Kristin Kaltenhäuser, Tijs Slaats, Thomas Gammeltoft-Hansen, Naja Holten Møller (2022): Deconstructing Gender in Asylum Categories: An Archival Perspective on a Practice with Limited Access. In: Proceedings of the 20th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Notes, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.48340/ecscw2022_n03

Deconstructing Gender in Asylum Categories: An Archival Perspective on a Practice with Limited Access

Kristin Kaltenhäuser, Tijs Slaats, Thomas Gammeltoft-Hansen, Naja Holten Møller
University of Copenhagen
Contact Authors: krka@di.ku.dk, slaats@di.ku.dk, tgh@jur.ku.dk, naja@di.ku.dk

Abstract. Public authorities make decisions that greatly impact both citizens and non-citizens. Decision-making on asylum, which is regulated by international law but administered by states, in particular is characterised by a higher level of secrecy than other public services. The 1951 Refugee Convention defines refugeehood as the fear of being persecuted for reasons of race, religion, nationality, social group, or political opinion. Although fear of gender-related persecution was not included as one of the grounds meriting asylum, state practice means that it is today generally recognised as such. The United Nations Refugee Agency (UNHCR) recommends that states "ensure a gender-sensitive interpretation of the 1951 Refugee Convention." Using natural language processing (NLP) to analyse an open dataset of Danish asylum case summaries, we first identify five empirical categories connected to gender in the case summaries: 1) gender-related persecution, 2) LGBT 3) sexual conditions, 4) marital conditions and 5) other gender-related forms of persecution. Secondly, we illustrate the relationship between these gender-related categories and other categories/topics in asylum motives. Finally, we discuss how data science techniques can be applied to better understand complex, cooperative work practices in an area where access for researchers is limited, but archival data is available.

1 Introduction

Asylum decision-making generates unique data about casework, because it reflects how national and international law is put into practice. In this domain, casework is centered around an interview process aimed at assembling the required documentation to determine refugee status (Nielsen and Møller (2022)). In this paper we present findings from a study of gender-related categories in asylum case data. Gender-related persecution was addressed relatively late by international law compared to other categories of asylum motives (e.g. persecution for race, religion, nationality, social group, or political opinion). Even when it was formally introduced in 2002 by a UN guideline, gender continued to have a somewhat implicit status. It remains up to states to interpret and apply it (Byrne and Gammeltoft-Hansen (2020)). Our motivation for this study is to deconstruct the formal UN category of gender-related persecution and study its empirical categories. Using data science methods and applying an archival perspective (Thylstrup et al. (2021)), we investigate gender-related categories in an open dataset of 9,075 asylum case summaries handled by the Danish Refugee Appeals Board. We take an archival perspective in this study by tracing the origin of data structures and categories and raising questions about the power structures that shape them, in our case International law, national decision-making and archiving practice.

Asylum decision-making is a highly politicised and securitised area (Bigo (2014)) for which it is hard to negotiate access for observational studies. National authorities are often concerned for the applicants' safety and at the same time seek to avoid unwanted critique. That means that decision-making in this area is, from an access standpoint, much less public as compared to other types of administrative decision-making. Seaver studies a domain with similarly limited access, algorithms in an environment of corporate secrecy (Seaver (2017)), and introduces the term *scavenge*, to describe their research tactic. Scavenging denotes tracing cultural practices across heterogeneous locations to empirically characterise an inaccessible object of work (algorithms) without directly addressing it. Seaver emphasizes how an exploration across multiple locations can provide a better understanding of the persistent context of a practice and its work objects. The entry point for our scavenging is the Danish archive (database) that contains information about asylum-decision making (see Figure 1).

Following this line of argumentation, we ask: How are gender-related categories presented in the asylum case summaries and how can they be probed to understand the practice of asylum decision-making?

The study applies data science and natural language processing (NLP) techniques to conduct a category and topic analysis of asylum case summaries. Applying data science methods with an archival perspective is performative in the sense that it can bring about a new understanding of, in our case, gender-related categorisation practices in asylum. Our first contribution is identifying five empirical categories connected with gender in the asylum motives of the case

summaries: 1) gender-related persecution, 2) LGBT 3) sexual conditions, 4) marital conditions and 5) other gender-related persecution[1]. Second, we illustrate the relationship between these gender-related categories and other asylum categories/topics, while showing how data science techniques used on archival datasets can serve as an entry point for studying practices in a context where access to the work domain is limited. We examine the empirical categories archived as part of a collaboration between the Danish national authorities and the Danish Refugee Appeals Board. When data is archived, power differences are present between the archiver who decides what is to be remembered and the individuals the records in the archive are about. While practices of categorisation and classification are a long-term interest of Computer-Supported Cooperative Work (CSCW) (Suchman (1993), Bowker and Star (1999), Møller and Bjørn (2011), Boyd and Crawford (2012), Pine and Liboiron (2015)), an archival perspective increases understanding of the contexts and power relationships that structure datasets. This can reveal something about underlying practices, in addition to observational studies which is a core strategy for understanding practices in CSCW (Randall et al. (2007)).

Engaging with a complex, collaborative work domain such as asylum, we argue, opens CSCW's long-standing interest in public sector decision-making for renewed considerations of how we as a research community can ensure that highly securitised areas of work are included in this strand of research. We as researchers can approach these sensitive areas with care for all stakeholders and act as intermediaries between interests of stakeholders with differential power, by increasing communication between the parties and making the power landscape visible.

