Compiling Open Datasets in Context of Large Organizations while Protecting User Privacy and Guaranteeing Plausible Deniability

Igor Jakovljevic[1,2][a], Christian Gütl[1][b], Andreas Wagner[2][c] and Alexander Nussbaumer[1][d]
[1]*ISDS, Graz University of Technology, Graz, Austria*
[2]*CERN, Geneva, Switzerland*

Keywords: Open Data, Privacy Protection, Plausible Deniability, Open Data Framework.

Abstract: Open data and open science are terms that are becoming ever more popular. The information generated in large organizations is of great potential for organizations, future research, innovation, and more. Currently, there is a wide range of similar guidelines for publishing organizational data, focusing on data anonymization containing conflicting ideas and steps. These guidelines usually do not focus on the whole process of assessing risks, evaluating, and distributing data. In this paper, the relevant tasks from different open data frameworks have been identified, adapted, and synthesized into a six-step framework to transform organizational data into open data while offering privacy protection to organisational users. As part of the research, the framework was applied to a CERN dataset and expert interviews were conducted to evaluate the results and the framework. Drawbacks of the frameworks were identified and suggested as improvements for future work.

1 INTRODUCTION

Releasing data generated in large organizations has been a great source of information for researchers, facilitating innovation and advances in various areas, and encouraging collaboration to bring new technology, insights, and capabilities to solve problems (Navarro-Arribas et al., 2012; Zhang et al., 2020). Promoting an open and transparent approach has provided diverse benefits and competitive advantages for organizations. One example is the move of the Obama administration to increase access to government data, by launching data.gov, to increase the visibility and the legitimacy of governmental data (Van Schalkwyk and Verhulst, 2017).

Large organizations generate a median of 300 terabytes (TB) of data weekly. The data is generated from the use of various methods of communication (chat, email, face-to-face, phone, SMS, social media) between organization members, data sharing tools, internal processes, different hardware units (mobile phones, tablets, laptops, etc.), and more (Jakovljevic et al., 2020). Public services, like non-governmental organizations (NGO) or governmental organizations (GO), have recognized the benefits of open data concepts. Open data initiatives in these organizations have resulted in greater availability of data, improved efficiency and effectiveness, improve decision making, increased transparency, accountability, citizen participation in NGOs, and economic and social value creation (Loenen et al., 2020).

Open data can be used and reused without financial, legal, intellectual, and technical obstacles. The reuse of open government data will create billions of Euros in economic value (European Commission and Directorate-General for the Information Society and Media, 2002). The Open Data Institute has determined that open data should be an integral part of organizational infrastructure as a key to building the future of the urban world (Yates et al., 2018).

Besides the benefits of sharing organizational data, there are also risks and drawbacks, such as exposing sensitive and private information, if not shared appropriately (Navarro-Arribas et al., 2012; Zhang et al., 2020). An example of exposing sensitive data is the Netflix Prize. It was an open competition for the best collaborative filtering algorithm to predict user ratings for movies, based on previous ratings without any other information about the users or movies. The participants produced algorithms that improved the recommendation system by as much as 10% per year

[a] https://orcid.org/0000-0003-1893-9553
[b] https://orcid.org/0000-0001-9589-1966
[c] https://orcid.org/0000-0001-9589-2635
[d] https://orcid.org/0000-0002-4692-5741

Jakovljevic, I., Gütl, C., Wagner, A. and Nussbaumer, A.
Compiling Open Datasets in Context of Large Organizations while Protecting User Privacy and Guaranteeing Plausible Deniability.
DOI: 10.5220/0011265700003269
In *Proceedings of the 11th International Conference on Data Science, Technology and Applications (DATA 2022)*, pages 301-311
ISBN: 978-989-758-583-8; ISSN: 2184-285X

(Zhang et al., 2020). In 2007, researchers were able to identify individual users by matching the Netflix datasets with movie ratings from the Internet Movie Database (IMDB), which lead to the cancellation of the competition due to privacy issues (Narayanan and Shmatikov, 2006).

