ROSKILDE UNIVERSITY

INFORMATICS

Privacy concerns of linking Danish digital public records

Authors

Daniel Șerbănescu
dase@ruc.dk
Student no.64513

Lukas Kucerik
kucerik@ruc.dk
Student no.66892

Anton Kjær Hansen
akjaerh@ruc.dk
Student no.64510

Tomas Nemecek
nemecek@ruc.dk
Student no.64509

Supervisor
Bo Holst-Christensen

December 17, 2019

Abstract

In recent years, there has been a strong focus on digital privacy both from governments and private entities. While legislations such as GDPR and industry guidelines like NIST and ISO 29100 are efforts to mitigate some of the issues regarding digital privacy, our society's approach to data privacy is still questionable. Even when data sources follow both legal and industry standards and reveal limited personally identifiable information as contextually necessary, it is still possible to combine records from multiple data sources to build a profile containing personally identifiable information. In this project, we investigate the potential issue of linkability by reviewing legal and industry standards, analyzing online Danish governmental public record data sources and their approaches to privacy, and attempting to create a theoretical method to connect public records from multiple data sources to create a collection of personally identifiable information.

Contents

List of Figures

List of Tables

1 Glossary

Term	Definition
API	Application programming interface - Standard way for services to communicate. In this project it refers to web services.
CSV	Comma separated values - A file format used to store tabular data.
PII	Personally identifying information
Sensitive personal data	A special category of personal data as defined by Article 9 by GDPR
Linkability	The possibility of associating records belonging to the same data subject
Data source	Anything that provides data. In this project it refers to websites, APIs and CSVs.
Data subject	Any individual person who can be identified, directly or indirectly, via an identifier.
Data controller	A person, company, or other body that determines the purpose and means of personal data processing.
Data processor	A natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller

Table 1: Glossary

2 Introduction

As part of an ongoing effort to ease citizens' interaction with governmental agencies, many governments encourage the digitalization of public services such as looking up of public records of land or company ownership, searching through court records or even finding suitable doctors based on desired criteria.

These measures are commendable, as citizens no longer have to travel to governmental offices in person and deal with bureaucratical headaches and delays. However, digitalization opens the door to another brand of possible issues. Without careful consideration of people's data privacy and measures against privacy violations, digital public records can attract malicious actors to anonymously collect and misuse personal data at a startlingly large scale without having to spend many resources.

The issue of digital privacy has been approached by governments with legislative measures such as GDPR and the Danish Data Protection Act in an attempt to codify privacy protections and the consequences of breaking them. The technology sector has also broached the topic of data protection by creating industry standards which advise how a companies should treat their users' personal data. While following the legal and industry norms greatly reduces the risks of privacy infringements, due to the nature of our digitized society, one of the lingering issues is the possibility of combining data from multiple data sources. This allows the construction of extensive profiles containing personal, possibly sensitive, data that can be used for malicious purposes.

2.1 Problem Formulation

What are the privacy concerns of linking Danish digital public records?

For the scope of this project we examine the possible privacy concerns of linking Danish digital public records which are contained in publicly-announced governmentally-tied data sources that do not require authentication (openly accessible).

The reason for the choice of this topic is that many possible linkability issues could be conceivably individually avoided by not providing personal data to untrusted companies.

However, as a resident when it comes to governmental data sources, there is no such option to opt-out. While open access to this data might be in the public interest, residents have to implicitly trust the government that it will protect their data.

We chose the Danish government specifically due to the benefit of our personal experiences with the digital Danish public services and because Denmark is the most digitized nation in the world, according to a study by the European Commission [1]. Therefore, we believe that Denmark makes for an ideal case study.

The reason for choosing data sources that require no authentication is because the focus of this project is data collection which can be done anonymously on a large scale by any actor.

2.2 Limitations

One of the limitations of this project is that none of the members of our team have any legal background or experience, thus any statements regarding legislations are subject to inaccuracies.

The examined data sources are not indicative of all Danish governmental data sources. During this project, we examine some of the data sources that match the publicly-announced governmentally-tied criteria which contained a significant amount of data. For further explanation of our methodology for identifying data sources continue to section 4.2.2.

One of the data source classification factors, harvestability, is heavily dependent on the work of Khelghati et al. [2], thus any inaccuracies propagate to our project as well. For discussion about harvestability as a judgment criterion turn to section 4.3.1.

The proposed methods of how to aggregate and possibly use data are entirely theoretical to avoid any ethical predicaments.

The demonstration of linkability is somewhat lacking, as it is based on some reasonable assumptions, due to the fact that to properly demonstrate linkability one would have to cross the ethical boundaries as described in section 2.3.

2.3 Ethical Considerations

When it comes to the topic of personal privacy, there are ethical dilemmas to consider. In an effort to not reveal any personal data and to not provide any possible "how-to guides", this project does not provide any actual code or step-by-step instructions. We simply provide plausible scenarios of data collection and warn about the ways the data can be misused.

2.4 Document Structure

We explore the possible privacy concerns of linkability by reviewing the current EU and Danish legislation, different industry standards, and methodologies regarding data privacy in section 3. In section 4, we asses Danish governmental data sources and the data they contain. We present how the information can be aggregated from the data sources in section 5 and how through linking different data points a personal profile can be created. In section 6, we contrast privacy to the public interest, demonstrate how data can be derived, explore different ways of misusing collected personal data, and provide some recommendations in regards to security and privacy.

3 Review of Data Protection Legislation, Standards and Methodologies

In this section, we review the European Union and Danish laws and look at their attempts to protect digital privacy. Further, we look into how this legislation fits into the current industry privacy standards and privacy protection methodologies. We conclude with a look at the concept of data ethics.

3.1 Data Protection Legislation

3.1.1 GDPR

The General Data Protection Regulation (GDPR) [3], also known as "persondataforordningen", is a European Union (EU) law. The law protects natural person's personal data across the EU, by standardizing legislation and regulating how companies and public bodies process personal data. The intent is to "... strengthen individuals' fundamental rights in the digital age". [4]

GDPR consists of 99 articles, each containing several provisions. Some of the articles and their provisions, that are particularly relevant for the scope of this project, are listed here.

Article 6 (1): "Processing shall be lawful only if and to the extent that at least one of the following applies:".

Article 6 (1) (e): "processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;".

Article 9 (1): "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited."

Article 9 (2): "Paragraph 1 shall not apply if one of the following applies: ".

Article 9 (2) (g): "[P]rocessing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject;"

Article 86: " Personal data in official documents held by a public authority or a public body or a private body for the performance of a task carried out in the public interest may be disclosed by the authority or body in accordance with Union or Member State law to which the public authority or body is subject in order to reconcile public access to official documents with the right to the protection of personal data pursuant to this Regulation "

3.1.2 The Danish Data Protection Act

The Danish Data Protection Act ("Databeskyttelsesloven" or DDPA) supplements the GDPR and was passed by the Danish parliament on the 17th of May, 2018 to be ratified as "Databeskyttelsesloven" [5]. The law details certain provisions of the GDPR which the individual EU member states are to interpret and implement. The law text is also translated to English [6] and while the translated text holds no official legal validity, it will be referenced throughout this report for readability reasons.

Part 6 of the DDPA, titled "Restrictions of the rights of data subjects", lists certain exceptions and situations for when derogation of the data subjects' data protection rights can be allowed. This includes "... essential considerations of public interests" [6] such as public safety and national security. Other particular cases are criminal investigations, the public and financial interest of the European Union (EU), judicial proceedings, statistical and scientific purposes, and the enforcement of civil law claims. Furthermore, the data subjects' right of access to data, processed by the public administrative authority, may be restricted in accordance with the Open Administration Act (Offentlighedsloven) of 2013 [7].

The data sources examined in this report are all administered by the government through its public administrative bodies. These data sources expose personally identifiable information, to the extent that in accordance with laws, decrees, and provisions, require this information to be public. Section 4.2.8 reviews each data source with reference to the applicable legal basis for processing the contained personal data.

3.2 Industry Standards

3.2.1 What Identifies a Person? (ISO 29100)

The International Organization for Standardization (ISO) has also created a framework that attempts to clarify some of the main areas regarding privacy and personally identifiable information (PII). The ISO 29100 standard [8] is better used as a clarification for various terms regarding privacy rather than a comprehensive framework for helping companies and organizations with creating their privacy policies and data management plans.

The standard defines PII as information that can be used to identify a person to whom it relates to or is in any way linked to them - directly or indirectly. For instance in Denmark, a CPR number is considered a PII since it is an identifier that refers to a natural person. Similarly, driving license or passport numbers are also considered PII since they can be related to a natural person. Email addresses, geographical locations or telephone numbers can be used to establish a communication with an individual and therefore are considered PII, too. Additionally, any data linking to any of the above-mentioned information are also considered PII.

Data can be combined to point to a specific person. Such combination might be harmless in a larger context of use (e.g. social media and online forums), but might be very specific when describing a person within a company. An example of such combination might be first name, age, and education level.

Throughout this document, we will be mentioning a few actors in relation to personally identifiable information. These actors are described in multiple standards and documents and therefore we decided to use these terms in our work as well. In the ISO standard, they are defined as follows.

- PII principles provide their PII for further processing (the website and API users).
- PII controllers determine how and why the PII is collected (the organization management).
- PII processors process the PII on behalf of the PII controller (people who directly process the information for analysis, validation, etc.).
- Third parties receive the PII from the PII processor or the PII controller but do not process the information on behalf of them. Third parties become a PII controller once they receive the principal's PII. We will not directly focus on this actor within the scope of this work, but we will point out their existence when applicable.

3.2.2 Privacy in Organizations (NIST 800-53)

Another industry standard, and arguably one of the biggest, is the document Security and Privacy Controls for Federal Information Systems and Organizations [9]. The document is published by the National Institute of Standards and Technology (NIST) under the name NIST 800-53 (revision 4) and it contains the latest updates from the year 2015. It describes various controls that can be implemented in order to assure that the organization adheres to industry standards regarding IT security and data privacy. Even though this document is intended for the United States, it is used by companies and organizations worldwide since it contains more specific information and guidelines for companies within the IT industry.

In Appendix J (Privacy Control Catalog), the document presents a number of controls that can be used to make sure that the organization is handling personal data in a globally-accepted manner. A privacy or a security control can be described as a safeguard or a countermeasure to avoid, detect, counteract or minimize a risk that can be present in an IT system. The authors have grouped the 26 different controls under 8 categories:

- authority and purpose
- accountability, audit and risk management
- data quality and integrity
- data minimization and retention
- individual participation and redress
- security
- transparency
- use limitation

The standard specifies that the organization should select the controls which are relevant for their case, ideally based on the organization's requirements and the need to protect personally identifiable information. Those controls should be reviewed by the security or information officers and people in risk executive functions [10]. Afterwards, the controls should be implemented according to the agreed plan.

Some of these controls have even influenced our classification in section 4, such as:

- AP-1: Authority to Collect
- AP-2: Purpose Specification
- DI-1: Data Quality
- DI-2: Data Integrity and Data Integrity Board

Control categories which have the most relevance to this project and public Danish systems are described in the following sections.

Data Quality and Integrity

This control category, if implemented correctly, might enhance the public confidence in the organization and that the collected personal information is accurate and complete. Since we are focusing on the systems maintained by various ministries, it is important that the citizens of Denmark feel confident about the organizations handling of their personal data. These systems are interconnected and sometimes rely on the other systems to be working well and provide accurate data, thus any inaccurate data could potentially result in compromised safety or erroneous evaluations of citizens. The Danish Agency for Digitization even mentions the citizens' confidence as one of its main goals in its plans for digitization [11] for the years 2016 to 2020. The systems should allow for both Danish citizens and foreigners to communicate securely and with confidence, regardless of whether they communicate with the public authorities or among each other.

This control category mentions 2 controls. The first one, *DI-1: data quality*, says that the PII collected

should be accurate, relevant, timely and complete. The organization should check for these aspects regularly and make sure that the data are collected directly from the citizens (as much as possible). The second control, *DI-2: data integrity and data integrity board*, suggests that processes ensuring data integrity must be properly documented and maintained. NIST also suggests creating a Data Integrity Board in the organization whose function would be to inspect and overlook the mentioned processes.

Individual Participation and Redress

Similarly to the European equivalents of this standard, the document also mentions individual participation as one of the main aspects of privacy. The citizens interacting with the public systems should be informed about which of their data is collected or further processed and should be able to make active decisions based on this information. The organization has the responsibility to provide individuals with access to their data and the possibility to edit or correct it. The systems should also make sure that the data its users entered are validated to the greatest extent possible since human errors are very common and very difficult to avoid.

The controls for this category consist of:

- *IP-1: Consent* - means of ensuring the users can agree or disagree with certain types of data being collected and act according to their decisions.
- *IP-2: Individual access* - makes sure that the individuals have access to all the PII that has been collected by the system.
- *IP-3: Redress* - provides a way for the users to amend the information that is incorrect. The systems should also make sure that the amended data are further distributed to the systems which use this data.
- *IP-4: Complaint management* - means for the citizens to communicate their dissatisfaction with the systems to the authorities.