2 Gender-related Persecution in International Asylum Law

While international law - notably the 1951 Refugee Convention - establishes a formal definition of what it means to be a refugee, states adapt and transform this category through both their national law and decision-making. The process of deciding asylum claims on a national level plays a major role in developing international refugee law. The lack of any dedicated court for refugees at the international level implies that states are the principal interpreters of international law in this area (Byrne and Gammeltoft-Hansen (2020)). The 1951 Refugee Convention is silent in terms of how states should design their asylum procedures. States implement diverse institutional and procedural frameworks for asylum decision-making, resulting in loosely coordinated, divergent practices across

[1] We note that the naming of the category *LGBT* in the dataset as a solitary initialism, lacking further description, is in itself questionable, since the asylum motive could in this case be read as *LGBT*, instead of *LGBT persecution*. Since part of the contribution of this study is the identification of empirical categories, we use the same naming, but raise this concern.

states. States maintain national databases of asylum decisions, establishing their own legal practices and categories on the basis of already decided asylum cases. In the case of Denmark, part of this database has been made publicly available. We analyse gender-related empirical categories in the public case data as a starting point for understanding the decision-making practices in this area.

The 1951 UN Refugee Convention describes a refugee as any person who:

> owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country (Article 1 A (2) Refugee Convention, UNHCR (1951))

Unlike some later human rights conventions, gender or sexual orientation is not a category explicitly listed in this formal definition. Since the 1980s, national asylum authorities have gradually come to recognise that gender may form the basis of an asylum claim; for instance, in some societies women can face repercussions for transgressing social mores, and sexual minorities can face widespread persecution. This position is backed by UNHCR guidelines, for example the 2002 guideline on gender-related persecution, which while not legally binding recommends states to "ensure a gender-sensitive interpretation of the 1951 Refugee Convention" as part of their national asylum procedures (UNHCR (2002)).

The UNHCR guidelines elaborate upon and specify the types of cases which belong within the category of gender-related persecution. As such, the UNHCR guidelines may be read as themselves constructing a category for seeking asylum, covering topics such as homosexuality, female genital mutilation, and so on. In the ensuing analysis, we systematically extract these topics, query the case summaries on this basis, and analyse their occurrence in the empirical categories applied by the Danish Refugee Appeals Board. The goal is to investigate the transformation of the formal category of gender-related persecution and its constituting topics when adopted in the national practice of asylum decision-making and reveal the topics that are omitted or articulated in the empirical categories that emerge in the Danish archive of asylum case summaries.

Gender has an implicit status in national asylum procedures, as it is not explicitly listed in the definition of refugeehood. According to the Convention, having fled one's country due to fear of persecution related to gender is not enough; one must have fled for one of five listed categories:

- race
- religion
- nationality
- membership of a particular social group
- political opinion

These categories are commonly referred to as *Convention grounds* in asylum law (see e.g. UNHCR (2002)) and provide an exhaustive list when deciding whether to grant a person refugee status based on the 1951 Refugee Convention.

For policy-makers to fit gender-related persecution into this framework, the introduction of a category hierarchy is required. Most commonly, gender is subsumed as a subcategory of *particular social group*, which has proven to be the most flexible category for types of asylum claims not originally foreseen when the Refugee Convention was drafted in 1951 (Hathaway and Foster (2003)). Legally, this subsumption is based on the argument that women and sexual minorities share the same type of core characteristics as regularly used to define other social groups. According to UNHCR:

> It follows that sex can properly be within the ambit of the social group category, with women being a clear example of a social subset defined by innate and immutable characteristics, and who are frequently treated differently than men. Their characteristics also identify them as a group in society, subjecting them to different treatment and standards in some countries. Equally, this definition would encompass homosexuals, transsexuals, or transvestites. (UNHCR (2002))

Historically, this type of subsumption has equally led to categorisation struggles, with some asylum lawyers questioning whether *a particular social group* can meaningfully be broadened to include half the world's population. Even today, categorisation struggles continue within this area as states apply widely different approaches to defining the contours of who is at risk (Hathaway and Foster (2014)). As Spijkerboer notes, female asylum applicants thus tend to be constructed as a "double other" by Western countries, simultaneously singled out based on their gender and non-Western background (Spijkerboer (2000)). More fundamentally, the recognition of gender-related claims in this manner reproduces a historical category hierarchy, in which gender has to be constructed and defined by reference to the broader category of *membership of a particular social group* or other Convention grounds. As we will show, this has implications for how caseworkers subsequently approach and label asylum applicants fleeing gender-based persecution. In the ensuing analysis, we investigate the connection between gender and the formal list of Convention grounds, and how this category hierarchy is applied in Denmark.

3 Related Work

An archival perspective can support the rebalancing of datasets in favour of people that are displaced, prior research in archival studies suggest. For example Gilliland (2017) finds that official records and archives of states do not address the existential needs for documentation of non-citizens, such as asylum seekers, who

seek protection in a country, but do not enjoy the same rights as citizens. Cakici et al. (2020) investigate the relationship between citizenship and the data practices that shape how people can make claims to protection. Claiming data that is considered credible by decision-makers is not a straightforward process (Nielsen and Møller (2022)). We thus turn our attention towards related work on archival data and the associated knowledge work as arenas of power (Star and Strauss (2004)).

3.1 Archived Data as an Arena of Power

Increasingly, data in asylum decision-making is saved to databases (both national and international) that come to form an archive, a long-standing interest of CSCW (Ackerman and Malone (1990), Bannon and Kuutti (1996), Ackerman et al. (2013)). Recent research co-created, together with volunteers of social clinics, a digital archive of oral histories to generate a counter narrative to the official records of the healthcare system in Greece, serving both asylum seekers and other vulnerable groups (Vlachokyriakos et al. (2021)). The archive, as we learn from this research, may not simply be left alone for us to engage with it years later, but can play an active part in shaping practices. This counters a dominant understanding of the archive as a mere historical database and window into the past (Thylstrup (2022)).