Another example of an attempt to encourage open information and collaboration that had negative consequences is related to American Online (AOL). In August of 2006, AOL, to support researchers in Information Retrieval (IR), provided a large query log from their search engine. The data represented around 650k users issuing 20 million queries. The troubling part of the data was the ease with which individuals could be identified with the logs. The result of this data release was the disclosure of private information for a number of AOL users, major reputational damage to AOL, and significant damage to the research efforts of academics who depend on such data (Adar, 2007).

Based on the previous examples it is evident that data privacy is an important aspect when it comes to sharing organizational data. It is necessary to protect persons, institutions, and organizations (Data Subjects) following laws and ethical rules during the life cycle of data (collecting data, processing and analyzing data, publishing and sharing data, preserving data, re-useing data) (Ergüner Özkoç, 2021). Many different organizations such as the European Union, PDPC Singapore, CERN, and others have created guidelines for sharing data. This research focuses on defining a framework, based on previously mentioned guidelines, for publishing organizational data as Open Data. Based on the observations stated above, more specifically, the main research questions are:

- **RQ1:** How to compile organizational datasets into open data and guarantee anonymization?
- **RQ2:** What are the benefits and drawbacks of the proposed framework for organizations?
- **RQ3:** What is the value of the data before and after applying anonymization methods?

The remainder of this paper is organized as follows: Section 2 covers the literature overview and discusses current topics in privacy-preserving data mining, open data, and possible use cases. In section 3 we propose a framework based on previous research. In Section 4, the framework is applied to a CERN dataset and the benefits and drawbacks of the framework are discussed and evaluated. The general characteristics of the CERN dataset are described in this section. Finally, we conclude the work in Section 5 with the discussion of the research questions and future works.

2 BACKGROUND AND RELATED WORK

Many institutions have recognized the need to regulate selecting data for sharing and sharing with the public. The primary goal behind these regulations is to protect individuals, organizations, and their sensitive data. Guidelines like the Federal Act on Data Protection (FADP), Personal Data Protection Act (PDPA), and General Data Protection Regulation (GDPR) have a commonality. They demand that sensitive data about individuals and organizations need anonymization to a certain degree before sharing with the public (European Commission, 2016; Personal Data Protection Commission Singapore, 2018; The Federal Assembly of the Swiss Confederation, 2019).

To make the data functional and useful, it is also crucial to find the balance between sharing too much and too little data. The Privacy-Utility trade off is based on the understanding that the more data we eliminate through anonymization, the more privacy we convey to users but the less useful that data becomes (Adar, 2007). Pseudonymization is defined as the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information (Personal Data Protection Commission Singapore, 2018). Such additional information is kept separately and subjects to technical and organizational measures to ensure non-attribution to an identified or identifiable individual. Data remain pseudonymous as long as the original identifying information is safeguarded by the publishers of the data. Various types of privacy-preserving methods, such as randomization, anonymization (k-anonymity, l-diversity, t-closeness), partition-based privacy, and differential privacy methods, are commonly used to solve the problem of deidentification. All these methods have different vulnerabilities, and researchers are continuing their research for updating them to adopt contemporary data (Ergüner Özkoç, 2021; Pramanik et al., 2021; Sousa et al., 2021).

On the other hand, if researchers share pseudonymized data without the related identifying keys, then those data are considered anonymous for the recipients. If applied properly, it may satisfy data protection requirements, since anonymized data is not considered as "personal" and therefore does not fall under the scope of data protection acts (Grace et al., 2016; Personal Data Protection Commission Singapore, 2018).