Security

Since the governmental organizations deal with a lot of information about Danish citizens, it is very important to make sure that this data is safe and not accessed or processed by parties that can misuse such information. As Denmark is in the process of digitalizing its systems and automating many processes that can be optimized, security is one of the top priorities for the years 2016 to 2020 [11]. Safe systems, however, are not only short-term goals, but should arguably be one of the most important goals of any governmental system.

The controls in this category are as follows.

- *SE-1: Inventory of personally identifiable information* says that the organization should maintain and update an inventory containing all the systems that are processing or storing any PII. It should also make sure that any updates to the systems are registered and communicated to the security officials.
- *SE-2: Privacy incident response* describes planning for privacy breaches and incidents and responding to them in an appropriate way.

Security itself is a very broad and complex topic with multiple standards [12] [13] describing how an organization should handle security and privacy risks. We will not describe any specific security measures or controls as they are not the goal of this project. However, our research indicates that poor data privacy decisions might pose direct risks to the safety of governmental institutions if not taken seriously.

3.2.3 Privacy Practices and Legal Issues

The industry standards we described in the previous sections are all dealing with different ways of enhancing privacy within government or state-related systems. However, since privacy is a very delicate topic and privacy frameworks and standards tend to explain these enhancements with a certain level of abstraction, many times resulting in rather vague explanations.

Nevertheless, this ambiguity is not unexpected. Industry standards provide guidelines for organizations world-wide and with privacy being very tightly coupled with a different set of laws in each state, the standard guidelines might considerably vary for different states and countries. In order to keep relevance for organizations in multiple countries, the standards tend to focus more on different ways of mitigating risks related to privacy rather than specifying which information can or cannot be exposed.

Since our assessment is performed in Denmark and is looking at the danish governmental data sources, we will place the danish law above the industry standards if we find any discrepancies. Similarly, if the GDPR guidelines do not provide enough information for us to assess some of the data sources, we will turn to the danish laws which are building on top of the GDPR.

3.3 Data Protection Methodologies

3.3.1 Privacy by Design

Privacy by Design (also known as PbD) consists of seven principles which describe the best practices to keep the IT systems privacy-oriented, while not sacrificing their functionality. The principles of PbD build and extend upon the principles of Fair Information Practices (FIPs) which share the same privacy goals [14]. The following section describes the principles in relation to our research and governmental institutions in Denmark.

Proactive, not Reactive; Preventative not Remedial. The system should be developed in a way that anticipates and prevents events that might cause damage to its users' privacy. This is especially the case with governmental websites, as they many times need to contain information about an individual's address, workplace or phone numbers. In short, the organization should not wait for incidents to happen, instead, it should make sure they will not happen.

Privacy as the Default. This principle should be especially focused on in regards to public systems. As the citizens often do not have the choice to not participate in a system run by the state and might not have the technical knowledge to change their privacy preferences. Organizations should make sure that their systems protect as much personal data as possible, without users needing to take any further action.

Privacy Embedded into Design. Privacy should be integral to the system, not added later as a feature or an optional preference. In the era of digitalization of public service systems, they should be built with a focus on the citizens' privacy from the beginning.

Full Functionality – Positive-Sum, not Zero-Sum. Since it is important for the state to keep their citizens' data accurate and safe, security is an important aspect of any IT system. Privacy and security should thus go hand-in-hand. The PbD principles suggest that it should be possible to make the systems both secure and privacy-friendly, without sacrificing public safety. It is however important that the organizations keep this in mind from the beginning (as mentioned in the *Privacy Embedded into Design* principle) and do not make any trade-offs which are unnecessary.

End-to-End Lifecycle Protection. When privacy is correctly embedded into systems, the data about its users should be protected through its whole life-cycle. This means that the organizations have to make sure that the individuals' PII is not disclosed during its collection, storage, amendment, and final disposal.

Visibility and Transparency. The system should make sure that its users can independently verify the organization's practices as they have been promised or guaranteed by the law. This principle brings an interesting discussion as Denmark already follows the tradition of an open government [15]. This brings about more visibility and transparency on the governmental level, however, there can be cases where this approach can undermine privacy. This dichotomy is discussed in more detail in section 6.1.

Respect for User Privacy. Privacy should be built into the systems in a similar fashion as many modern products are developed - by keeping its features user-centric. The preferences should be user-friendly and privacy notices easy to follow. The citizens' privacy should be placed above all other requirements. However, this point poses a challenge to many state-run institutions. When developing public systems, the state is required to involve information agencies and security officials who can have different opinions on what data should be kept secret and not accessible to the authorities or persisted over long periods of time. A similar challenge can be found in the case of the introduction of the automatic number plate recognition system in Denmark [16].

3.3.2 Privacy-Enhancing Technologies

Privacy by Design describes guidelines that a privacy-oriented system should be built in accordance with. However, when developing such a system, slightly more technical descriptions are needed. Privacy-Enhancing Technologies (PET) provide a set of high-level technical explanations of various technologies that can be used to protect the digital privacy of individuals. These principles are more of a technical nature and can be used as a set of guidelines for the developers and IT specialists who create and maintain the IT systems.

There are multiple descriptions of privacy-enhancing technologies, some of which may vary in the technologies described. Wang and Kobsa provide a comprehensive explanation of many different privacy enhancements mentioned in OECD [17] and other standards [18], while adding principles of anonymity and other techniques worth considering when developing in IT system.

The most important PET principles related to our project are:

- *Data Minimization* - minimize the impact on users' privacy when collecting and processing their data by choosing the least privacy-invasive options.
- *Purpose Specification* - state the purpose of the data collection and its usage before such a process begins.
- *Collection Limitation* - the data from the citizens should be obtained lawfully and its collection should be limited to the organization's purpose.
- *Use Limitation* - personal data should not be disclosed to unspecified third parties or used for other purposes than those specified by the organization.
- *Onward Transfer* - transfer of personal data to third parties should not happen unless the data is adequately protected.
- *Integrity/Accuracy* - the data stored and collected should be accurate and timely relative to its usage context and purpose.
- *Enforcement/Redress* - the organization must ensure access of individuals to their personal data and make sure they can amend them. Complaints regarding users' data should be handled as well and acted upon accordingly. The organization should also verify that it is still compliant with the laws and its promises to the public and failure to comply should have consequences for the organization.

Seničar, Jerman-Blažič and Klobučar described various threats to privacy and how they can be tackled in

IT systems development [19], outlining the specific techniques for protecting a website's privacy of its users. Those include privacy mechanisms like *encryption*, *key management* and *certification services* or projects like *P3P* (Platform for Privacy Preferences). Other research focuses on client-side defenses like *PGP*, *onion routing* or *anti-phishing tools* [20].

Other privacy-enhancing technologies include anti-scraping techniques used by many modern websites today. Even though these techniques are not directly protecting users' privacy, they are a valuable server-side defense, especially in the recent years where the website bots and scrapers (or crawlers) can automate the process of data collection about individuals. In their work, Haque and Singh present a number of such techniques and mitigation strategies [21]. Some noteworthy mentions would include network traffic analysis, white-listing or blacklisting of IP addresses, technologies for preventing automated behavior like CAPTCHA, or honeypots and honeynets which are widely used by a number of major companies in the networking-related business.

3.3.3 LINDDUN Methodology

LINDDUN is a privacy threat framework [22], which describes a systematic approach for privacy threat modeling to quickly reveal threats to privacy and possible countermeasures.

LINDDUN is an acronym made from the primary privacy threats it aims to counter: **L**inkability, **I**dentifiability, **N**on-repudiation, **D**etectability, **D**isclosure of information, content **U**nawareness, policy consent, and **N**on-compliance.

As the LINDDUN methodology is based on STRIDE [23] which focuses primarily on security, both LINDDUN and STRIDE can be used within the Security Development Lifecycle (SDL) [24].

Hard and Soft Privacy

Two distinctive privacy properties are hard privacy and soft privacy. The goal of Hard Privacy is data minimization. It focuses on the data subject that concerns themselves with minimal trust required for processing of their data. The goal of Soft Privacy is to provide data security and process data for the tasks which the user consented to. Soft privacy starts with the premise that the data subject already lost control of personal data and has to entrust the data controllers in order to protect their privacy.

Wuyts et al. [22] proposes that the following hard privacy properties have the most relevance within the LINDDUN framework:

- Unlinkability - hiding the link between two or more actions, identities or data
- Anonymity - hiding the link between identity and action or information
- Pseudonymity - identifier of a person in a certain context, thus multiple identifiers could be allowed
- Plausible Deniability - ability to have no irrefutable evidence about an event or action
- Undetectability - hiding user activities in a way that an attacker cannot distinguish their existence
- Unobservability - hiding users' activities in a way that it cannot be distinguished who they belong to
- Confidentiality - hiding the data content by means of encryption

Among the soft privacy properties, they name the following:

- User Content Awareness - the user's consciousness regarding his own data
- Policy and Consent Compliance - rules regarding data protection determined by the stakeholders and users

Other privacy properties are also important and they are usually covered during the security assessment phase:

- Integrity
- Availablity
- Forward security

3.4 Data ethics

Data ethics is about responsible and sustainable use of data. It is about doing the right thing for people and society [25]. It is a step further than just compliance with certain data protection laws.

Data is a very valuable asset but at the same time, it is also a risk [26]. Therefore, data ethics play a significant role in companies that deal with any kind of data. Tranberg et. al. [25] gathered a set of data ethics principles by which we abide when we look at the ethical considerations of treating data within this project. These principles are as follows:

- Human Being at the Centre - the main beneficiary of data processing is a human being
- Individual Data Control - the human being should be able to control every aspect of data about self
- Transparency - every automated decision made should be fully transparent to the individual
- Accountability - the data processor should take the accountability to reduce risks of data collection
- Equality - by paying special attention to vulnerable groups of people, algorithm's bias is reduced

During the span of this project we took measures that abide by the aforementioned data ethics principles. The human is at the center and therefore we did not include any personal information about data subjects nor stored any PII during the research for this project. Since we took this measure, we followed the individual data control and transparency principles.

Accountability is avoided as well as we do not keep any PII and we merely point out issues with the original data processors' systems.

Equality is one principles that influenced most of our work as we identified the groups of people who are most vulnerable to the publication of their own personal information within the governmental public services. We, therefore, ensured that every resident would be treated equally and the methodology bias would be at a minimum level.

3.5 Implications of Theory

In recent decades, the public sector has become increasingly digitalized. As data plays an integral part in our lives and is gathered in centralized repositories, concerns have been raised about the citizens' privacy. GDPR and the Danish Data Protection Act came into effect because of these concerns and raised the bar of privacy protection in the digital space. However, one could argue that these laws and legislations are still ambiguous to some extent. This is, nonetheless, understandable since it is difficult to create laws to fit all industries and sectors. Even though we believe that introducing these laws was a good starting point, they should be subject to regular improvements as time goes by.

Similarly to the previous point, the guidelines in the industry standards tend to be quite generalized too. Despite this fact, it is obvious that exposing personal data in the form of PII is something that no organization should be doing without the users' consent. Initially, this might seem like a minor obstacle, however, the issue gets more difficult when one starts considering linking of publicly available data. The organization might release data that is harmless on its own, but when it is linked in certain ways, this data might become personally identifiable. In this work, we try to point out that such linking is possible and that it can be potentially dangerous if it is not managed carefully.

4 Classification of Data Sources

In order to prove that linking of publicly available data might pose risks to Danish citizens, we first needed to assess data sources of governmental institutions. We categorized data sources by which governmental institution owns them, what data are they contain and how easy it is to automate harvesting them.

4.1 Classification Table

Datasource	Access Method	Input	Output	Impact	Harvest-ability Score	Size estimate
DAWA	CSV	Address, Region	Address, Region	Negligible	100%	3746107
Tinglysningen	API	Address	Company name, Address, Company address, Worth of Property, Worth of Land, Morgage Information	Limited	95%	N/A
Statstidende	API	Name, Date Range	Name, Education, CVR, CPR (proclamations of death)	Significant	93%	N/A
CVR	API	CVR, Company name, Name, Address	CVR, Company Address, Company board members, Company Capital information, Person's ties to companies	Limited	95%	1535526
Sogn.dk	API	Free Text, Name, Church name, Address	Name, Work address, Work Phone number, Work title	Limited	100%	N/A
Sundhed.dk	API	Free Text, Region, Gender, Age group, Address	Name, Work title, Work address, Work phone, Webpage, Gender, Age group	Significant	95%	N/A
Folketingets telefonbog	Website	Name, Department / Party, Work title	Name, Work title, Work e-mail, Work phone, Facial Image	Significant	86%	N/A
Borgerforslag	Website	Free text, Proposal ID	List of proposers, Name, E-mail, Phone	Limited	86%	N/A
Avokatnøglen	Website	Name, Company Name, Area of expertise, Jurisdiction, Region	Name, Company Name, CVR, Work title, Appointment Date, Work Phone, Work Email, Can appear before the High courts, [...] The Supreme Court	Limited	93%	4829
CHR	Website	Name, Adress	Name, Company Name, Address, Company Address	Limited	86%	N/A
Forsknings-databasen	Website	Free text, Name, Educational Institution	Name, Work address, Educational institution, Work Email, Work Telephone, Facial Image	Significant	88%	N/A
Jobnet.dk	Website	Freetext, Work title, Address, Region, Education	Name, Work title, Work phone, Work email, Work Address	Limited	88%	N/A
Domstol	Website	N/A	Judge names, Defendant names, Advocate names, Plaintiff names, Case id, Brief case description	Significant	73%	N/A
Arbejdstilsynet	CSV	Company Name, Safety rating	Company Name, Safety rating	Negligible	100%	N/A
Autorisations-registret	CSV	Name, Birthdate, Authorization date, Work title, Valid authorization	Name, Birthdate, Authorization date, Work title, Valid authorization	Significant	100%	N/A
DK-Hostmaster	API	Domain name	Name, Company name, Address, Company address, Phone, Work phone, E-mail, Work e-mail	Limited	91%	1316367

Table 2: Data source classification table

4.2 Classification Methodology

4.2.1 Qualification of data sources

The data sources that we have chosen to investigate for this project are maintained by governmental bodies or organizations very closely tied to it. The data sources also contain publicly-announced data and require no registration or authentification.