Following Foucault, an archive can be an empirical site functioning as an administrative tool for data production about e.g. populations and nations, but also an analytical concept or a lens (Foucault (1972)). Power from this perspective is *generative* and understands communities as subjects through data, not simply as subjected to a power-over reality imposed onto them. Taking an archival perspective means paying close attention to various forms of power structures that shape archives (D'Ignazio and Klein (2020)) and thus become scrutable (Thylstrup (2022)). Amelia Acker in Thylstrup et al. (2021) suggests that investigating hidden origins of database structures (archives) and the motivation behind their implementation can provide insights into the power of cultural practices. She contends that categories of archives are continuously changing and subject to political and social change.

The archivists of the asylum decision-making process are the caseworkers (knowledge workers) who establish credibility, negotiate a status with a displaced person and construct them as an asylum seeker. Recognizing that categories have politics, as they are a fundamental device by which societies constitute their social order (Suchman (1993)), this form of knowledge production is a question of power, as specific categories are assigned to a person applying for asylum.

In a CSCW-context, we study practices of categorisation that come to form the origins of database structures (Møller and Bjørn (2011)). However, not all work practices are readily available for us to study in CSCW. The sensitive and highly politicised area of work in asylum decision-making is an example of a domain with limited access, where archived data can serve as en entry point for further study.

3.2 Data Shadows in Archives

Archived data casts *shadows* on topics that are omitted in the categories that make up the database structure and function as a gatekeeper for later access. People that are displaced rely on documentation and public records to establish credibility in order to be categorised as *persecuted* (Nielsen and Møller (2022)). However, these records are dispersed over several national and international archives, and people that are displaced often resort to "irregular forms and uses of records" to claim refugee status, e.g. washing out stamps in their passports that would make it impossible to enter certain countries or transmitting photographs of personal documents via mobile phones (Gilliland (2017)). Following Gilliland, publicly available and internationally accessible archives can play a critical role to address the needs of the displaced. In this sense, the public archive of asylum case summaries, we are analysing, may serve as a resource for asylum seekers and their advocates to navigate the Danish asylum decision-making process.

Data shadows are one example of an analytical concept that can make power differentials in datasets visible (Møller et al. (2021)). They appear when the realities of an actor with less power are not captured in a dataset. By inspecting data produced by asylum casework for data shadows instead of uncritically using it as ground truth for further processing (Muller et al. (2021), Aragon et al. (2022)), we make omitted and articulated topics visible and contestable.

Little research in CSCW has investigated how scholars can consider the relation between communities and the data is produced by and about them, to enable the construction of their own interpretations, as opposed to data technologies imposing a reality onto them (Irani et al. (2010), Taylor et al. (2015)). Caselli et al. (2021) argue for a participatory approach to NLP in which *language* is not considered as data, but as the product of people and advocate for involving producers of text in its sensemaking process (Ibid). We position our research in relation to debates in CSCW on changing the discourse around databased technologies that address how we respond to global connectivity and mobility (Irani et al. (2010)). Combining data science methods with an archival perspective allows us to scrutinize both the omitted and articulated gender-related categories in an archive of asylum cases, to study power structures that shape the national asylum decision-making process in Denmark.

4 Topic Analysis of Asylum Case Summaries

The dataset and archive we investigate makes summaries of asylum cases available for query on the website of the Danish Refugee Appeals Board. Figure 1 shows how categories are listed with check boxes so the user can search for entries associated with that category. In this case the categories take on the role of a gatekeeper to the archive. Topics that are articulated in the categories are easily accessible, while others that are not, require explicit searches or are otherwise forgotten.

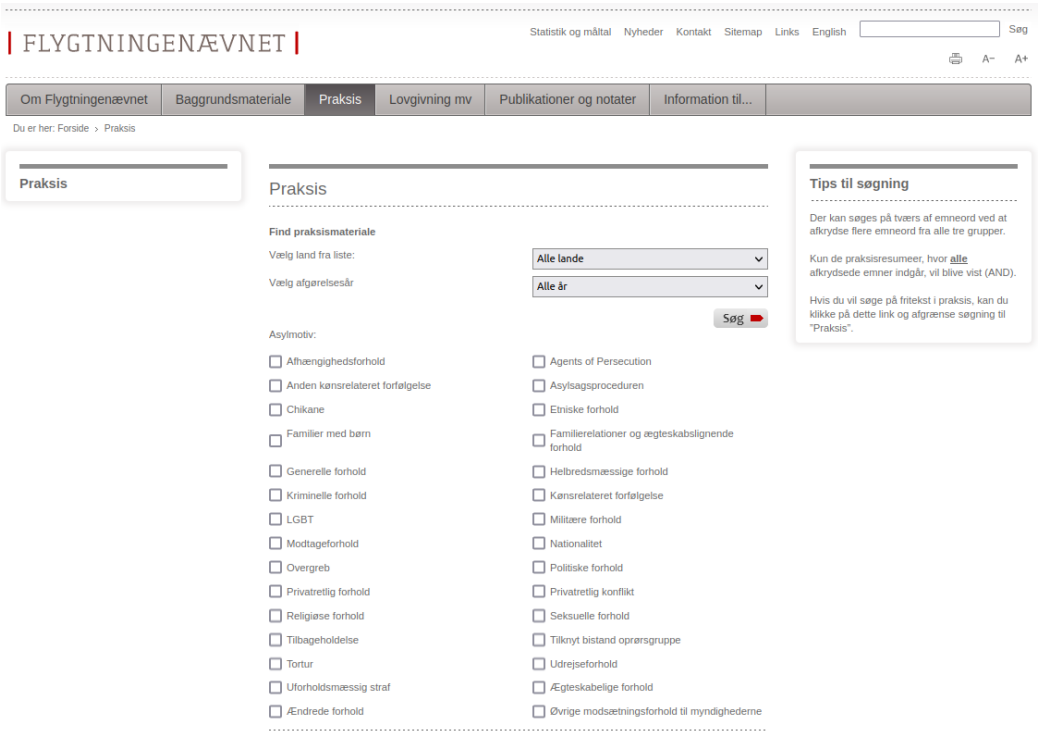

Figure 1. This Figure shows the interface to access the public dataset on the website of the Danish Refugee Appeals Board - https://fln.dk/praksis and the predefined categories that can be chosen to query the data.

