thesis heralds them to ponder the implications and to embrace the opportunities at hand. Ultimately, this work motives to research cultural heritage collection with the method of data analysis to gain new insights on the collections and importantly also about the way how knowledge about objects is organized and managed.

5 Literature

Bruce *Altshuler* ed., Collecting the New: Museums and Contemporary Art (Princeton, New Jersey 2007).

Jan *Assmann*, Das kulturelle Gedächtnis: Schrift, Erinnerung und politische Identität in frühen Hochkulturen. 2., Durchges. Aufl. (München 1997).

Charlotte *Barat*, Darby *English*, The Artist Wasn't Present: On MoMA's Fumbled First Showing of Black American Art. ARTnews, 07/17/2019, online at <https://www.artnews.com/art-news/news/among-others-blackness-at-moma-excerpt-12972/>.

Matthew *Battles*, Michael *Maizels*, Collections and/of Data: Art History and the Art Museum in the DH Mode Chapter Author(s): MATTHEW BATTLES and MICHAEL MAIZELS. In: Debates in the Digital Humanities 2016, edited by Matthew K. Gold, Lauren F. Klein, 325–344 (2016), doi:10.5749/j.ctt1cn6thb.

Tony *Bennett*, The Birth of the Museum: History, Theory, Politics Culture : Policies and Politics (London ; New York 1995).

Michael *Berthold*, KNIME The Konstanz Information Miner. Java (Zürich 2004), online at <https://www.knime.com/>.

Blagoy *Blagoev*, Sebastian *Felten*, Rebecca *Kahn*, The Career of a Catalogue: Organizational Memory, Materiality and the Dual Nature of the Past at the British Museum (1970–Today). Organization Studies 39, no. 12 (12/2018) (12/2018) 1757–1783, doi:10.1177/0170840618789189.

Christian *Chabot*, Chris *Stolte*, Andrew *Beers*, Pat *Hanrahan*, Tableau (2003).

Patricia *Cohen*, Family's Claim Against MoMA Hinges on Dates. The New York Times, 08/23/2011.

Catherine Nicole *Coleman*, Managing Bias When Library Collections Become Data. International Journal of Librarianship 5, no. 1 (07/23/2020) (07/23/2020) 8–19, doi:10.23974/ijol.2020.vol5.1.162.

Fernando *Domínguez Rubio*, Preserving the Unpreservable: Docile and Unruly Objects at MoMA. Theory and Society 43, no. 6 (11/2014) (11/2014) 617–645, doi:10.1007/s11186-014-9233-4.

Johanna *Drucker*, Graphical Approaches to the Digital Humanities. In: A New Companion to Digital Humanities, edited by Susan Schreibman, Ray

Siemens, John Unsworth, 238–250 (Chichester, UK 2015), doi:10.1002/9781118680605.ch17.

———, The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Scholarship. 1st ed. (First edition. | Abingdon, Oxon ; New York : Routledge/Taylor & Francis, 2021. 2021), doi:10.4324/9781003106531.

Gabrielle *Foreman*, Labanya *Mookerjee*, Computing in the Dark: Spreadsheets, Data Collection and DH's Racist Inheritance. Always Already Compuational: Collections as Data, 2019, 108–109.

GallerySystems, GallerySystems. TMS Collections Guide. Gallerysystems.Com, 06/17/2023, online at <https://ideas.gallerysystems.com/rs/962-HZY-660/images/TMS%20Collections_web.pdf>.

Haidy *Geismar*, Museum Object Lessons for the Digital Age (2018), doi:10.14324/111.9781787352810.

Lisa *Gitelman*, Virginia *Jackson*, Introduction. In: "Raw Data" Is an Oxymoron, edited by Lisa Gitelman, 1–14 Infrastructures Series (Cambridge, Massachusetts ; London, England 2013).

Ernst *Gombrich*, The Story of Art (London 1979).

Antony *Griffiths*, Collections Online: The Experience of the British Museum. Master Drawings 48, no. 3 (2010) (2010) 356–367.

Sabine *Haag*, Veronika *Sandbichler*, *Kunsthistorisches Museum Wien*, *Schloss Ambras Innsbruck* eds., Ferdinand II: 450 Jahre Tiroler Landesfürst: Jubiläumsausstellung: eine Ausstellung des Kunsthistorischen Museums Wien in Kooperation mit der tschechischen Nationalgalerie und dem Institut für Kunstgeschichte der Akademie der Wissenschaften der Tschechischen Republik, 15. Juni bis 8. Oktober 2017 (Innsbruck 2017).

Rebecca *Kahn*, Laura *Gibson*, Digital Museums in the 21st Century: Global Microphones or Universal Mufflers? Museological Review, no. 20 (2016) (2016) 39–51.