By governmental data sources, we mean websites such as Tinglysning or datacvr.virk.dk, which are administered by the Danish government.

For our scope, we have also decided to include data sources which are administered by organizations closely tied to the government. That includes data sources such as sogn.dk or advokatnoeglen.dk. Our reasoning for their inclusion being, that one cannot be part of a church or practice law, without a membership to one of these organizations.

Publicly announced data sources are sources which indicate what data they contain. Meaning we do not use data sources which contain data they should not, either due to a technical error or other reasons.

The reason for not choosing data sources that require any registration or log-in is a way for us to prevent any issues regarding the terms and conditions of any data sources.

Another reason for choosing data sources with these specific characteristics is that data contained in these data sources is generally curated by governmental institutions. As a citizen or resident, you cannot ask for your data to be removed. This data can thus be freely accessed with minimal obstacles and we can avoid most of the legal and ethical issues during our project.

4.2.2 Methodology for finding data sources

At the beginning of the project, the methodology for finding data sources was mostly ad-hoc as we were investigating this topic and still determining our scope. During this stage, we looked at data sources that we already knew about and interacted with in our daily lives. We also thought about different registries there are for people such as researchers, practicing lawyers, and doctors.

As our project became more concrete, our approach became more systematic. We looked at Danish Ministries and which governmental websites they are responsible for. The administrator of the Danish internet space, dk-hostmaster.dk, helped us significantly, as during this project, we were able to look up governmental institutions and what other domains they own. However, as of 25.11.2019, this functionality has been disabled on dk-hostmaster.dk due to privacy concerns [27].

Another way that we looked for data sources is looking at danish municipalities. We did not have much success, with this approach as it did not provide us with general enough data sources that would be worthwhile to integrate into our data aggregation process.

During this process, we created a mind map to keep track of the data sources that we have found. In figures 1, 2 and 3 an example of unfolding the map about the data sources connected to the Ministry of Justice can be seen. For the full interactive mind map please visit this link [28].

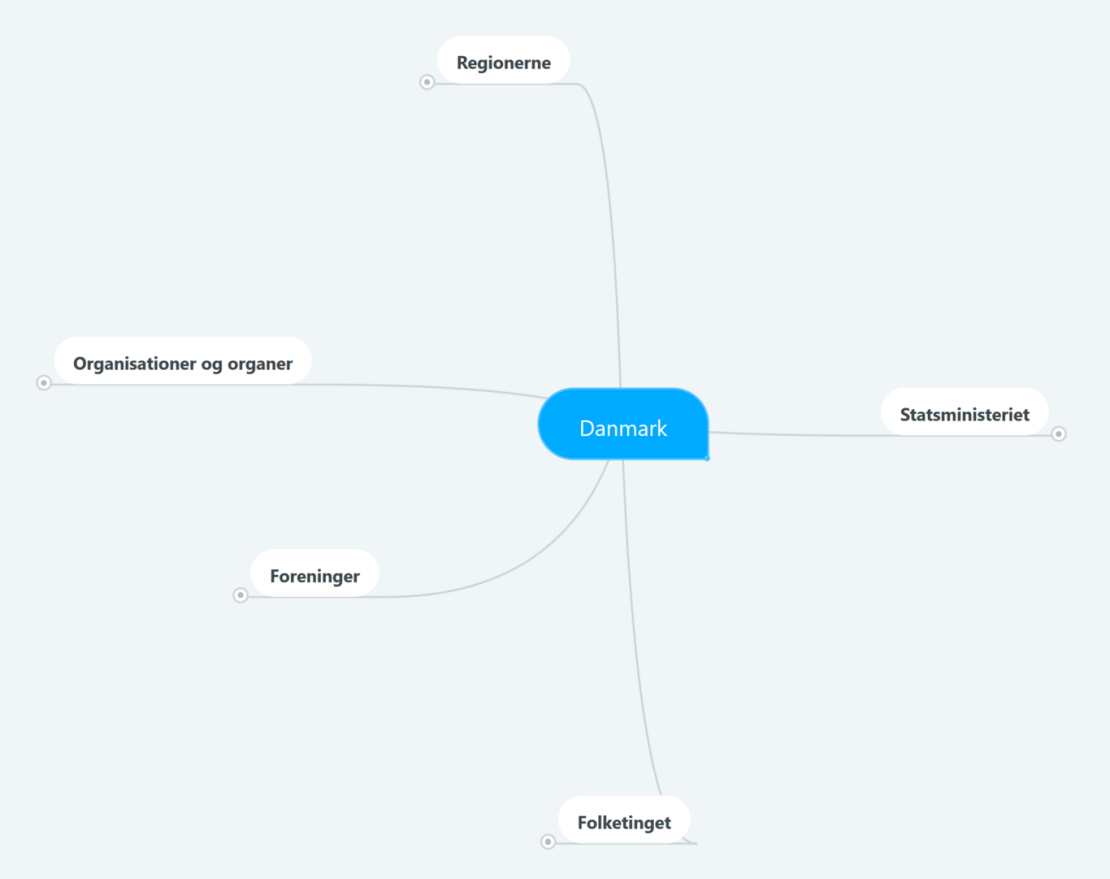

Figure 1: General overview of the mind map

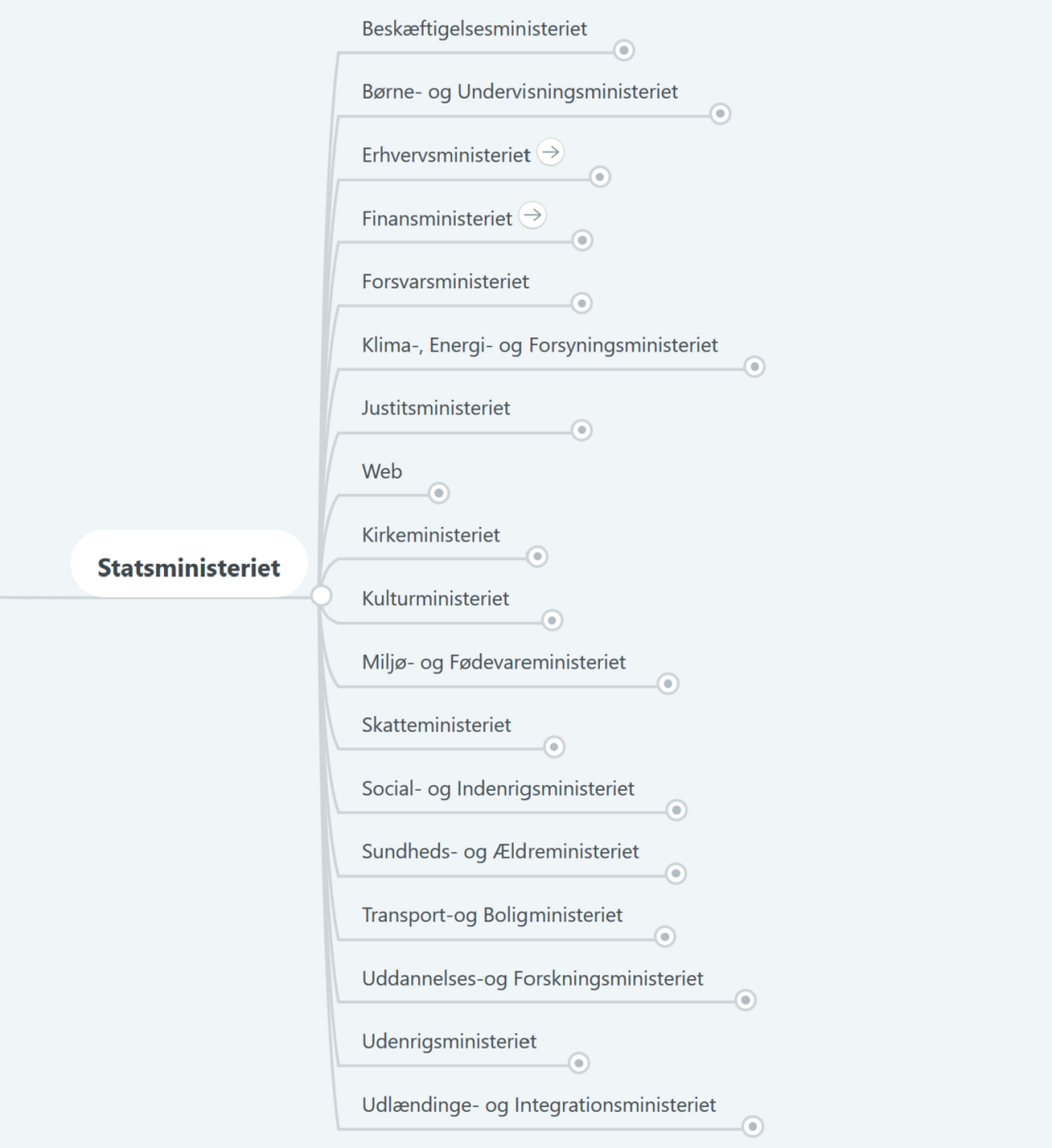

Figure 2: Expanded view of the "Statsministeriet"-category

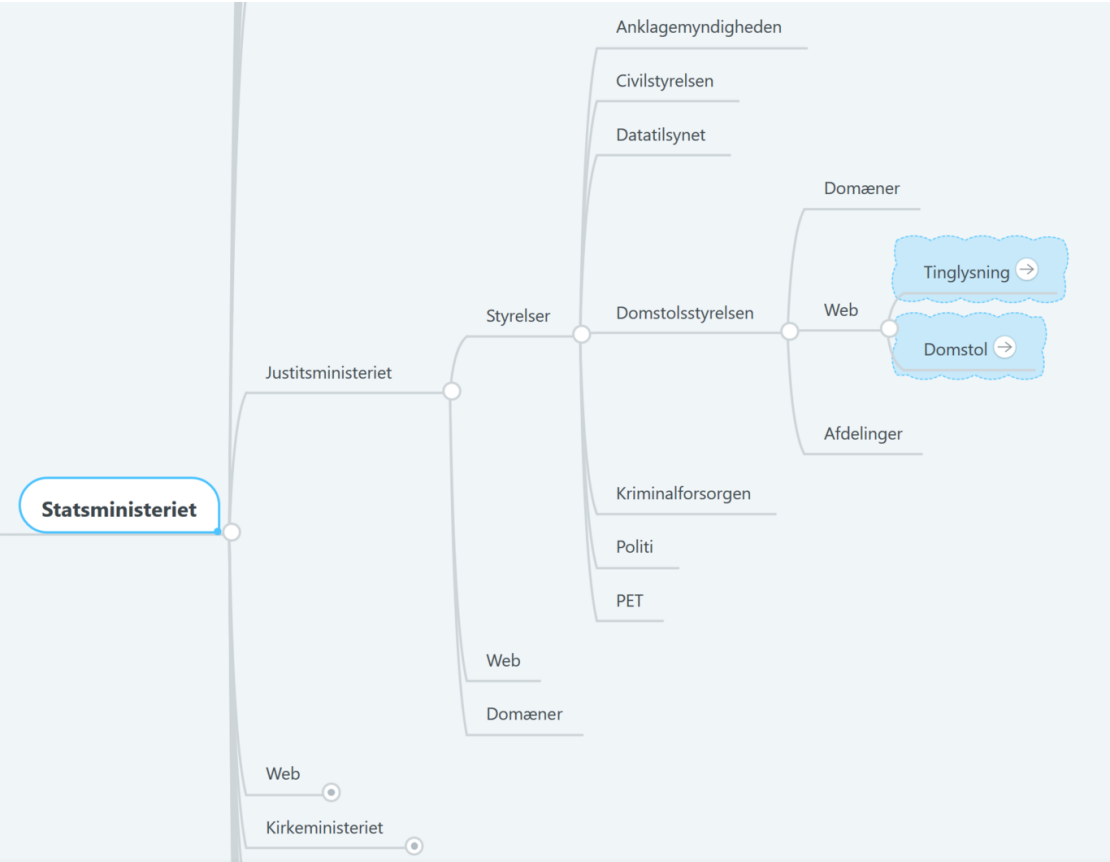

Figure 3: Expanded view of the "Justitsministeriet"-category

The Danish Agency for Data Supply and Efficiency provides a list of public data sources [29]. However many of the listed data sources do not fit our criteria, as they require authentication for usage. The list also does not contain all the data sources that we were able to find, so overall this list has not been very useful for our project.