4.1 Data Extraction and Analysis

Using techniques from NLP and a qualitative analysis of judicial text, we combine several analysis methods in this study. The data was scraped using the Python libraries *beautifulsoup4* and *Selenium* on the 19. October 2021 and yielded 9,075 case summaries. We performed the analysis of the category of gender-related persecution in the following steps:

1. After an initial exploratory analysis of asylum motives in the data and an investigation of the UNHCR definition of gender-related persecution, we matched and selected the categories in the Danish dataset with that definition for further analysis. The following are the five empirical categories:

    - gender-related persecution
    - LGBT
    - sexual conditions
    - marital conditions
    - other gender-related persecution

    Following the UNHCR guidelines, claims based on sexual orientation contain a gender element. We therefore add the categories *LGBT* and *sexual*

conditions to our analysis. In an initial investigation, we also included the category of *abuse* as gender-related. After a qualitative study of a sample of the related case summaries, however, we discarded it. In this context, abuse mostly pertained to political and military contexts. For this study, we translated the categories into English. It is also important to note that in the Danish language, the same word is used for *sex* and *gender*.

2. We then analysed the other asylum motives with which gender-related categories are tagged. We first assessed if there are cases that only received a gender-related category and then calculated a co-occurrence matrix, as displayed in Table III. A co-occurrence matrix is a tool often used in NLP and image analysis. It represents the number of times terms appear in the same context with other terms (or pixels in the case of image analysis). In our case, we created a table in which each column is assigned to an empirical category pertaining to gender-related persecution in the Danish dataset. Each row represents one of the 29 asylum motive categories used by the Refugee Appeals Board.

   The number in each cell of the table represents the number of cases to which a gender-related category is applied, together with another asylum motive category. The shading of the cell indicates the magnitude of the proportion of the number compared to the other categories that were tagged with it. We applied the shading for clarity to illustrate the composition of the co-tagged categories in the table.

3. We then systematically extracted topics from the UNHCR guidelines on gender-related persecution that are explicitly stated as claims related to gender. Table I provides a list of the extracted topics in the first column. We arrived at the list topics by identifying themes stated as examples relevant to gender-related persecution in the UN guideline. We decided not to include more general themes such as "transgression of social mores" (UNHCR (2002)), since it can include several topics, such as abortion, extramarital affairs, which sometimes overlap with other topics, which we included.

4. Finally, we conducted a topic analysis by querying the case summaries for words related to the extracted topics, see Table IV. For example, to detect the topic of *homosexuality*, we query the cases for the terms *lesbisk* and *homoseks*. To get from the UN topics to the Danish search terms that are applied in the case summaries, we performed a qualitative analysis on a sample of the cases (n=30) to select the keywords corresponding to the UN topics. This analysis showed a consistent use of the Danish terms as shown in Table I, except for transvestism, for which no cases were found. We employed manual stemming of the query words to reduce words to their basic form or stem. This technique is often applied in NLP to account for different forms of words in a text. Each case that contains any of these terms is counted as one occurrence. That means a case can contain several topics.

We then calculated the percentage composition of each category to illustrate which topics constitute a category.

Topic	**Queried terms**
Homosexuality	homoseks, lesbisk
Human trafficking	menneskehandel, traffick
Female genital mutilation	omskåret, omskæring
Forced marriage	tvangsgift
Forced prostitution	prostitution
Rape	voldt
Transgenderism	transkøn
Forced abortion	abort
Forced sterilisation	sterilis
Bisexuality	biseks
Transvestism	transvest

Table I. List of topics extracted from the 2002 UN guideline on gender-related persecution that constitute the category. The second column lists the terms that were used in this study to query the Danish case summaries for the topics. The selection of the query words is based on a translation into Danish of the terms used in the UN guideline.

To guarantee reproducibility and to provide the research community with a tool for comparing new methods with the ones evaluated in this study, we provide the code used in this study, as well as the sense-making process on GitHub: *https://github.com/KristinKalt/ecscw2022_dk_asylum_analysis*.

4.2 Characteristics of the Data

The object of this analysis is the publicly available dataset by the Danish Refugee Appeals Board (*https://fln.dk/praksis*). The data consists of summaries of asylum decisions by the Refugee Appeals Board in Denmark between 2004 and 2021. The Refugee Appeals Board is the second institution to assess applications for asylum in Denmark. Only cases rejected by the Danish Immigration Service, the first instance in the process, are automatically referred to the Refugee Appeals Board.

We conducted the analysis on a subset of appeals cases, namely the cases that are publicly available, for which the selection strategy is unknown. The publication follows an *Executive order on rules of procedure of the Refugee Appeals Board* (Justitsministeriet, Denmark (2016)), which states that the secretariat of the Refugee Appeals Board "regularly updates the board's website www.fln.dk with, among other things, summaries of the board's decisions, the board's background material and other relevant information about the work of the Refugee Appeals Board." (translated by the first author). This guideline is rather broad and does not specify the criteria of cases that are selected to be made public. Therefore the data is not representative of the full set of Danish asylum seeking

cases. From a statistical perspective, we cannot make valid general statements about Danish asylum cases of the scraped data. However, an analysis of the origins of the applied categories provides insights into the archival practices in the Danish asylum system.