Rebecca *Kahn*, Rainer *Simon*, Feast and Famine: The Problem of Sources for Linked Data Creation. In: Graph Technologies in the Humanities - Proceedings, 86–100, 2020.

Sybil *Kantor*, Alfred H. Barr, Jr., and the Intellectual Origins of the Museum of Modern Art, 2002.

Tiffany-Quan *Le*, MoMA and Nazi-Era Art Restitution: Contexts and Thoughts for the Future (Master Thesis Concordia University 2017), online at <https://spectrum.library.concordia.ca/id/eprint/982899/1/Le_MA_F2017.pdf>.

Alexander *Loth*, Datenvisualisierung mit Tableau. 1. Auflage. (Frechen 2018).

Glenn D. *Lowry*, Introduction. In: MOMA Highlights: 350 Works from the Museum of Modern Art, New York, edited by Museum of Modern Art (New York, N.Y.), Harriet Schoenholz Bee, Cassandra Heliczer, 2nd ed., 16–21 (New York 2004).

Sharon *Macdonald*, Collecting Practices. In: A Companion to Museum Studies, 81–97 Blackwell Companions in Cultural Studies (Malden, MA 2006).

André *Malraux*, Das imaginäre Museum (Frankfurt New York 1987).

Rhiannon *Mason*, Cultural Theory and Museum Studies. In: A Companion to Museum Studies, edited by Sharon Macdonald, 17–32 Blackwell Companions in Cultural Studies 12 (Malden, MA 2006).

MoMA, MoMA Collection - Automatic Monthly Update (11/01/2022), doi:10.5281/ZENODO.7269353.

Sabra *Moore*, Openings: A Memoir from the Women's Art Movement, New York City 1970-1992. First edition. (New York, NY 2016).

Franco *Moretti*, Conjectures on World Literatur. New Left Review 1, no. 1 (2000) (2000) 54–68.

S. *Münster*, F. I. *Apollonio*, P. *Bell*, P. *Kuroczynski*, I. *Di Lenardo*, F. *Rinaudo*, R. *Tamborrino*, Digital Cultural Heritage Meets Digital Humanities. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W15 (08/23/2019) (08/23/2019) 813–820, doi:10.5194/isprs-archives-XLII-2-W15-813-2019.

Tamara *Munzner*, Visualization Analysis and Design A.K. Peters Visualization Series (Boca Raton 2015).

Daniel G. *Murray*, Tableau Your Data! Fast and Easy Visual Analysis with Tableau Software®. Second edition. (Indianapolis, Indiana 2016).

Museum of Modern Art, Collections Management Policy (04/20/2020), online at <https://www.moma.org/momaorg/shared/pdfs/docs/about/Collections-Management-Policy-2020-04-20.pdf>.

Museum of Modern Art (New York, N.Y.), Charlotte *Barat*, Darby *English*, Mabel *Wilson*, Glenn D. *Lowry* eds., Among Others: Blackness at MoMA (New York 2019).

Jeffrey K *Olick*, Collective Memory: The Two Cultures. American Sociological Association 17, no. 3 (1999) (1999) 333–348.

Thomas *Padilla*, On a Collections as Data Imperative, 2017.

———, Responsible Operations: Data Science, Machine Learning, and AI in Libraries, doi:10.25333/W8SG-8440, (05/05/2023).

Thomas *Padilla*, Laurie *Allen*, Hannah *Frost*, Sarah *Potvin*, Elizabeth Russey *Roke*, Stewart *Varner*, Always Already Computational: Collections as Data, 2019.

Erwin *Panofsky*, Early Netherlandish Painting. 9. [print]. The Charles Eliot Norton Lectures 1947–48 (New York 1987).

Erwin *Panofsky*, Irving *Lavin*, William S. *Heckscher*, Three Essays on Style (Cambridge, Mass 1995).

Daniel *Rosenberg*, Data Before the Fact. In: “Raw Data” Is an Oxymoron, edited by Lisa Gitelman, 16–40 Infrastructures Series (Cambridge, Massachusetts ; London, England 2013).

Leslie F. *Sikos*, Dean *Philp*, Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs. Data Science and Engineering 5, no. 3 (09/2020) (09/2020) 293–316, doi:10.1007/s41019-020-00118-0.

Zachary *Small*, MoMA Survived Ten Weeks of Protest. But Inside the Museum, Some Employees Are Feeling the Strain. A Protest Movement Questioning the MoMA Board’s Ties to “Toxic Philanthropy” Came in the Midst of a Staffing Crisis. Artnet News, 07/19/2021, online at <https://news.artnet.com/art-world/moma-survived-ten-weeks-protest-strike-moma-1990049>.