4.2.3 Data Source Description

To provide a larger context to the classification table in section 4.1, below you can find the descriptions for all of our selected data sources.

Tinglysning: The Land Registration Court ("Tinglysningsretten") is in charge of keeping a central register of "[...] titles to land, mortgages and other charges, marriage settlements etc." [30]. Some of the records are digital and entirely public while others are only available by visiting the court in person.

Statstidende: A governmental publication published on the internet by The Department of Civil Affairs under the Ministry of Justice. Statstidende processes notifications that in accordance with the law shall be declared. These notifications usually have a legal effect on citizens and public bodies.

Det Centrale Virskomhedsregister (CVR): CVR is the Danish central business register allowing users to search for businesses and their participants. CVR is part of the "Virk Data"-data catalog, managed by the Danish Business Authority, and is "[...] promoting the use of open data from the Danish Government in the business community" [31]

Sogn.dk: Also known as "Sogneportalen" communicates information about the parishes in the Evangelical-

Lutheran Church (Folkekirken) to the public. One of the portals' feature is to search for vicars and other administrative employees.

Sundhed.dk: This website is the access point for the Danish health care services on the internet (eHealth). The initiative for the portal was taken in 2001, with the intent of collecting and centralizing information about the health care sector in a single place. The parties behind Sundhed.dk are Danske Regioner, the Ministry of Health and Kommunernes Landsforening (KL). These parties have established a Board of Directors and a Steering Committee [32].

Folketingets telefonbog: The Danish Parliament publishes a vast amount of information on the website https://www.ft.dk. This includes general information and legislative documents that allow insights into the legislative process. The website has an online phone book, that allows users to search for employees of the parliament's administration, including the elected representatives and other employees.

Borgerforslag: Another website governed by the Danish Parliament is a citizens' proposal system (https://www.borgerforslag.dk). This website allows Danish citizens to see, support and suggest proposals for something that the citizen believes should be changed in law or in society. If a proposal gets 50.000 declarations of support, the Parliament is obliged to discuss and treat the proposal.

Advokatnøglen: Advokatnøglen is a registry for all the lawyers that are members of the Danish Bar and Law Society. This means that every lawyer that is allowed to practice law in Denmark, has to be a member and thus can be found in this registry. The bar society consists of around 6000 lawyers [33].

Det centrale husdyrbrugssregister (CHR): This register contains information about animal livestock and their owners. The register is operated by the Ministry of Environment and Food of Denmark. The intent is to provide effective detection and response to disease outbreaks [34].

Forskningsdatabasen: The Danish National Research Database, operated by the Ministry of Higher Education and Science, contains "Nearly 1 million research publications and more than 75 thousand researcher profiles collected from the research databases of 14 Danish universities and research institutions." [35].

Jobnet.dk: The jobcentres in Denmark assist job seekers with their job search. The intent is to assist by shaping the applicants' CV and expand their network. Jobnet is a portal that connects job seekers and employers, and contains a job bank, a CV repository, and information about the public jobcentres [36].

Domstol: The Danish courts' website contains, among other things, court lists for all of Denmark's courts. The information available on the court lists varies based on the type of court and case. Names of defendants and plaintiffs are, for example, not available on the internet in criminal cases, but still available for court attendants showing up in person.

Autorisationsregistret: 19 different groups of health care professionals, requires to be registered and authorized to perform health care work. These people are registered, with their authorization status, full name, date of birth, profession, date of registration, ID number and specialist title. The register is operated by the Danish Patient Safety Authority [37].

DK Hostmaster: Is the administrator of the ".dk" top-level domains (TLD). DK Hostmaster maintains a part of the Domain Name Service (DNS) infrastructure in the Danish internet space and a public WHOIS-database of registrants of ".dk"-domains [38].

DAWA: Is an acronym for 'Danmarks Adressers Web API' and offers simplified access to Denmark's authoritative addresses. The information available through DAWA is aggregated from multiple governmental registers [39].

Arbejdstilsynet: The Danish Working Environment Authority is an agency under the Ministry of Employment that "[Works for] a safe and healthy working environment in all companies in Denmark."[40]. This data source describes data from a part of the agency's website titled 'Smiley søgning' that rates the work environment in different places of work. This data can be downloaded as a CSV-file.

Attributes	Definition
Access method	Describes whether the data source can be accessed as a website, API or a CSV
Input parameters	What can be searched by in the data source
Output	The data that the data source contains
Impact	Describes how much the data can be misused. The range is Negligible, Limited, Significant, Maximum
Harvestability Factor	Estimate of how easy it is for a data source to be mined
Size estimate	Estimate of number of records contained within the data source

Table 3: Data source classification attributes

4.2.4 Input/Output

The values of the input and output attributes describe the different data elements that can be collected from the data sources. These attributes are standardized in such a way that unless otherwise specified, data attributes which contain multiple other attributes are under one umbrella term. For example, address which can contain multiple data attributes such as street name, street number, city, zipcode is simply marked as address.

Data sources marked as CSV, also have identical values for input and output. This is due to the fact that one can access all data contained in a CSV and you can search by anything contained within the CSV.

4.2.5 Impact

Impact describes the degree to which the data contained within the data source can be exploited. The range of values is: Negligible, Limited, Significant, Maximum. This range is based on privacy threat modeling done by Furnell et al. [41].

4.2.6 Size Estimate

Size estimate describes the number of records contained in the data source. With some sources, we can only estimate. Other sources provide the number themselves or it can be obtained by using a query using an empty string or a space character, which returns the entire database.

This estimate can be fairly incorrect, as we cannot accurately control for duplicates in the data source. For example in the Researcher Database [42], while the empty space query returns results for ca. 91 000 entries, it contains duplicates, where researchers working for two different universities will have 2 separate researcher profiles. Thus the number of actual persons contained in the data source can be significantly lower.

4.2.7 Harvestability Score

To judge how easy it is to collect data from data sources, we use the Harvestability score. This score is represented by a percentage and is based on the harvestability factor (HF) presented by Khelgati et al. [2].

This factor represents the degree of difficulty for a data source to be harvested or crawled, meaning how easy it is to gather data from the data source through an automated process.

The proposed factor is defined for websites only and takes into consideration the characteristics of the website which influence how difficult it is to harvest it.

For this project, we use the harvestability as defined for our website data sources, and use a scaled-down version of it to fit the assessment of APIs.

For our own ease of comparison, we translated the harvestability factor into a percentage-based system. To be able to compare the harvestability of APIs and websites, we used the full version of the Harvestability calculation which is then calculated into a percentage.

To see the full Harvestability score calculation table please turn to section 8.1 in the Appendix.

Harvestability of Websites

Harvestability factor is defined as:

$$HF_W(w) = \sum_{i=1}^{n} \left(1 - \left(w_{p_{fi}} \cdot \left(\frac{1}{m} \sum_{j=1}^{m} (Fa_{fj})\right) \cdot Cr_{fi}\right)\right)$$

As per the formula, given a website w, for each of its feature Cr_{fi} is multiplied by the average performance of a harvester Fa_{fj}. wp_{fi} symbolizes whether the feature is present on the website or not.

n is the number of features and m is the number of harvesters.

Due to the scope of this project, the harvester performance (Fa_{fj}) that is used in our classification is based on the performance of a general harvester for other websites with similar characteristics as defined by Khelghati et al. [2]. The number of harvesters (m) is 1 because this theoretical harvester is used.

The characteristics that determine the complexity of website harvesting fall into thee categories: web development techniques, website policies, and data and content [2].

Web development technique category consists of website features such as embedded scripting languages in the page HTML which create dynamic content, complex URL redirections, usage of Applet or Flash or inconsistent coding practices. All of these features can make harvesting more difficult.

The website policies category consists of two main characteristics: search policies and security, privacy and legal policies. Search policies define the website's query interface, indexing, and navigation, while security, privacy, and legal policies describe such things as whether a login is required to access data, or if there are other security methods used such as CAPTCHA, IP address blocking, etc.

Data and content category represent website features such as the data type, format, layout, presence of metadata, content language and how scattered the data is across pages.

There are 15 elements that define the website harvestability factor. The maximum score is 15, which indicates that the website is very easy to harvest.

Harvestability of APIs

The overall idea for API harvestability is similar to the harvestability of websites. The difference is the number of elements that the data source is evaluated on. Since many of the characteristics that apply to websites do not apply to APIs, we simply removed them and thus reduced the number of elements to 7, meaning the maximum harvestability score for an API is 7.

The factors that are considered for APIs are: data structure and data layouts, navigation, search policies, indexing policies, and CIA+ policies.

Harvestability of CSVs

Some of the obtained data sources were provided from a website as a CSV. By nature, CSVs are easily harvestable for the data contained within. Thus no calculation is provided for these data sources, as the discussion regarding them turns more towards the quality or usability of the data instead of harvestability itself. Therefore all CSVs automatically get a 100% harvestability score.

4.2.8 Lawful Basis for Processing

One factor by which we classify data sources, that is not in the classification table, is the legal basis of processing. It explains the reason why the data is publicly available

Tinglysning: The provisions for the registers governed by Tinglysningsretten are established in the "Bekendtgørelse af lov om tinglysning"[43]. Chapter 8 details the overall purpose of keeping and publicizing such registers. The provisions below are especially relevant in the context of digitization and privacy.

(§ 50) establishes that "tingbogen", "bilbogen", "andelsboligbogen", and "personbogen" should be kept digitally. (§ 50 a.) details the overall purposes for these registries, which are to support the judiciary system, laying the foundation for advertising land registry information to the public, and for statistical purposes.

(§ 50 c.) lists the legal and illegal applications of information obtained from the digital registries, and further states that CPR-numbers are prohibited from disclosure.

(§ 50, stk. 8) states the following as legal use cases for the disclosed information: conveyance, insurance, prosecution, and other legal matters, mortgaging of real estate and movable property, credit rating, management of real estate, and counseling in connection with these purposes.

Statstidende: (§ 6) in "Lov om statstidende" [44] states that everyone freely can access the messages published in Statstidende. (§ 5, stk. 2) specifies that personal information, which in other contexts is considered confidential, should be available no more than one year after initial publication. (§ 5, stk. 3) dictates that civil registration numbers (CPR-numbers) only should be published in proclamations of death.

Det Centrale Virskomhedsregister (CVR): Everyone should have access to the information about the legal entities that are registered in CVR as per (§ 11) in "Bekendtgørelse om Det Centrale Virksomhedsregister og www.cvr.dk"[45]. (§ 2) lists the information available in the register, which includes:

- an entities CVR-number
- start date and possibly end date
- type, name, and address of the entity

- name, address, position, and CPR-number or CVR-number of participants, founders, and executive members. CPR-numbers are registered but not public.
- phone and e-mail

Sogn.dk: "Bekendtgørelse af lov om menighedsråd" [46] sets the provisions for the governing of parishes' by the parochial church council. (§ 12, stk 2) mandates that the newly elected chairmen of the church council to publish a list of the chosen council members. We have not been able to identify a national law allowing the processing of personally identifiable information of staff which is not part of the council. Article 6 (1) (e) of the GDPR may, however, establish the legal basis for processing.

Sundhed.dk: As per (§ 57 d, stk. 2) in "Sundhedsloven"[47] general practitioners are mandated to publish information about their medical practice concerning the citizens' choice of doctor.

Folketingets telefonbog: We contacted "Folketingets Oplysning"[48] as it wasn't immediately clear to us what legal basis allowed such processing. We were told the basis for processing the information in the telephone book can be found in Article 6 (1) (e) of the GDPR [3]. Folketingets Oplysning furthermore pointed us to a manual published by The Danish Data Protection Agency[49], that loosely translated bears the title "Guidance on data protection in connection with employment"[50], which provided us insights into the legal basis for publishing information for work purposes.

Borgerforslag: (§ 2, stk. 7) in "Lov om etablering af en ordning for borgerforslag med henblik på behandling i Folketinget" [51] establishes the lawful basis for collecting, using and storing information about persons. It is noted, as per the data processing policy [52], that the main proposer and co proposers' contact information is published.

Advokatnøglen: The Danish Bar and Law Society's legal validity is established in "Retsplejeloven" [53]. Chapter 15 details certain provisions for the Danish Bar and Law Society but doesn't clearly establish the legal basis for processing the information found in this data source. We inquired The Danish Bar and Law Society and learned that the basis for processing is found in (§ 37, stk. 3) in "Bekendtgørelse om godkendelse af ændringer af vedtægt for Det Danske Advokatsamfund" [54], that states the Danish Bar and Law Society operates an internet website, where citizens can search for lawyers in a particular geographical area and for lawyers with a specific area of expertise. Processing may also be covered by Article 6 (1) (e) of the GDPR.

Det Centrale Husdyrbrugssregister (CHR): (§ 13) of the "Bekendtgørelse om registrering af besætninger i CHR" [55] clearly states that everyone has access to information from CHR, with the exception of the provision described in (§ 13, stk. 2). This provision prohibits the public to access phone numbers and e-mail addresses of the data subject.