Following the data conceptualisation strategy described by Kitchin (2014), we describe the data as follows: The dataset is semi-structured. It consists of an unstructured qualitative part composed of free text summaries. It is supplemented by structured and quantitative metadata comprised of three attributes: decision year, country of origin, and asylum motive. Accordingly, while the attributes *decision year* and *country of origin* have single discrete values, the attribute *asylum motive* can be assigned multiple categories. These types of overlapping categories are also referred to as *tags* in data science.

The categories of the attribute *asylum motive* are collectively exhaustive in the sense that each case fits into at least one given category. To make this system work, residual categories such as *other* are often needed to describe cases that do not fit in the carefully designed boxes of the classification scheme (Bowker and Star (1999)). The Danish dataset utilises residual categories as well, such as *other gender-related persecution* or *other conflicts with authorities*.

As the data points are summaries of asylum cases, the data is derived from other data traces of the process of an asylum application in Denmark. The authors are using this data as secondary data, which according to Kitchin (2014) is data made available for reuse by someone other than the people who generated it and for a different purpose. The data is anonymised by substituting names with letters or a broad descriptive noun (e.g. boyfriend, agent) and is therefore considered attribute data, which represents an aspect of the phenomenon of asylum applications in Denmark, but is not uniquely identifiable (Kitchin (2014)).

5 Findings

	Cases	% of total cases
Gender-related persecution	303	3.3
LGBT	211	2.3
Sexual conditions	366	4.0
Marital conditions	474	5.2
Other gender-related persecution	175	1.9

Table II. This table illustrates the distribution of case summaries among the empirical gender-related categories in the dataset of the Danish Refugee Appeals Board.

In what follows, we illustrate how we make inferences about asylum decision-making and archiving practices. Of the 9,075 extracted cases, 1,876 are tagged with at least one of the five categories that are the subject of this analysis. 20% of the total cases have an asylum motive that falls under a category related to gender (see

Table II). An overall look at the distribution of case summaries among the empirical gender-related categories in the case summaries demonstrates that gender is indeed an asylum motive that we need to take into consideration as we work to understand asylum decision-making practices. Our findings show that 1) gender in asylum case data is a co-dependent category, and 2) that some topics related to gender are omitted in the asylum motive categories, while others are articulated.

5.1 Gender-related Persecution as Co-dependent Category

A first observation and finding of our data analysis is that gender appears to be a co-dependent category. We found that none of the empirical categories on gender-related persecution are applied as a standalone motive for asylum in the Danish data. They always occur together with at least one other category, such as e.g. *religious conditions*. We see in Table III how gender-related categories and the other asylum motives are interconnected. The table illustrates the categories that have been applied together in individual cases. The cells with darker shading highlight categories that are tagged together the most. We find that especially the categories of *ethnic*, *political* and *religious conditions*, as well as the more general categories of *private law matters* and *general conditions* occur together with the categories of gender-related persecution.

	Gender-related persecution	LGBT	Sexual conditions	Marital conditions	Other gender-related persecution	Total
Private law matters	116	37	102	171	39	465
Agents of Persecution	79	14	71	89	63	316
General conditions	53	15	46	67	42	223
Religious conditions	20	21	51	57	4	153
Political conditions	32	17	35	29	6	119
Abuse	21	12	41	21	4	99
Ethnic conditions	29	4	23	17	7	80
Harassment	26	13	19	12	1	71
Criminal conditions	10	3	15	7	2	37
Private law conflict	6	1	7	20	3	37
Military conditions	5	2	9	6	1	23
Nationality	4	4	2	10	1	21
Torture	2	3	9	6	1	21
Departure conditions	6	0	5	5	1	17
Connection to opposition groups	0	1	4	6	4	15
Other conflicts with authorities	0	1	4	6	0	11
Detention	5	0	4	1	0	10
Disproportionate punishment	1	0	4	4	0	9
Recipient conditions	0	3	2	1	1	7
Family relations	2	0	1	2	0	5
Health conditions	0	0	0	2	0	2
Asylum procedure	0	0	0	1	0	1
Families with children	0	0	0	0	0	0
Changed conditions	0	0	0	0	1	1

Table III. This table shows the co-occurrence matrix of the five empirical gender-related categories (columns) and how often cases were tagged together with the other 24 asylum motives (rows) in the public dataset of the Danish Refugee Appeals Board.

This confirms the general legal understanding that there is a category hierarchy in refugee law, as exemplified by the UNHCR guidelines, on how states may apply the category of gender-related persecution. It means that asylum is granted (or not), by subsuming gender aspects into the formal set of Convention grounds (race, religion, nationality, membership of a particular social group, political opinion). Gender-related persecution does not constitute a valid claim for asylum on its own. Rather, it is used in connection with other categories to establish a sufficient asylum motive.

5.2 Omitted and Articulated Gender-related Topics in Asylum Motives

A second observation and finding of our analysis of the origins of the empirical categories associated with the formally defined category of gender-related persecution is, that some topics are made visible, while others are omitted in the process of archiving. Table IV shows the distribution of the 11 topics (rows) among the five empirical categories (columns) of the Danish asylum cases. The percentages illustrate how each empirical gender-related category is composed of several topics. Some categories, such as *LGBT* persecution, are more homogeneous than others, in the sense that they are mainly composed of a few topics that contain a high share of cases in that category. The empirical topic of gender-related persecution on the other hand is broader in scope and more heterogeneous, since it it is composed of many topics that all have a comparably low share of the cases in that category.