William S. *Smith*, Dissident Modernism Meets Peak Philanthropy at the New MoMA. Art in America, 10/25/2019, online at <https://www.artnews.com/art-in-america/features/moma-reopens-modern-art-politics-protests-63665/>.

Matthew *Stanley*, Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations. In: “Raw Data” Is an Oxymoron, edited by Lisa Gitelman, 77–88 Infrastructures Series (Cambridge, Massachusetts ; London, England 2013).

John Wilder *Tukey*, Exploratory Data Analysis Addison-Wesley Series in Behavioural Science (Reading, Mass 1977).

Kirk *Varnedoe*, Introduction. In: Modern Contemporary: Art since 1980 at MoMA, edited by Kirk Varnedoe, Paola Antonelli, Joshua Siegel, Museum of Modern Art (New York, N.Y.), 2nd ed., 11–15 (New York 2004).

Elaine *Velie*, Protesters Crash MoMA Gala Over Board Chair's Fossil Fuel Ties. Climate Activists Are Asking the Museum to Remove Board Chair Marie-Josée Kravis, Whose Husband's Private Equity Firm Has Invested Billions in Oil and Gas Projects. Hyperallergic, 06/06/2023, online at <https://hyperallergic.com/826458/protesters-crash-moma-gala-over-board-chairs-fossil-fuel-ties/>.

Isabel *Vincent*, These Famous Artworks Were Looted by Nazis — and Are on Display at Met, MoMA. New York Post, 08/22/2022, online at <https://nypost.com/2022/08/22/new-law-requires-new-york-museums-to-label-nazi-looted-works/>.

Florian *Windhager*, Paolo *Federico*, Gunther *Schreder*, Katrin *Glinka*, Marian *Dork*, Silvia *Miksch*, Eva *Mayr*, Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. IEEE Transactions on Visualization and Computer Graphics 25, no. 6 (06/01/2019) (06/01/2019) 2311–2330, doi:10.1109/TVCG.2018.2830759.

With New Lawsuits Against MoMA and the Santa Barbara Museum of Art, the Heirs of a Holocaust Victim Are Seeking to Reclaim a Pair of Schieles. The Pieces Were Once Owned by Austrian Jewish Performer Fritz Grünbaum. Artnet News, 12/21/2022, online at <https://news.artnet.com/art-world/with-new-lawsuits-against-moma-and-the-santa-barbara-museum-of-art-the-heirs-of-a-holocaust-victim-are-seeking-to-reclaim-a-pair-of-schieles-2234437>.

Christopher *York*, Exploratory Data Analysis for the Digital Humanities: The Comédie-Française Registers Project Analytics Tool. English Studies 98, no. 5 (07/04/2017) (07/04/2017) 459–482, doi:10.1080/0013838X.2017.1332024.

Sandra *Zalman*, Unpacking the MoMA Myth: Modernism under Revision. Modernism/Modernity 29, no. 2 (04/2022) (04/2022) 283–306, doi:10.1353/mod.2022.0009.

Chiara *Zuanni*, Theorizing Born Digital Objects: Museums and Contemporary Materialities. Museum and Society 19, no. 2 (07/30/2021) (07/30/2021) 184–198, doi:10.29311/mas.v19i2.3790.

6 Image References

7 Zusammenfassung

Diese Arbeit zeigt, dass die Erforschung von Kulturerbesammlungen, insbesondere Kunstsammlungen, durch Datenanalysemethoden neue Erkenntnisse zutage bringt, Zusammenhänge aufdeckt und Lücken und Einschränkungen aufzeigt. Die Studie nutzt zwei öffentlich zugängliche Datensätze des Museum of Modern Art in New York (MoMA) und zeigt die Ergiebigkeit dieser Methode bei der Erforschung von größeren Kunstsammlungen auf. Dabei wird nicht nur ein tieferes Verständnis der Sammlungen und der darin enthaltenen Kunstwerke aufgezeigt, sondern auch interne Mechanismen der Institution und deren Art Objekte und das vorhandene Wissen über sie zu katalogisiert und zu organisiert veranschaulicht. Die Arbeit befasst sich mit der Bedeutung der Entscheidung des MoMA, Teile seiner Datenbank zu veröffentlichen, die es Forscher*innen ermöglicht, die Sammlung des Museums unabhängig zu untersuchen und neue Erkenntnisse zu gewinnen. Damit leistet diese Arbeit einen Beitrag für alle Disziplinen, die sich mit Kulturerbesammlungen auseinandersetzen, insbesondere Kunstgeschichte, Collection Studies und Digital Humanities.