2.1 Open Data and Open Data Initiatives

Open Data is the term used to describe data available freely for anyone to use for analysis and research (Antony and Salian, 2021). There have been different initiatives for collaboration based on open data, such as the previously mentioned Netflix Prize, OpenStreetMap, CERN Open Science Initiative, Open City Initiatives, and more. All of these collaboration projects faced a common issue. Sharing of data that contains identifies, quasi-identifies, and sensitive attributes. Besides these issues, political factors such as structures, regulations, and ways of working become challenges for sharing data even within an organization (Antony and Salian, 2021; Runeson et al., 2021). Open Data Repositories (ODR) are structures, whether academic or non-academic, that host data and allow free access to them. Examples of open repositories are Zenodo, arXiv, CiteSeerX, UK Data Archive, and Figshare (Costa et al., 2021).

An important aspect of making data usable is the ability to identify, locate, and retrieve the correct data together with understanding the context of the data. This can be achieved by providing metadata about the dataset, in the form of a data dictionary or a metadata repository. Metadata is the summary information describing the data, including the availability, nature, meaning, and type of each attribute of the dataset. It provides context about the data that helps users understand their meaning. The open data initiative requires a uniform data publishing approach to ensure interoperability between different datasets Data.Gov.IE (2015); De Bie et al. (2022).

Open Data Ecosystems (ODE) is an emerging concept for data sharing under public licenses in software ecosystems. A study done by Runeson et al. (2021), interviewed 27 participants from 22 different private companies and public authorities on conceptual ideas about ODE. Their qualitative analysis of data and interviews concluded that the value of ODE lies in the data they produce and in the collaboration around the data. Furthermore, they concluded that identifying data (e.g, identifiers and quasi-identifiers) is challenging from a legal point of view, and liability issues are also unclear. Trust in the data and the governance of an ODE is also a challenge.

As seen in the previous sections, there is a need for organizations to share data and/or make them publicly available. To correctly open internal organizational data it is necessary to assess potential risks, evaluate if the data contains sensitive information, determine to which ODR to distribute the data, evaluate which license to use for data sharing, and more. When distributing sensitive information, it is important to follow guidelines, which can be given by countries, public entities, or even the organizations themselves. Most of these guidelines require a level of privacy and anonymization for sensitive data (Antony and Salian, 2021; Personal Data Protection Commission Singapore, 2018; Van Schalkwyk and Verhulst, 2017). Privacy-preserving and anonymization methods use some form of transformation on the data. Naturally, such methods reduce the granularity of representation and remove information. This results in a loss of effectiveness for data management, data processing methods, and algorithms created from this data (Adar, 2007).

2.2 Data Transformation and Anonymization Methods

Depending on the type of data, different methods for anonymization can be applied (Ergüner Özkoç, 2021; Pramanik et al., 2021). Past research indicates that the most used methods for anonymization of datasets are:

- **Randomization Methods -** add noise to data to conceal the attribute values of records. The added noise is large enough so that individual records cannot be recovered (Ergüner Özkoç, 2021).
- **Cryptographic Approaches -** are based on applying a cryptographic function over data that is presented in raw format. This raw data is also called plaintext. Applying a cryptographic function to plaintext produces cyphertext. It is hard to reproduce the original raw data, from the cyphertext. This is why it is used for anonymization of identifying data (e.g., names, addresses, etc.) (Sousa et al., 2021).
- **k-anonymity -** follows the idea that the release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals. Let $T(A_1,...,A_n)$ be a table and Q_T be the quasi-identifiers associated with it. T is said to satisfy k-anonymity if and only if for each quasi-identifier $Q \in Q_T$ each sequence of values in $T[Q]$ appears at least with k occurrences in $T[Q]$ (Samarati and Sweeney, 1998).
- **l-diversity -** is an improvement to k-anonymity and aims to mitigate possible defects of k-anonymity like homogeneity and background Knowledge attacks. In homogeneity attacks, all the values for a sensitive attribute within a group of k records are the same. Even if the data is k-anonymized, the value of the sensitive attribute for

that group of k records can be predicted exactly (Aggarwal and Yu, 2008)..