	Gender-related persecution	LGBT	Sexual conditions	Marital conditions	Other gender-related persecution	Total
Homosexuality	9%	89%	12%	1%	1%	269
Rape	17%	12%	21%	30%	19%	262
Female genital mutilation	8%	0%	2%	17%	65%	177
Forced marriage	7%	0%	5%	41%	14%	143
Human trafficking	6%	1%	1%	1%	11%	48
Forced abortion	1%	1%	3%	10%	2%	41
Bisexuality	0%	15%	1%	0%	0%	39
Forced prostitution	4%	2%	2%	1%	5%	35
Transgenderism	0%	2%	0%	1%	0%	6
Forced sterilisation	1%	0%	0%	0%	1%	6
Transvestism	0%	0%	0%	0%	0%	0
Total	303	211	366	474	175	

Table IV. This table shows the 11 topics extracted from the UNHCR guideline on gender-related persecution (rows) and the five empirical gender-related categories in the public dataset of the Danish Refugee Appeals Board (columns). It illustrates the prevalence of different topics in the categories. The percentages indicate the proportion of overall cases in the category that are tagged with the topic. One case can be tagged with several topics and some cases are not tagged with any of the extracted topics. Thus, the percentages of the categories (in the columns) don't add up to 100%. The darker the cell shading, the larger the proportion of cases in that particular category that mention a particular topic.

Taking an archival perspective, and attending to power relationships, we analysed which topics are omitted in the case data and which topics are articulated and thus made visible.

As seen in Table IV, some topics are dispersed over several categories, such as *human trafficking*, *rape*, or *forced prostitution*. These topics are omitted in the database structure. If the database is accessed through one of the empirical categories, only a partial picture of all cases related to that topic is provided.

The topic of female genital mutilation is predominantly found in the residual category of *other gender-related persecution*. This topic is also omitted. Classification schemes often depend on residual categories to render themselves complete. Residual categories cover the cases that do not fit into *pure* categories (Møller and Bjørn (2011) following Bowker and Star (1999)). One thus expects a wide scope of topics to be included in that category. However, the empirical residual category in the Danish asylum case data, *other gender-related persecution*, is dominated by the topic of female genital mutilation. 65% of the 175 cases in that category cover this topic. In the context of archiving as the act of organising and preserving knowledge, members of a residual category are more invisible than other categories.

Other topics are given prominence by being subsumed under a homogeneous category with a narrow scope. Table IV illustrates that the topics of *homosexuality* and *bisexuality* are articulated in the category of *LGBT*. This category is mainly composed of these two topics.

In the context of an archive, categories take on the role of gatekeepers. They make the data records tractable and serve as the basis for further processing and access. In the context of data science, archival categories turn into features of computational models that aim to predict and inevitably also shape the future. A topic that has been omitted and thus rendered invisible in an archive, has therefore a reduced potential to influence the future when seen in a context of data-driven technologies. Individuals that are part of that category are less likely to be seen and engage with shaping the world. Møller et al. (2021) refer to data shadows as realities that are not captured in tracking data. Here we show how data shadows can also occur in classification schemes.

6 Discussion: Categorisation in Asylum Data and the Context and Practices that Shaped It

Star and Strauss (2004) argue that no category or work practice is inherently visible or invisible but something that we construct. As we see a shift towards casework becoming data-driven (Ammitzbøll Flügge et al. (2021), Saxena et al. (2020)), it is even more critical to recognize the visibility of categories as an arena of power and to deconstruct the origins of data structures (Star and Strauss (2004)). In this sense, the perspective of the archiver (in our case the the Refugee Appeals Board)

is privileged, as they decide which topics are made visible and accessible by being turned into a category and other topics are omitted and therefore difficult to access.

Such omitted topics are characterised by prior CSCW research as data shadows. Møller et al. (2021) establish them as an effective approach for sensemaking of data about a work practice in which stakeholders have differential power. By showing omitted topics (e.g. female genital mutilation) that are not immediately searchable in the database of asylum motives, data shadows appear in our data produced in and through the casework in the area of asylum. We were able to identify these omitted topics by tracing the origins of the five empirical gender-related categories: 1) gender-related persecution, 2) LGBT 3) sexual conditions, 4) marital conditions, and 5) other gender-related persecution (see Table IV). What we learn from the analysis of these empirical gender-related categories is that omitted topics, that do not emerge as a category on their own, are as important to map out as the articulated categories. These are the constituents of residual categories, but also topics dispersed over several categories such as e.g. human trafficking, that enable us to point to specific limitations of a large dataset.

Data shadows are often cast by the categories that are articulated (see Table IV), for example, the topic of homosexuality. Our analysis shows that 89% of the cases in the LGBT category are about *homosexuality*, which thus constitutes a homogeneous category and makes the topic of homosexuality visible. Heterogeneous categories with a wide scope can also cast shadows on the topics they are composed of. The empirical category of gender-related persecution in the Danish dataset casts such a shadow, as it is composed of most of the topics that were subject to this analysis, e.g. rape, human trafficking, homosexuality, and female genital mutilation.

Studying the use of large-scale data in work domains characterised by sensitivity and security concerns such as asylum (Bigo (2014)) is important to ensure they are included in the research on public sector decision-making. We apply NLP to deconstruct gender-related categories as a starting point for engaging with asylum decision-making as a practice. We take an archival perspective to acknowledge power as a factor that shapes the application of gender-related categories. A possible next step is to trace the effects of omitted topics. For example, in regard to female genital mutilation: Is the recognition rate of asylum lower for this particular group or has this topic been omitted because asylum has rarely been granted based on this motive? What does it mean to connect this insight to practice by taking a participatory approach? How can a participation-inspired NLP unfold in a complex, collaborative work domain such as asylum, where researchers act as intermediaries between interests of stakeholders with differential power?