Forskningsdatabasen: We were unable to identify a distinct provision that establishes the lawfulness of processing PII in connection with Forskningsdatabasen. There is, however, a provision (§ 19.46.02.11) in the National Budget for 2017 [56] that is aimed for activities that seek to create knowledge about and promote research integrity. Further legal base can be found in Article 6 (1) (e) of the GDPR [3] and (§ 10) in Databeskyttelsesloven.

Jobnet.dk: Chapter 6 of the "Lov om organisering og understøttelse af beskæftigelsesindsatsen m.v." [57] establishes the legal base for IT-systems in the employment sector. (§ 33 and § 34) specifically describe provisions for the functioning of Jobnet. We have not been able to find the legal base for the processing of PII in job postings, but these may be covered by Article 6 (1) (c) (e) and Article 9 (2) (b) (f) (g) of the GDPR.

Domstol: (§ 5, stk. 3) of "Bekendtgørelse om retslister og om massemediernes aktindsigt i og opbevaring

af kopier af anklageskrifter og retsmødebegæringer mv." [58] specifies that court lists are published on the internet. The court lists published on the internet may not contain the names of the accuser and the accused or other information that is covered by Article 9 (1) of the GDPR. The provision furthermore specifies that court lists published on the internet are to be deleted two weeks after the court lists' period of validity has expired.

Autorisationsregistret: (§ 1) of "Autorisationsloven" [59] notes the purposes of the law are to strengthen patient safety and the promoting quality of health care services through the authorization of particular groups of healthcare professionals. (§ 2) establishes that the Danish Patient Safety Authority has the mandate of authorizing persons, that have completed a particular education, and keeping registers containing the different groups of authorized healthcare professionals. (§ 2, stk. 4) gives the Danish Patient Safety Authority the mandate to establish rules concerning the publication of authorizations.

DK Hostmaster: (§ 1) of "Domæneloven" [60] explains the purpose of the law is to provide the framework for accessing, using and managing internet domains. (§ 18) instructs the administrator (DK-hostmaster) to create and maintain a WHOIS-database containing domain registrants' name, address, and phone number. (§ 18, stk. 2) imposes the administrator to ensure that the data in the WHOIS-database is accurate, updated and publicly available.

DAWA: (§ 12) of "Adresseloven" [61] details that the Danish Ministry of Climate, Energy and Utilities shall establish, operate and maintain a public register of addresses. This register is part of the danish 'basic data' (Grunddata)[62]. (§ 12, stk. 2) establishes the aforementioned ministry to be the Data Controller of the register. (§ 16) states that the information contained within the register is made digitally available as common basic data for everyone. It is worth to note that there is no explicit personal data in DAWA as such, but as an address can be linked to a data subject this data may still be considered to be personally identifiable.

Arbejdstilsynet: We were unable to find a separate law providing the legal basis for processing the data contained in the 'Søg smiley'-function of the agency's website. The basis for processing may be found in Article 6 (1) (e) of the GDPR [3].

4.3 Classification Observations

4.3.1 Harvestability as a Classification Criterion

Harvestability is a very useful measurement for this project, as it describes how much prior preparation a harvester would require for a website or an API.

However, when trying to implement this concept into our project, we encountered a few obstacles.

One of the issues is accuracy because in our work no actual harvester is being used. Therefore the harvestability calculation is entirely theoretical. Our general harvester performance is entirely dependent on Khelghati et al. [2], which we use as our benchmark. To properly evaluate the harvestability of a website, one should create multiple harvesters and attempt to harvest the website, unfortunately, creation of such harvesters is outside the scope of this project.

Another concern is the comparability of harvestability between websites and APIs. Khelghati et al. focus primarily on the websites, thus many of the measurements do not apply to APIs. If we apply the same standard to APIs as to websites, APIs are significantly more harvestable. That is usually the case, however Khelghati et al. do not take into consideration cases where APIs have elements, that do not apply to websites, that make them harder to harvest. One such case could be for example not following the standard REST API practices [63], trying to figure out query parameters without documentation or a website to give an example would cause difficulties. Another example of different criteria could be performance. If the database has to make complex operations and is not very performant, the response can take significant time, which decreases the harvestability.

One of the other issues with harvestability is also the timeliness of some of the judgement criteria. For example, Flash or Applet, being an element to determine harvestability, becomes less and less relevant with each passing year. In the year 2014, when this measure was defined, there were significantly more websites using these technologies. However, during our research, we came across only one. It is similar to the case of the Multi-frame page factor, we did not come across a single website using this design pattern. We believe that criteria such as these somewhat inflate the overall harvestability score and thus could be worthwhile to re-adjust these measurement elements in the future and possibly remove them. We tried to validate the results of our harvestability score calculation, by looking at the data source harvestability score and comparing it to our ad-hoc estimates for harvesting based on our experience.

For example, when looking at the Harvestability score for APIs, sogn.dk has a harvestability score of 100%. We agree with this calculation as it, the API can be searched by multiple parameters and has very few limitations. On the other hand the least harvestable API, Statstidende, has 93.33% harvestability score. We believe that the score should be somewhat lower, as the two features that is has that lower the harvestability score are different data layouts and somewhat strict search policies, which make harvesting significantly more difficult and require special fine-tuning.

When it comes to the harvestability score for websites we agree with the harvestability ranking order as well. The two most harvestable websites are Advokatnølgen and DK-hostmaster with a score of 93.33% and 91.66% respectively. These two websites have very simple and consistent pages, with very little dynamic content, which makes harvesting fairly easy. The least harvestable website is domstol.dk with 73.33%. We believe that this score should also be lower, as the website contains Flash elements, as well contains differing data layouts, some of the data is structured and data formats can different from record to record. All of these elements make it almost impossible to harvest in our opinion.

Overall while we might disagree with some of the finer details such as the feature weights as proposed by Khelgati et al., we still believe that the harvestability factor provides a good general overview of difficulty

of harvesting data sources and even though there were a few hindrances in trying to use harvestability, it is very valuable resource for classifying data sources based on how easy it is to automate data harvesting. This knowledge can then be used when trying to design an automated process for aggregating data from these data sources.

4.3.2 Issues with Assessing Data Quality

As specified in the NIST standard, data quality is one of the important privacy controls [9]. Originally we aspired to include it as one of our classification criteria, however, we faced the issue of how to correctly measure it and validate the data.

Measuring data correctness and quality externally is a fairly difficult problem as confirmed by Clauß [64]. He argues that one would require a general master source of information to properly compare data contained in a dataset. Since, in our case, there is no such master source, the only possible way would be to compare multiple data sources. However, our data sources have only some overlap in the data they contain, so the only plausible option would be to download all data from all the data sources and compare them against one another. As mentioned in section 2.2 and 2.3, our data aggregation is entirely theoretical to avoid any ethical issues, and therefore we do not conduct such evaluation.

Another issue with measuring data quality in our case is that the government is the authority in charge of providing accurate data. Any third party systems providing similar types of information receive the data from governmental systems, thus any inaccuracies would be contained there as well. Therefore, we can only assume that while there can be inaccuracies in these data sources, they are as accurate as possible in this context.

Additionally, Pipino, Lee, and Wang argue [65] that data quality is actually a multi-dimensional concept and it is very difficult to address it as one simple classification criterion like we originally intended to. They proposed a table of 16 different dimensions which data quality can be defined by. Furthermore, they suggest that the quality evaluation is dependent on who is currently evaluating it and the results of such evaluations might differ to a large extent. Therefore, if an organization needs to evaluate the quality of their data, they should perform both the objective assessment based on metrics relevant to the organization's needs and a subjective assessment which would derive from concerns of different employee groups, clients and other system users. Based on these assessments, the organization can see what dimensions of their data quality should be improved. As we already pointed out, such evaluation would be too time-consuming, if not impossible from our standpoint, and we decided not to include it in our classification.

4.3.3 Photographs as Biometric Data

When trying to assess the impact of the data contained in the data sources, one of the big questions that we came across is the issue of photographs and whether they count as sensitive data.

Under GDPR images can be images are considered personal data as per Article 4 [66] personal data is defined as: "*any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*".

The image of a person can also fit under the definition of biometric data which is defined as *"personal data resulting from specific technical processing relating to the physical, physiological or behavioral characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data" [66].*

However that Recital 51 also states that photographs do not automatically mean biometric data: "*The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person" [67].*

While in the cases of Folketingets telefonbog and Forskningsdatabasen the personal photographs are optional and they are not used to identify natural persons through technical means, the photographs can be harvested from the data sources and then technically processed by a third-party actor for identification or other possible malicious purposes.

When it comes to Danish law, since 2002 it was the practice of the Danish Data Protection Agency to differentiate between 'situational photos' and 'portrait photos' [68]. This differentiation led to portrait photos being considered as sensitive and personal information, whereas situational photos would be considered as regular personal information. This practice changed in September 2019 and there is no longer differentiated between these types of photos. Thus both types of photos are considered sensitive personal information.

Based on the GDPR and the Danish data protection laws, we, therefore, decided to consider photographs as biometric data not based on how it is used by the data source, but on how the image can be misused, which is reflected in our impact rankings.

5 Data Aggregation and Processing

5.1 Data Aggregation and Processing Introduction

In this section, we demonstrate that it is theoretically possible and programmatically feasible to develop a system that can aggregate, organize and process large amounts of data from multiple publicly available data sources. We examine different ways such a theoretical system could function, present a visualization showing the relationships between the selected data sources and show how such a system could be used.

5.2 Data Aggregation Techniques

As mentioned in sections 4.2.7 and 4.3.1, one of the criteria that we judged the data sources on is their harvestability, which indicates how easy it is to gather data from them. As detailed in the aforementioned sections, websites are generally harder to harvest compared to APIs, and CSVs are the easiest to harvest.

As per Krotov, web scraping can be defined as an automatic approach to extraction and organization of data from the Web for further analysis. It consists of three phases: website analysis, website crawling, and data organization [69]. While this description is focused on websites, it can be applied to other data sources as well, such as APIs and CSVs.

Data source analysis is a pre-requisite for any harvesting because first, one needs to understand the underlying technologies of the data source and data structure of the information contained therein. For websites, this means analysis of the architecture of the website by examining HTML, CSS, Javascript, etc. For APIs, this means inspecting the design of the query interface and the structure of the returned data. CSVs need to be investigated regarding the structure of their data. We have conducted such data source analysis as part of our classification. For more detail turn to the harvestability score break down in section 8.1 in the Appendix.

Data source crawling is performed by a script or a program that gathers data from the data source. Websites are crawled through by automatic browsing of the site and collecting of the data. APIs can be harvested by simple clients that make requests and handle the responses. The data in CSVs can be accessed without any limitations, therefore extracting it is a very simple process.

Data organization is the last phase of data aggregation. After harvesting data sources, the data has to be pre-processed and organized meaning it has to be transformed into something that can be used. For example, even if the collected data contains similar attributes, they might have various naming conventions or the data values are formatted differently. Therefore it needs to be standardized so that it can be processed properly.

The most popular programming languages for crawling data sources and data organization are Python and R, this is due to the popularity of these languages in the data science community and the plethora of libraries and frameworks to simplify these processes. While all the phases can be automated, they still need a large degree of human oversight [69] to make the data aggregation and the further processing function.

5.3 Data Processing Methodology

Data aggregation can be achieved by three different approaches. One being real-time data harvesting, which has real-time data harvesting as its driving force. The second approach comprises of data harvesting, storing and indexing and the third approach is a mixture of the previously outlined approaches.

A data pipeline is a method of processing data from multiple sources in a predefined order starting with a single query.

A real-time harvesting pipeline can leverage only real-time data so a minimum amount of storage is necessary. Therefore, regardless of the size of the pipeline, most of the ordinary personal computers nowadays would be able to use only temporary memory during the entire process. We call this type of pipeline a data processing pipeline.

In contrast to the data processing pipeline, we define data collecting pipeline to be the process of collecting the maximum amount of data extractable from a data source without having defined a clear scope or value beforehand.

The pipeline can be built following either the processing or collecting principles depending on its purpose. For a higher efficiency, a layered architecture that combines both processing and collecting principles can also be taken into consideration.

5.3.1 Comparison of Processing and Collecting Pipelines

Implementation Difficulty

Pure data processing pipeline does not deal with any disk storage operations or communication with relational databases, however, it has its own complexities. The intricacy of data processing lies in the organization of dataflow of the different inputs and outputs of the data sources. Another issue is that the data processing pipeline cannot scale very easily, as each data attribute and data source increase the pipeline complexity.

Time to Yield

Data processing pipeline requires a minimum amount of setup compared with a collection pipeline which would require all relevant data sources to be harvested beforehand. The harvesting would take proportional time to the amount of data available in the data sources. Therefore the processing pipeline would yield results faster.

Storage Considerations

For data processing pipeline just the temporary memory (RAM) is necessary for operations and therefore no disk storage is required. On the other hand for a collecting pipeline one needs to house all the entirety of all the relevant data sources needed for a pipeline.

Quality of Data

For data processing pipeline the data would always be fresh (up to a couple of minutes older than its data source) and therefore might be closer to reflect the current state of the item it represents. However, a downside of the processing pipeline is the missing change history.