- **t-closeness -** While k-anonymity protects against identity disclosure, it does not protect in general against disclosure of a sensitive attribute corresponding to an external identified individual. t-Closeness is another extension of k-anonymity which tries to solve this issue. t-closeness requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of a sensitive attribute in the overall dataset (Li et al., 2007).
- **Partition based Privacy -** For an aggregate function $f : D \rightarrow R$, a dataset D with n records of n individual users, and a privacy preference $\phi = (\varepsilon_1, \cdots, \varepsilon_n)\,(\varepsilon_1 \leq \cdots \leq \varepsilon_n)$, where ε is the privacy parameter. Let Partition(D, ϕ, k) be a procedure that partitions the original dataset D into k partitions $(D_1 \cdots D_k)$. The partitioning mechanism is defined as $PM = B\left(DP^{f}_{\varepsilon_1}(D_1), \cdots, DP^{f}_{\varepsilon_k}(D_k)\right)$ where $DP^{f}_{\varepsilon_i}$ is any target ε_i-differentially private aggregate mechanism for f, B is an ensemble algorithm (Li et al., 2007)

2.3 Differential Privacy

Differential Privacy can be understood as a randomized function k which is applied to document collections or query results before their public release. According to Sousa et al. (2021), for all subsets S in the range of k, and a document collections D and $D^{'}$ differing on at most one element, k provides ε-differential privacy if:

$$Pr[k(D) \in S] \leq exp(\varepsilon) Pr[k(D') \in S]$$

2.4 Existing Open Data Publishing Solutions

Different governments and organizations have adopted open data principles and created guidelines for publishing such data. For instance, Open Data Handbook provides an introduction to the concept of Open Data and essential guidance for its publication. On the other side, the Open Data Ireland: Best Practice Handbook provides more detailed recommendations, additionally, it compares current international and Irish practices (Lee et al., 2014; Open Knowledge, 2015). Kučera et al. (2015) have analyzed 16 different guidelines and extracted the main processes and activities. These main processes recognized in them are the development of an open data publication plan, preparation of publication, realization of publication, and archiving. While the main activities identified are data quality management, communication management, risk management, and benefits management. Based on the analysis of different guidelines, key issues to be considered before publishing open data are:

Table 1: Main Questions for Publishing Data (Data.Gov.IE, 2015; Kučera et al., 2015; Lee et al., 2014; Open Knowledge, 2015).

Question
Which dataset should be published?
Does your organization produce, manage and maintain the dataset?
How to create datasets descriptors (recommended metadata schema) to help in the identification, location, and retrieval of online resources?
Is the data published online or offline? Is the dataset updated? How frequently?
Is anonymization and/or aggregation required?
Is there any other reason why data cannot be published (confidentiality clauses, third-party copyright, etc.)?
Can this dataset be associated with the recommended open Licence?
Can the dataset be published in a structured machine-readable format?
Does your dataset use a recognized international or national standard?

Table 2 describes the main steps in popular frameworks for open data publishing (Carrara et al., 2018; Ontario Human Rights Commission, 2022; Van Herreweghe, 2021).

Table 2: Analysis of Frameworks for Open Data Publishing.

Title	Steps
Open Data Manual Practice	1. Choose your dataset(s) 2. Choose a model licence 3. Open up your source data 4. Document your dataset(s) 5. Make your dataset(s) discoverable 6. Evaluate your Open Data Practice
What is involved in collecting data	1. Identify issues and/or opportunities for collecting data 2. Select issue(s) and/or opportunity(ies) and set goals 3. Plan an approach and methods 4. Collect data 5. Analyze and interpret data 6 Act on results

One commonality for all open data publishing frameworks is the usage of FAIR Data Principles[1]. They define a range of qualities a published dataset should have to be findable, accessible, interoperable, and reusable (Data.Gov.IE, 2015; Open Knowledge, 2015; Van Herreweghe, 2021). When publishing Open Data, international standards described by reputable organizations, such as ISO, the European Commission, W3C, IETF, OGC, and OASIS should be used for classifying data attributes and creating metadata repositories and/or data dictionaries. If international standards are inconvenient, national standards can be used (Data.Gov.IE, 2015; Lee, 2021).