In a design context, *categories* become features of data models that shape caseworker systems. Taking practice as our starting point, CSCW researchers need to find new ways of enabling practitioners to have a reasonable say over the design of these caseworker systems. In our case, that means taking a participatory approach to NLP as suggested by Caselli et al. (2021) and continuing encounters with the stakeholders of the Danish asylum system, such as the NGOs and Danish

national authorities, to further develop the questions we raise in relation to the data. The opportunistic nature of the study in this area with limited access means for example discussing our assumptions around specific categories while being walked through the asylum registration procedure in the Danish reception center.

7 Concluding Remarks

In this paper we set out to investigate how gender-related categories are presented in Danish asylum case summaries and how they can be probed for understanding the practices of asylum decision-making.

We illustrate how data science techniques and NLP can be useful for exploring omitted and articulated topics in asylum case summaries. We uncover data shadows; that is, realities that are not captured by data, such as the topic of female genital mutilation, which is rendered invisible by being organised under the residual category of *other gender-related persecution*. While these omitted topics have a reduced potential to engage in shaping the world in the context of data-driven technologies, being included in official archives is not the only way of being recognized. There are alternative archives that enable communities whose stories haven't been included in official archives to have an impact on the future (Vlachokyriakos et al. (2021)).

We also show how to investigate the empirical reality of a category hierarchy introduced by the historical omission of gender-based persecution as part of the definition in the 1951 Refugee Convention. Despite subsequent state support to include gender as a basis for granting asylum, our analysis shows how category hierarchies are nonetheless routinely reproduced in everyday asylum decision-making. Using a co-occurrence matrix, the analysis shows that in the Danish system, asylum motives related to gender always occur in connection with other categories. As socio-legal research confirms, this has implications for the practice of this domain, specifically for how cases are handled and evaluated by authorities (Spijkerboer (2000)).

This study has some limitations. We only studied a subset of appeal cases, for which the criteria for selection was not known. The NGO we worked with (the Danish Refugee Council) confirmed that not all cases are recorded in the archive. Furthermore, we are cis-gendered scholars with European citizenship and an academic education. That means we are speaking from a position of privilege and engage in an act of knowledge reorganising that is shaped by our standpoint (Harding (2004)). A critical next step for this research is thus to enable the contestation of our findings through participation of stakeholders of the asylum decision-making process.

We invite other researchers to engage with our code (published on GitHub) used for the analysis - as well as the open dataset - as a strategy for knowledge production in this area. Normatively, this study calls on the CSCW community to combine datasets and include data science techniques to study work practices, even if access is limited.

Acknowledgments

We thank Nanna Bonde Thylstrup for the insightful discussion about archival theory and Anna Højberg Høgenhaug for providing her first-hand knowledge into the publication practices of the Refugee Appeals Board. A big thank you to Hubert Zając for his great advice on the interpretation of the topic analysis. We are also thanking Vasilis Vlachokyriakos, Asbjørn Ammitzbøll Flügge and Trine Rask Nielsen for providing valuable feedback on the overall study. And last but not least, thanks to David Struthers and Colleen Jankovic for copy editing.

References

Ackerman, M. S., J. Dachtera, V. Pipek, and V. Wulf (2013): 'Sharing Knowledge and Expertise: The CSCW View of Knowledge Management'. *Computer Supported Cooperative Work (CSCW)*, vol. 22, no. 4, pp. 531–573.

Ackerman, M. S. and T. W. Malone (1990): 'Answer Garden: A Tool for Growing Organizational Memory'. In: *Proceedings of the ACM SIGOIS and IEEE CS TC-OA Conference on Office Information Systems*. New York, NY, USA, pp. 31–39, ACM Press.

Ammitzbøll Flügge, A., T. Hildebrandt, and N. H. Møller (2021): 'Street-Level Algorithms and AI in Bureaucratic Decision-Making: A Caseworker Perspective'. *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1.

Aragon, C., S. Guha, M. Kogan, M. Muller, and G. Neff (2022): *Human-Centered Data Science - An Introduction*. MIT Press.

Bannon, L. J. and K. Kuutti (1996): 'Shifting perspectives on organizational memory: from storage to active remembering'. In: *Proceedings of HICSS-29: 29th Hawaii International Conference on System Sciences*, Vol. 3. pp. 156–167, IEEE.

Bechmann, A. (2019): 'Data as Humans: Representation, Accountability, and Equality in Big Data'. *Human Rights in the Age of Platforms*.

Bigo, D. (2014): 'The (in)securitization practices of the three universes of EU border control: Military/Navy – border guards/police – database analysts'. *Security Dialogue*, vol. 45, no. 3, pp. 209–225.

Bowker, G. C. (2010): 'The Archive'. *Communication and Critical/Cultural Studies*, vol. 7, no. 2, pp. 212–214.

Bowker, G. C. and S. L. Star (1999): *Sorting Things Out: Classification and Its Consequences*, Inside Technology. MIT Press.

Boyd, D. and K. Crawford (2012): 'Critical Questions for Big Data'. *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679.

Byrne, R. and T. Gammeltoft-Hansen (2020): 'International Refugee Law between Scholarship and Practice'. *International Journal of Refugee Law*, vol. 32, no. 2, pp. 181–199.

Cakici, B., E. Ruppert, and S. Scheel (2020): 'Peopling Europe through Data Practices: Introduction to the Special Issue'. *Science, Technology, & Human Values*, vol. 45, no. 2, pp. 199–211.