Security Considerations

In the case of anti-harvesting techniques implemented by the data sources, the data processing pipeline would have an advantage over the data collection pipeline. For example in case of a request limit data processing pipeline is less likely to hit it, because it does not do many requests to a single data source.

Legal Concerns

The legal liability for using the processing pipeline is dependent entirely on the intended usage, while the tool itself would conform to the law. On the other hand, the tool from a collecting pipeline can easily violate the law.

Ethical Concerns

A processing pipeline would have fewer ethical concerns because it involves no storage of data and is a mere advanced search process. The ethical gray and black zones start as soon as one begins to store data which might have not gotten proper permission for.

Comparison Conclusion

Based on the comparison factors a processing pipeline has fewer issues than a collecting one. The only downside is the inability to look at historical data or track data changes.

5.4 Linking Data Sources

In the previous sections, we described three different approaches for collecting and processing data about individuals from public sources. We refrained from implementing these solutions, due to ethical reasons as specified in section 2.3. However, to provide proof that the linking of public data sources is possible, we have created a visualization of the various data sources and the contained data attributes.

Figure 4 presents a high-level overview of the relationships among various data sources and their attributes. Each data source can take multiple attributes as inputs (red arrows on the diagram) and output data (blue arrows on the diagram). The diagram shows all the data source types - the APIs, websites and CSV files. However, it is important to note that the diagram does not contain all the identified attributes since many of the attributes do not provide value for the purpose of this project and would make the diagram unnecessarily complex.

The diagram in figure 4 can be read by starting at any data source and see what attributes this data source provide as outputs. Following the output attribute, we see that it is connected with another data source as its input. By repeating this pattern, we would traverse the diagram and find multiple data sources and attributes, collecting important information with each step. Such a process would result in a pipeline which could be repeated on multiple records that represent various people. We define these records, containing personal information, as profiles. The ways that these profiles can be theoretically misused are discussed in section 6.3.

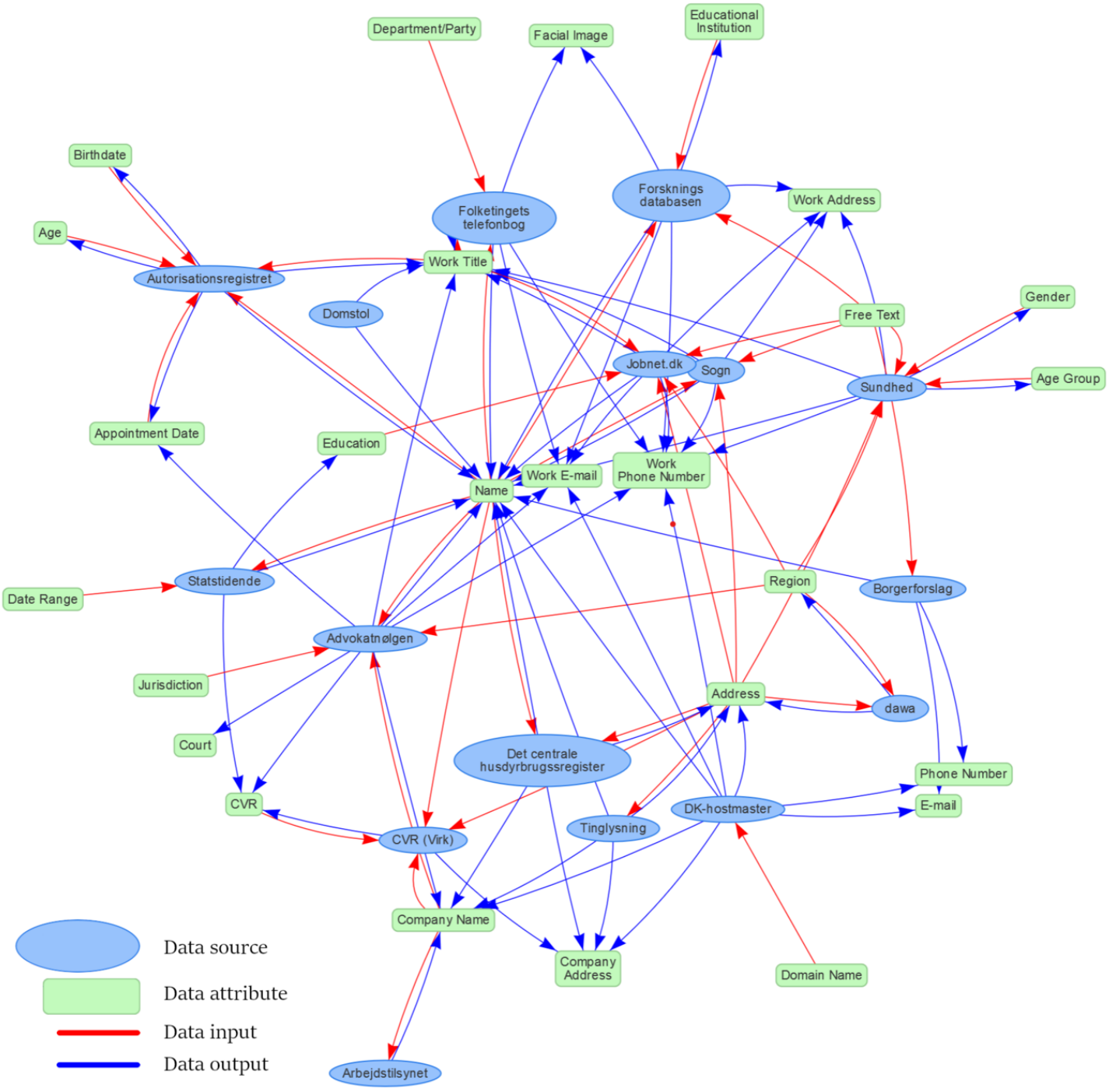

Figure 4: Visualization showing the relationships between the data sources

5.4.1 Optimizing the Pipelines

Because of time limitations and scope constraints, we did not create a more elaborate solution for the pipeline. However, we can already foresee a number of ways to make the pipeline more efficient with higher chances of finding more information about various people. In this section, we shortly describe the different ways of increasing the linkability of public records.

- Attribute Mapping - some attribute values could be mapped to other attributes as well. An example of such mapping could be found between the *company owner* and a *name*, since the information

about who the company owner is is actually in the form of a person's name.
- Conditional Attributes with Shared Values - similarly to the previous point, some attribute values can be used by other attributes as well, if certain conditions are met. This means that once such an attribute is found and a specified condition fulfilled, the attribute's value can be shared by other attributes. An example of such a scenario can be found in the case of a *company address*, which can potentially serve as a *work address* for a person who is this company's owner.
- Implicit Attributes - some attributes could be implied or derived from others without a direct mapping similar to the one mentioned in the two previous points. An example could be deriving the *gender* of a person from their *name* since, in Denmark, the male and female names are generally relatively easy to categorize. More examples of implicit attributes can be read in section 6.2.

Contrary to the mentioned improvements, there are also the aspects of this pipeline which would decrease its accuracy, efficiency, or could potentially present false positives. These aspects would have to be considered before the process would begin and the pipeline would have to be adjusted accordingly.

- Data quality - in this project, we place our trust fully in the public data sources. We assume that the records are always up to date and relevant to the people that they concern. However, this does not have to be the case with every data source and assessing this aspect would be fairly difficult, as we have already pointed out in section 4.3.2.
- Attribute similarities - our theoretical pipeline is very dependent on the attributes being linkable. Based on the fact that they contain the same values, they can be mapped to the same profile which would point to a specific person. However, not being careful with the attribute selection could create a number of false positives. An example could be finding a lot of information about a person with a name that is very common in Denmark - it would be unlikely that all this information is pointing to that specific person since the same name is shared by many other people. In that case, one can use Web Person Search techniques that utilize machine learning algorithms to distinguish between persons [70].

5.4.2 Demonstration of Linkability

The basis for the theoretical pipeline is the idea that one person can be found in multiple data sources, which have at least one overlapping attribute. This attribute can then be used to link the records from one data source to the other. The effectiveness of such linking is dependent on the uniqueness of the shared attribute. For example, a commonly used name is not an optimal linking attribute, as it is not exclusive to a singular person. On the other hand, a work email or a telephone number are unique identifiers and would make for a more suitable pipeline input parameter.

During the linking process, it might get to a point where the input parameter returns multiple matches. To ensure that we are not adding incorrect attributes to our profile, found attributes need to be corroborated with initial attributes. Even when a single result is returned, corroboration is required.

Thus to demonstrate linkability one would have to prove that data sources share data attributes and records regarding a person. To prove that the data sources contain records regarding a person, one would have to examine the content of the data sources, which is not in the scope of this project due to the ethical considerations. However, it is reasonable to assume that there are records referring to the same person in multiple data sources as they are managed by governmental organizations. Since we have proven in figure 4 that data sources share data attributes, we have thus demonstrated the possibility of linking digital public records as much as we can within the constraints of this project.

5.5 Data Aggregation and Processing Observations

As previously mentioned, our selection of data sources is very limited to avoid any ethical issues. However, if one would consider more data sources which are also owned by private organizations or even purchased access to repositories which are closed to the wider public, the amount of data sources linking to various attributes would increase by a large number. This would naturally result in increased linkability and therefore make this process even easier.

There would be a number of issues regarding the implementation of the proposed pipelines. One such issue would the false positives as mentioned in section 5.4.1, another would, of course, be the development of harvesters. While some data sources could theoretically use the same harvester, many of them would have to be custom-made. This would take up quite a bit of resources as thorough testing of such harvesters would be required. Any changes to the data sources can potentially break the harvester and introduce even more complexity into the development.

6 Discussion

6.1 Privacy in contrast to Public Interest

We outline in section 3.1, that legislation regarding privacy establishes that 'public interest' can override the individual's right to privacy. Since the definition of the term 'public interest' is contentious, parts of the law are up to interpretation. Therefore we approach this discussion from different perspectives.

We approach the subject from the perspective of media law and ethics. Further, we also argue that public interest is tightly coupled to the conception of democracy and thus we relate it to a broader discussion on how different types of democratic governance can influence the way the term public interest is perceived and understood. We conclude this section by relating public interest to 'Offentlighedsloven' and other relevant legislation.

6.1.1 Media Law and Ethics

The term "fourth estate" refers to the press and the media due to their ability to communicate and frame political issues. As a consequence of this power, the media is obliged to be responsible in their reporting. Media ethics address this responsibility and are particularly relevant in the conversation about the public interest. When reporting information about a private person, that person's privacy rights must be balanced against the public interest. However, that is a term that is hard to define.

In a case from 2015, a danish media outlet covered a tragic fire, where five people lost their lives. The media published pictures of the victims (four of whom were children), obtained from the victims' mother's Facebook profile without her consent. The chief editor, Karen Bro, argued that she had the right to publish the photos because it was a matter that public interest, which she defined by what interests society. Vibeke Borberg, a media lawyer, argued that 'public interest' is not dependent on the amount of people interested in the subject and is a principle from media law and ethics [71].

Ejvind Hansen, a philosopher of the public sphere, claims that these views may be a product of our different perceptions of democracy [72].

6.1.2 Public Interest in Democratic Theory

Public interest can be defined as "[T]he general welfare and rights of the public…". We assert that 'general welfare' and 'rights of the public' are important characteristics in a democracy.

We will show how the views of Borberg and Bro can be reflected in democratic theory.

Consumer Democracy

The core of consumer democracy is the individual citizens' (the consumer's) wishes and is fundamentally anti-elitist. The idea of a 'good democracy' is largely dependent on whether or not the consumer gets what they want. The opinions of experts, who may be particularly qualified in some areas, are equated with the opinions of the general citizen.

It is in this democratic ideal that Karen Bro (representing the tabloid media) can argue that the 'public interest' is defined by what interests society [72].

We argue that in a consumer democracy the citizens' preference, on whether to have personal information published online, should be respected. Reflection of this ideal may be found in legislation respecting non-disclosure of name and address.

Deliberative Democracy

At the core of the deliberative democracy is the "...collective decision making with the participation of all who will be affected by the decision or their representatives..."[73]. Jesper Strömbäck lists some important points in the deliberative democracy [74] :

- Discussions take place in the public sphere and in smaller settings.
- Discussions focus on 'rationality', 'impartiality', 'intellectual honesty' and 'equality'.
- Discussions are valuable by themselves and useful when coming to an agreement and to understand conflict.

This ideal is seen by Vibeke Borberg, and media law in general, as they represent a more deliberatively oriented view of democracy [72].

We argue that this democratic model enables citizens to engage in public debate, for example in discussions related to private information in public data sources, as long as their arguments are based on reasoning and objectivity.

Competitive Democracy

In a competitive democracy there is little trust that the average citizens are able to govern society. Political elites are therefore elected, as it is the general belief that their ability to govern is better than the general majority's. "[It] is the political elites that *act*, whereas the citizens *react.*"[74]. The public needs to be informed in order to make the "right" decision as to who should be elected to govern. The citizens' role is to passively absorb information and leave the decision making about society to the experts since they are the most qualified.

It could be argued, that in a competitive democracy, the representatives can decide what data should be publicly available without the input of the public.

6.1.3 Legislation for the Public Interest (Offentlighedsloven)

The official title, 'Lov om offentlighed i forvaltningen', is a danish law effectuated from the 1st of January, 2014. The law regulates the public's access to documents. It was presented by the then Minister of Justice with the intent to increase public transparency and to support democratic control of the administration [75].

Despite the politician's intent to make the government more transparent, certain provisions of the law were met with a fair amount of criticism from journalists, researchers and transparency advocates as they argued the effects of the law actually minimize democratic control and aim to restrain the press. [76] [77]This shows a conflict between political agendas and the media's goals of informing the public.

We believe that the data sources we investigated in this project uphold the principles of public interest such as transparency. However, also believe they would benefit from stricter privacy controls. We assert that it is possible to both maintain the privacy of data subjects and provide data for public transparency. For further discussion regarding privacy, security and public availability turn to section 6.4.

6.2 Implicit Data Attributes

In section 5.4.1, we have outlined a few strategies for improving the data processing pipeline and increasing the linkability of data sources. We have mentioned that certain data attributes can be derived from others and provided an example of how the gender of a person can be implied from their name.

Although name and gender are often not considered sensitive information in this context of use (see section 3.2.1 for an explanation of the context of use), there are other pieces of information having a stronger impact on one's privacy that can still be implied from publicly available information. In this section, we discuss these implicit attributes.

One such example would be mortgage information from the Tinglysning [30] portal. Upon entering an address, the website presents information not only about the real estate itself but also about mortgages taken in connection to the specific real estate. If one were to save historical data of the mortgage, it is theoretically possible to compare the data to give an indication to what extent a person was able to pay off their mortgage.

Another example could be information about an employee's salary. In section 5.4, we show that there are multiple data sources providing information about a person's work title, work address, and name of the company they are working for. Combining this information with the yearly publicly accessible financial reports from Virk, one could figure out the salaries of people in certain positions in various companies [31]. It is debatable whether the salaries of employees should be publicly available, but it is worth considering since, in combination with other derived information, this piece of data can provide insight into one's personal economy.

Another example is the indication of people's political beliefs. According to the ISO 29100 and Article 9 of GDPR, this information is considered sensitive and should not be disclosed without an informed consent [12] [4]. In Borgerforslag, a citizen proposal system, it is possible to gather information on which proposals certain people support. With modern information retrieval techniques like text classification and intent analysis (reference here?), it could be possible to derive information about one's political beliefs.

6.2.1 Danish Personal Identification Numbers (CPR)

Even though the data sources do not directly disclose peoples' CPR numbers, it is, unfortunately, one of the attributes that can be derived with quite a high accuracy, given a sufficient amount of information. Date of birth and gender are the attributes which are used to determine a person's CPR number, and as we showed in section 5.4, both of these attributes can be sufficiently obtained from publicly-available data sources.

The CPR number is a ten-digit number - first six digits represent the date of birth in the format of DDMMYY and the last four digits are generated by a combination of different rules based on gender, century which the person was born in, sequential numbers based on three series and a control checksum number which was made optional in the year 2007 [78]. For a detailed description of how the Danish personal identification numbers are generated, see the official documentation [79]. While the exact CPR number of a person cannot be pinpointed based on this data, it significantly reduces the number of possibilities. A website that is used to authenticate based on a CPR number, can then be used to verify whether a CPR number is valid or not. Using this validator in combination with the strategies to reduce the number of possible CPR numbers, allows one to find authentic CPR numbers without having to spend many resources.

Because a CPR number is based on determinable data, it should not be used as a means of authentication.

6.3 Exploitability of Profiles

There are many ways that the profiles collected from public records can be abused. This discussion does not detail all the ways that profiles can be exploited, but only provides a few examples of the possible misuses.

To maximize the use of these profiles, malicious actors can use the techniques to either derive implicit data attributes as described in section 6.2, or they can leverage these profiles into gathering even more information about them from their social media. Another way to extend these profiles is to use social engineering techniques [80] to target the victims' co-workers, friends and family to reveal even more information about the victim.

One of the use cases is stalking. Malicious actors can figure out the working hours of their victims, based on their occupation and place of work. If they also know the home address of their targets, it is trivial to figure out the daily routes of their victims.

Gathering enough information allows building a sufficiently detailed profile to commit identity fraud. This allows the malicious actors to impersonate their target in the digital space. A typical example would be getting a credit card in their victim's name.

Another use for these profiles is spear phishing. Spear phishing is a targeted phishing attack customized to an individual or set of individuals, to make the target take an action such as clicking a link or opening an attachment containing malware [81]. The ultimate goal behind spear phishing differs based on who it is done. It can range from criminal organizations' committing blackmail to governmentally-funded espionage and psychological warfare to influence countries' elections [81][82]. The first step in spear phishing is to collect data about their target(s). This data is then used to customize an attack for the specific person to increase the likelihood of success. The governmental public records provide a very simple first step to collect data on a massive scale regarding a specific set of individuals such as people in a specific company or a specific field such as politicians, lawyers or doctors.

6.4 Security and Privacy Control Recommendations

Currently, information is heavily used by governments to fulfill their duties to the citizens. This information can also pose a threat to the citizen when misused.

Authentication

Authentication is a way to prove your claim of identity and thereby gives you access to the information you are entitled to. Data sources use it to limit who can access their data which is a useful tool to prevent harvesting at a large scale by crawlers.

However, if one wanted to keep data public it should not be behind a login. To keep data public and prevent harvestability at a large scale is a challenge-response test such as CAPTCHA [83] to prove that the agent behind the request is human. While this technique can be circumvented by the real-time data processing pipeline, OCR, and text-to-speech, it does prevent large scale harvesting.

In the exploratory phase of data sources, we did not encounter any type of challenge-response test that would hinder harvestability and that raises issues regarding privacy.

Anti-Scraping Techniques

One approach in protecting information from machine processing is to ensure that personal data is only represented in a human-readable format when it faces the eyes of the public. This can be achieved by converting text to lossy image formats when presented to the public through the websites and public APIs, and other channels of communication.

This approach will increase the difficulty of harvesting using current technologies as the harvester would need to involve OCR software which is unreliable and needs a great amount of configuration and training.

However, converting text to images might result in negative user experience if it makes text illegible. Using the same approach, an audio version of the text should be provided to increase the accessibility of the system. This will guard privacy and protect the PII of the principal while hindering the automatized processes of bulk information collection.

Digital Permanence

Another threat to private information is archiving projects like The Internet Archive [84] which crawl the web archiving websites and their public content periodically. Some of the data sources we have investigated disclose information that by law should not be public for more than a certain amount of time.

The presentation of this time-sensitive private information should be done in a way that hinders any automated processing and archiving tools. Archiving can not be avoided by converting the text data to another format like image or sound. Therefore the site owners should ensure that this kind of information remains on the deep web (not indexed by search engines) and not linked to from any other place. In fact not even implementing a URL for the specific information would be a great step towards avoiding archiving. Traditional archiver tools save the contents of static pages so dynamically loaded content is not archived.

Index and Search Policies

Another aspect that would decrease the disclosure of information is the implementation of strict search policies. The majority of data sources we investigated allowed for empty string or stop word searches [2] which are usually used to get as much data as possible with the least amount of requests. Hence usually data can be collected page by page starting from the results of such a search query. We used this technique to get the approximate size of the database behind the interface presenting the data.

Stricter search policies would invoke a character limit for the search queries and their effect area would be limited to single properties. In any case, cross-property search functionality should be avoided.

Privacy Assessment of the APIs

A regular assessment is an important part of any system's life-cycle and privacy should be a part of such an assessment. As an example, we can look at the case of Battery Status API [85], where a feature was implemented into web browsers based on a W3C standard that provided the battery level information to them. However, after its implementation, the feature was re-assessed and privacy issues were discovered. It was found that the feature has been misused for tracking purposes. Based on this case, the following recommendations would benefit the governmental APIs as well:

- The specification process should include a privacy review of the implementations. This could apply very well to both closed and open governmental APIs. The API specification authors and developers should conduct a review process where they examine the privacy of their implementation.

- API use in the wild should be audited after implementation. Any use or misuse by the API clients connected to the governmental API needs to be audited from a privacy standpoint.
- Specification authors should carry out privacy assessments with multiple threat models and model countermeasures for mitigation. LINDDUN methodology can be used as a guideline.

6.5 Knowledge contributions

One of the knowledge contributions produced by this project is the mapping of public bodies and publicly accessible data sources that they are responsible for. In our research, we could not find any existing resource that clearly answers this question. Registerindsigt.cpr.dk provides a way to track parts of your own digital footprint based on which private and public services you allowed to handle your CPR data. However, we believe that the government should provide a directory of the data sources and what governmental body is in charge of it. This would go hand-in-hand with the principles of open government and transparency. We think that such an overview would be a great resource for both research and awareness purposes to let people know who and how is handling their data. The methodology described in section 4.2.2 provides a possible way such a registry of data sources could look like.

Another contribution of this project is the classification of data sources. In section 4, we present a data source classification table with criteria which indicate how the data sources can connect within the context of a data pipeline. During our literature review, we were unable to find sources that would support or disagree with this approach. Whether this is due to the fact that this is an incorrect approach to the problem or the lack of academic research in this area is up for discussion. However, we believe that by building on top of harvestability research by Khelghati et al. [2] and taking into consideration more data source attributes when trying to build a data pipeline, our work can be beneficial for future research.

6.6 Future work

While we consider our data source classification scheme to be useful, we acknowledge that it requires more work. One possible improvement would be to provide a more technical specification for each of the data sources. This would provide a way to categorize each data source based on which harvesters can be used to collect data from them. This allows the developers of a data pipeline to quickly decide which harvester can be re-used, whether it needs to be modified to suit the data source or whether a new harvester has to be built from scratch.

As mentioned throughout this project, the proposed pipeline is entirely theoretical. The seemingly obvious next step would be trying to implement such a pipeline, however, one would run into the ethical issues we were trying to avoid throughout this project. A possible solution to this dilemma would be to design a pipeline which does not actually save any data and does not allow any person to see the processed data. The only output would be a description of possible connections between data sources based on the input.

7 Conclusion

As demonstrated in this project, Danish digital public records can be harvested, linked and potentially misused. The scale of such misuse is dependent on the data sources used, the number of people they contain and what type of data is stored. While industry standards and legislation define ways to handle privacy, what matters the most is the actual implementation of privacy controls.

With the growing trend of digitalization and open data, linkability is only going to be a larger issue in the future if not considered from the inception of projects. During our research, we found most of the data sources to be lacking some of the aforementioned privacy controls. While implementing such controls would not decrease linkability as such, it could reduce its impact. Unfortunately, as long as multiple data sources publish information regarding a data subject, linkability is always going to be an issue.

8 Appendices

8.1 Harvestability table

Datasource vs. Measures	Embedded Scripts	Flash	Different Data Layouts	Navigation	Multipage Data Sources	Search Policies	Indexing Policies	Bad HTML Coding Practices	CIA+ Policies	URL Redirections	Residing Data (Text)	Session Management	Query Interface Type	Non-Persistent Data Formats	Multi-frames	Harvestability factor	Harvestability score
Tinglysningen	N/A	N/A	No	Yes	N/A	No	No	N/A	No	N/A	No	N/A	N/A	No	N/A	6.25	95%
Statstidende	N/A	N/A	Yes	No	N/A	Yes	No	N/A	No	N/A	No	N/A	N/A	No	N/A	6	93%
CVR	N/A	N/A	No	Yes	N/A	No	No	N/A	No	N/A	No	N/A	N/A	No	N/A	6.25	95%
Sogn.dk	N/A	N/A	No	No	N/A	No	No	N/A	No	N/A	No	N/A	N/A	No	N/A	7	100%
sundhed.dk	N/A	N/A	No	Yes	N/A	No	No	N/A	No	N/A	No	N/A	N/A	No	N/A	6.25	95%
Borgerforslag	Yes	No	No	No	No	No	No	Yes	No	No	No	No	Yes	No	No	13	86%
Folketingets telefonbog	No	No	No	No	No	No	No	Yes	Yes	No	No	Yes	No	No	No	13	86%
Advokatnølgen	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	14	93%
Det centrale hysdyrsregister	Yes	No	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	13	86%
Forsknings-databasen	No	No	Yes	No	No	No	Yes	No	No	No	No	No	Yes	No	No	13.25	88%
Jobnet.dk	Yes	No	No	No	No	No	Yes	No	No	No	Yes	No	No	No	No	13.25	88%
Domstol.dk	No	Yes	Yes	No	No	No	No	Yes	No	No	Yes	No	Yes	Yes	No	11	73%
DK-Hostmaster	Yes	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	13.75	91%
Arbejdstilsynet	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100%
DAWA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100%
Autorisations-registret	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100%

Table 4: Harvestability of data sources

In the above table we asses the harvestability of our data sources according to the following measures:

Embedded Scripts

Embedded scripts in HTML that show/hide content, redirect, dynamically generate navigations make harvesting more difficult.

Weight: 0.75

Flash

Harvesting is nearly impossible when a website uses Flash or Applets.

Weight: 1

Different Data Layouts

If data is represented (or defined) in different ways, harvesters have to know about it in advance.