Caselli, T., R. Cibin, C. Conforti, E. Encinas, and M. Teli (2021): 'Guiding Principles for Participatory Design-inspired Natural Language Processing'. In: *Proceedings of the 1st Workshop on NLP for Positive Impact*. Online, pp. 27–35, Association for Computational Linguistics.

D'Ignazio, C. and L. F. Klein (2020): *Data Feminism*, Strong Ideas. MIT Press.

Foucault, M. (1972): *The archaeology of knowledge*. Tavistock Publications.

Gilliland, A. J. (2017): '"A Matter of Life and Death: A Critical Examination of the Role of Official Records and Archives in Supporting the Agency of the Forcibly Displaced"'. *Journal of Critical Library and Information Studies*, vol. Retrieved from https://escholarship.org/uc/item/5787j3qd.

Gitelman, L. (ed.) (2013): *"Raw Data" Is an Oxymoron*, Infrastructures. MIT Press.

Harding, S. G. (2004): *The Feminist Standpoint Theory Reader: Intellectual and Political Controversies*. Routledge.

Hathaway, J. C. and M. Foster (2003): 'Membership of a particular social group'. *Int'l J. Refugee L.*, vol. 15, pp. 477.

Hathaway, J. C. and M. Foster (2014): *The Law of Refugee Status*. Cambridge University Press, 2 edition.

Irani, L., J. Vertesi, P. Dourish, K. Philip, and R. E. Grinter (2010): 'Postcolonial computing: a lens on design and development'. In: *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2. pp. 1311–1320, ACM Press.

Ismail, A. and N. Kumar (2018): 'Engaging Solidarity in Data Collection Practices for Community Health'. *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, pp. 1–24.

Justitsministeriet, Denmark (2016): *Bekendtgørelse om forretningsorden for Flygtningenævnet*.

Karusala, N., A. Vishwanath, A. Kumar, A. Mangal, and N. Kumar (2017): 'Care as a Resource in Underserved Learning Environments'. *Proceedings of the ACM on Human-Computer Interaction*, vol. 1.

Kitchin, R. (2014): *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE.

Muller, M., C. T. Wolf, J. Andres, M. Desmond, N. N. Joshi, Z. Ashktorab, A. Sharma, K. Brimijoin, Q. Pan, E. Duesterwald, and C. Dugan (2021): 'Designing Ground Truth and the Social Life of Labels'. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA, ACM Press.

Møller, N. H. and P. Bjørn (2011): 'Layers in Sorting Practices: Sorting out Patients with Potential Cancer'. *Computer Supported Cooperative Work (CSCW)*, vol. 20, no. 3, pp. 123–153.

Møller, N. H., G. Neff, J. G. Simonsen, J. C. Villumsen, and P. Bjørn (2021): 'Can Workplace Tracking Ever Empower? Collective Sensemaking for the Responsible Use of Sensor Data at Work'. *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 1–21.

Neff, G., A. Tanweer, B. Fiore-Gartland, and L. Osburn (2017): 'Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science'. *Big Data*, vol. 5, no. 2, pp. 85–97.

Nielsen, T. R. and N. H. Møller (2022): 'Data as a Lens for Understanding what Constitutes Credibility in Asylum Decision-making'. *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 1–23.

Pine, K. H. and M. Liboiron (2015): 'The Politics of Measurement and Action'. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. pp. 3147–3156, ACM Press.

Randall, D., R. Harper, and M. Rouncefield (2007): *Fieldwork for Design: Theory and Practice*. Springer.

Ring, A. (2014): 'The (W)Hole in the Archive'. *Paragraph*, vol. 37, no. 3, pp. 387–402.

Saxena, D., K. Badillo-Urquiola, P. J. Wisniewski, and S. Guha (2020): 'A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System'. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA, p. 1–15, ACM Press.

Seaver, N. (2017): 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems'. *Big Data & Society*, vol. 4, no. 2.

Spijkerboer, T. (2000): *Gender and Refugee Status*. Ashgate.

Star, S. L. and A. Strauss (2004): 'Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work'. *Computer Supported Cooperative Work (CSCW)*, vol. 8, pp. 9–30.

Suchman, L. (1993): 'Do Categories Have Politics? The Language/Action Perspective Reconsidered'. In: *Proceedings of the Third Conference on European Conference on Computer-Supported Cooperative Work*. USA, p. 1–14, Kluwer Academic Publishers.

Taylor, A., D. Sweeney, V. Vlachokyriakos, L. Grainger, J. Lingel, S. Lindley, J. Lingel, and T. Regan (2015): 'Data-in-Place: Thinking Through the Relations Between Data and Community'. In: *CHI '15, Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. pp. 2863–2872, ACM Press.

Thylstrup, N. B. (2022): 'The ethics and politics of data sets: deleting traces and encountering remains'. *Media, Culture & Society*.

Thylstrup, N. B., D. Agostinho, A. Ring, C. D'Ignazio, and K. Veel (eds.) (2021): *Uncertain Archives: Critical Keywords for Big Data*. MIT Press.

UNHCR (1951): *The Refugee Convention, 1951*. Geneva, Switzerland.

UNHCR (2002): *Guidelines On International Protection: Gender-Related Persecution within the context of Article 1A(2) of the 1951 Convention and/or its 1967 Protocol relating to the Status of Refugees*. Geneva, Switzerland.

Vlachokyriakos, V., C. Crivellaro, H. Kouki, C. Giovanopoulos, and P. Olivier (2021): 'Research with a Solidarity Clinic: Design Implications for CSCW Healthcare Service Design'. *Computer Supported Cooperative Work (CSCW)*, vol. 30, pp. 757–783.