Weight: 0.5

Navigation

The result of a query should be clearly differentiated from a detail page.

Weght: 0.75

Multipage data source

To get all the details of a resource, one must make multiple requests (or go to multiple pages)

Weight: 0.5

Search policies

Limitiations on number of queries and search policies such as removal of stop words make harvesting more difficult .

Weight: 0.5

Indexing policies

Indexing only some fields of the data, and thus being able to search only by those fields make harvesting harder. For example, being able to search books only by titles and not authors.

Weight: 0.5

Bad HTML Coding Practices

Havesters rely on tags, attribues etc in HTML, if bad or inconsitent coding practices are used, harvesting is more difficult.

Weight: 0.5

CIA+ Policies

Using captcha, blocking of IP addresses, disabling of AP, ant-bot services makes harvesting harder.

Weight: 0.75

Client Side Redirection

Harvesters have harder time handling redirections started by embedded scripts.

Weight: 0.5

Residing Data (Text)

Non-structured data is more difficult for harvesters to handle.

Weight: 0.5

Session Management

If client environment changes or a session expires, harvesters cannot easily access resources.

Weight: 0.75

Query Interface Type

Different search options make harvesting more efficient.

Weight: 0.75

Non-Persistent Data Formats

Different text patterns, date patterns, boolean representations etc make harvesting more difficult.

Weight: 0.75

Multi-frames

Multi-frame pages can be difficut to distinguish which frame contains the right data.

Weight: 0.75

References

[1]E. Commission, "International Digital Economy and Society Index 2018", 2018.

[2]M. Khelghati, M. Van Keulen, and D. Hiemstra, "Designing a General Deep Web Harvester by Harvestability Factor.", in *SDSW@ ISWC*, 2014.

[3]"EUR-Lex - 32016R0679 - EN - EUR-Lex". https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298&uri=CELEX%3A32016R0679.

[4]"Data protection in the EU". https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en, 17AD.

[5]"Lov om supplerende bestemmelser til forordning om beskyttelse af fysiske personer i forbindelse med behandling af personoplysninger og om fri udveksling af sådanne oplysninger (databeskyttelsesloven) - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=201319.

[6]Datatilsynet, "The Data Protection Act". https://www.datatilsynet.dk/media/6894/danish-data-protection-act.pdf.

[7]"Offentlighedsloven - Lov om offentlighed i forvaltningen". https://www.retsinformation.dk/forms/r0710.aspx?id=152299.

[8]"ISO/IEC 29100: Information technology - Security techniques - Privacy framework", no. ISO/IEC 29100. 2011-Dec-15AD.

[9]"NIST SP 800-53: Security and Privacy Controls for Federal Information Systems and Organizations", no. NIST SP 800-53 r4. 0AD.

[10]"Risk executive (function) - Glossary — CSRC". https://csrc.nist.gov/glossary/term/risk-executive.

[11]"A stronger and more secure digital Denmark". .

[12]"ISO/IEC 27005: Information technology - Security techniques - Information security risk management", no. ISO/IEC 27005. 0AD.

[13]"NIST SP 800-37: Risk Management Framework forInformation Systems and Organizations", no. NIST SP 800-37 r2. 0AD.

[14]A. Cavoukian, "Privacy by Design: The 7 Foundational Principles: Implementation and Mapping of Fair Information Practices", 2010.

[15]D. Schuler, "Online Deliberation and Civic Intelligence. In Open government: collaboration, transparency, and participation in practice", 2010.

[16]J. Lund, "Denmark: Targeted ANPR data retention turned into mass surveillance". 2017.

[17]"The OECD Privacy Framework". .

[18]Y. Wang and A. Kobsa, "Privacy-Enhancing Technologies", *Handbook of Research on Social and Organizational Liabilities in Information Security*, 2008.

[19]V. Seničar, B. Jerman-Blažič, and T. Klobučar, "Privacy-Enhancing Technologies - approaches and development", *Computer Standards & Interfaces*, 2003.

[20]I. Goldberg, "Privacy Enhancing Technologies for the Internet III: Ten Years Later", *Digital Privacy: Theory, Technologies, and Practices*, 2007.

[21]A. Haque and S. Singh, "Anti-Scraping Application Development", 2015.

[22]K. Wuyts, *LINDDUN: a privacy threat analysis framework*. https://people.cs.kuleuven.be/~kim.wuyts/LINDDUN/LINDDUN.pdf, 2014.

[23]S. L. Michael Howard, *The Security Development Lifecycle: SDL: A Process for Developing Demonstrably More Secure Software*. Microsoft Press, 2006.

[24]"Security Development Lifecycle", in *Web Security*, Auerbach Publications, 2015, pp. 467–486.

[25]B. K. O. & C. S. B. Pernille Tranberg Gry Hasselbalch, *DATAETHICS – Principles and Guidelines for Companies, Authorities & Organisations*. Spintype.com, 2018.

[26]G. H. & P. Tranberg, *Data Ethics - The New Competitive Advantage*. Publishare / Spintype, 2016.

[27]"User IDs removed from WHOIS search results — DK Hostmaster". https://www.dk-hostmaster.dk/en/news/user-ids-removed-whois-search-results.

[28]"Mind map - Danmark". https://mm.tt/1373371043?t=9rhc1bUWMA.

[29]"Dataoversigt". https://datafordeler.dk/dataoversigt/.

[30]Justitsministeriet, "The Land Registration Court". http://www.domstol.dk/om/otherlanguages/english/thedanishjudicialsystem/landregistrationcourt/Pages/default.aspx.

[31]"What is Virk Data?". https://data.virk.dk/what-is-virk-data.

[32]"Parterne bag sundhed.dk". https://www.sundhed.dk/borger/service/om-sundheddk/om-organisationen/parterne-bag-sundheddk/.

[33]"Advokatsamfundet". https://www.advokatsamfundet.dk/Service/English/Organization.aspx.

[34]"Om CHR". https://chr.fvst.dk/chri/faces/about?_adf.ctrl-state=1dj5lgraaa_3.

[35]"Danish National Research Database". https://www.forskningsdatabasen.dk/en.

[36]"Jobnet". https://info.jobnet.dk/om-jobnet/jobnet-in-english.

[37]"Information about the Online register - Danish Patient Safety Authority". https://en.stps.dk/en/health-professionals-and-authorities/online-register-registered-health-professionals/information-about-the-online-register/.

[38]"DK Hostmaster". https://www.dk-hostmaster.dk/en.

[39]"DAWA". https://dawa.aws.dk/dok/om.

[40]"The Danish Working Environment Authority". https://at.dk/en/about-us/about-the-wea/.

[41]S. Furnell, H. Mouratidis, and G. Pernul, Eds., *Trust Privacy and Security in Digital Business*. Springer International Publishing, 2018.

[42]"Danish National Research Database / Den Danske Forskningsdatabase". https://www.forskningsdatabasen.dk/.

[43]"Tinglysningsloven - Bekendtgørelse af lov om tinglysning - retsinformation.dk". https://www.retsinformation.dk/forms/r0710.aspx?id=142900.

[44]"Statstidendeloven - Lov om Statstidende - retsinformation.dk". https://www.retsinformation.dk/forms/r0710.aspx?id=6190.

[45]"CVR-bekendtgørelsen - Bekendtgørelse om Det Centrale Virksomhedsregister og www.cvr.dk - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=164539.

[46]"Bekendtgørelse af lov om menighedsråd - retsinformation.dk". https://www.retsinformation.dk/forms/r0710.aspx?id=152415.

[47]"Bekendtgørelse af sundhedsloven - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=210110.

[48]"Folketingets Oplysning". https://www.ft.dk/da/organisation/folketingets-adminstration/folketingets-oplysning.

[49]"The Danish Data Protection Agency". https://www.datatilsynet.dk/english/.

[50]"Vejledning om databeskyttelse i forbindelse med ansættelsesforhold". https://www.datatilsynet.dk/media/6931/databeskyttelse-i-forbindelse-med-ansaettelsesforhold.pdf.

[51]"Lov om etablering af en ordning for borgerforslag med henblik på behandling i Folketinget - retsinformation.dk". https://www.retsinformation.dk/Forms/r0710.aspx?id=196897.

[52] "Behandling af persondata på borgerforslag.dk". https://www.borgerforslag.dk/persondata/.

[53] "Retsplejeloven - Bekendtgørelse af lov om rettens pleje - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=209542#id3286b241-4f99-4882-8c87-9354274ec34e.

[54] "Bekendtgørelse om godkendelse af ændringer af vedtægt for Det Danske Advokatsamfund - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=194137.

[55] "Bekendtgørelse om registrering af besætninger i CHR - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=205432.

[56] "Finanslov for finansåret 2017 - Finansministeriet". https://www.fm.dk/~/media/publikationer/imported/2017/afl17/fl17a19.ashx?la=da.

[57] "Lov om organisering og understøttelse af beskæftigelsesindsatsen m.v. - retsinformation.dk". https://www.retsinformation.dk/forms/r0710.aspx?id=167202.

[58] "Bekendtgørelse om retslister og om massemediernes aktindsigt i og opbevaring af kopier af anklageskrifter og retsmødebegæringer mv. - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=1850.

[59] "Autorisationsloven - Bekendtgørelse af lov om autorisation af sundhedspersoner og om sundhedsfaglig virksomhed - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=209811#idc0e4a1fc-0272-4002-8ad5-2dd3f3ad454c.

[60] "Domæneloven - Lov om internetdomæner - retsinformation.dk". https://www.retsinformation.dk/forms/r0710.aspx?id=161869.

[61] "Adresseloven - retsinformation.dk". https://www.retsinformation.dk/Forms/R0710.aspx?id=186325#idb85323e4-a1cb-4d2c-b03d-00cef781ade2.

[62] "Grunddata". http://grunddata.dk/english/.

[63] "API design guidance - Best practices for cloud applications". https://docs.microsoft.com/en-us/azure/architecture/best-practices/api-design.

[64] S. Clauß, "A framework for quantification of linkability within a privacy-enhancing identity management system", in *International Conference on Emerging Trends in Information and Communication Security*, 2006, pp. 191–205.

[65] L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment", *Communications of the ACM*, vol. 45(4), pp. 211–218, 2002.

[66] "Art. 4 GDPR – Definitions — General Data Protection Regulation (GDPR)". https://gdpr-info.eu/art-4-gdpr/.

[67] "Recital 51 - Protecting Sensitive Personal Data — General Data Protection Regulation (GDPR)". https://gdpr-info.eu/recitals/no-51/.

[68] D. D. P. Agency, "Ændret praksis i forhold til billeder på internettet". https://www.datatilsynet.dk/presse-og-nyheder/nyhedsarkiv/2019/sep/aendret-praksis-i-forhold-til-billeder-paa-internettet/.

[69] V. Krotov and L. Silva, "Legality and ethics of web scraping", 2018.

[70]"Personal Data Retrieval and Disambiguation in Web Person Search", *IEICE Transactions*, vol. 102-D, 2019.

[71]"Presselogen: EB's dækning af brandtragedie er over grænsen". https://nyheder.tv2.dk/2015-02-08-presselogen-ebs-daekning-af-brandtragedie-er-over-graensen.

[72]E. Hansen, "De gode offentligheder". http://offentligheder.dk/de-gode-offentligheder/.

[73]J. Elster, "Introduction", in *Deliberative Democracy*, Cambridge University Press, 1998, p. 8.

[74]J. Strömbäck, "In Search of a Standard: four models of democracy and their normative implications for journalism", *Journalism Studies*, vol. 6, no. 3, pp. 331–345, Aug. 2005.

[75]L. Barfoed, "Kronik: Moderne offentlighedslov". https://jyllands-posten.dk/debat/kronik/article4412802.ece/.

[76]T. I. Danmark, "Offentlighedsloven Transparency International Danmark". https://transparency.dk/publikationer/offentlighedsloven/.

[77]FORSKERforum, "Forskere frygter tilbageskridt i ny offentlighedslov — FORSKERforum". https://www.forskeren.dk/forskere-frygter-tilbageskridt-i-ny-offentlighedslov/.

[78]"Personnumre uden kontrolciffer (modulus 11 kontrol)". https://cpr.dk/cpr-systemet/personnumre-uden-kontrolciffer-modulus-11-kontrol/.

[79]"Personnummeret i CPR-systemet". https://cpr.dk/media/17534/personnummeret-i-cpr.pdf.

[80]K. Krombholz, H. Hobel, M. Huber, and E. Weippl, "Advanced social engineering attacks", *Journal of Information Security and Applications*, vol. 22, pp. 113–122, Jun. 2015.

[81]M. Bossetta, "The weaponization of social media: Spear phishing and cyberattacks on democracy", *Journal of International Affairs*, vol. 71, no. 1.5, pp. 97–106, 2018.

[82]F. L. Goldstein and B. F. Findley, "Psychological operations: Principles and case studies", AIR UNIV MAXWELL AFB AL, 1996.

[83]"The Official CAPTCHA Site". http://www.captcha.net/.

[84]"Internet Archive: About IA". https://archive.org/about/.

[85]L. Olejnik, S. Englehardt, and A. Narayanan, "Battery Status Not Included: Assessing Privacy in Web Standards", in *International Workshop on Privacy Engineering*, 2017.