2.5 Discussion

Even though various guidelines exist that illustrate the steps necessary to publish organizational data, they focus on data anonymization and contain conflicting ideas and steps. The mentioned guidelines often do not focus on the whole process of assessing risks, evaluating, and distributing data.

There is a need to develop a more generic framework for releasing organizational data as open data, by merging ideas, concepts, and steps from multiple frameworks. Based on previous research and analysis of different frameworks, instead of focusing only on data anonymization, data privacy, and sensitive information release, a generic framework should answer the following questions first: What are the risks of sharing the data? What are the benefits and drawbacks of disclosing the data? Who is going to be the gatekeeper of the data? What are the applicable data governance rules? Where will the data be distributed? From what sources was the information compiled? Are there any restrictions for the use of data?

The following section describes a step-by-step framework for the compilation of organizational data into open data.

3 ORGANIZATIONAL FRAMEWORK FOR OPEN SOURCING DATA - DATALIFT

Based on the previous section, relevant task have been identified, adapted, and synthesized into a framework to transform organizational data into open and sharable data. The main steps of this framework are: Define the Purpose And Scope of Data, Data Classification, Risk Assessment, Data Transformation and Anonymization, Evaluation, and Publishing.

3.1 Define Purpose and Scope of Data

The first step for compiling organizational data into open data is determining why and which data should be distributed and for how long. Data.Gov.IE (2015); Kučera et al. (2015); Lee et al. (2014); Open Knowledge (2015); Personal Data Protection Commission Singapore (2018) describes that answering the following questions allows to formulate a clear and concise plan with a specified purpose and scope for the data:

- Q1: What is the intent of its collection and processing?
- Q2: Which type(s) of data is being processed (e.g. machine information, user input data, user data, sensor information, etc.)?
- Q3: To which audience (public or internal organizational shareholders) will the data be distributed?
- Q4: What is the data retention period (e.g. GDPR suggestion is 6 years, indefinitely for internal data)?
- Q5: Where applicable, are there details regarding transfers of data (e.g. what are the necessary actions before moving the data to a different repository or a new governing body)?

3.2 Data Collection and Classification

In the second step, it is required to collect the data, and evaluate and classify the data based on the level of sensitive attributes. Before collecting the data it is necessary to determine which data format (eg. pdf, CSV, XML, JSON, etc.) to use and where to temporarily securely store the data before publishing. The selection of the format and storage depends on the organizational requirements. Data can be collected from multiple sources such as newly generated data or data from another internal or external source, which implies that it can be in different formats also. The data collection step focuses on aggregating data from different formats and transforming them into a single predetermined format (Data.Gov.IE, 2015; Kučera et al., 2015; Open Knowledge, 2015).

According to Data.Gov.IE (2015), metadata information such as datatypes should be assigned to data attributes. It is recommended that the value of the property is taken from a well-governed set of resource types, such as the DCMI [2]. Besides basic datatype information, metadata should include up-to-date additional information such as the context, qualities, and meaning of each attribute of the dataset (De Bie et al.,

[1] https://www.go-fair.org/fair-principles/

[2] http://dublincore.org/documents/dcmi-terms/#section-7

2022). Based on table 3, it is necessary to assign privacy classification classes, that fulfill the definitions for Identifiers (ID), Quasi-identifiers (QID), Sensitive attributes (SA), and Insensitive attributes (IA), to data attributes from the mentioned dataset..

Table 3: Privacy Data classification (Pramanik et al., 2021).

Type	Description
Identifiers (ID)	Information that uniquely and directly identifies individuals (e.g. full name, driver license, and social security number, etc.)
Quasi-identifiers (QID)	Identifiers that, combined with external data, lead to the indirect identification of an individual (e.g. gender, age, date of birth, zip code, etc.)
Sensitive attributes (SA)	Contains data that is private and sensitive to individuals, such as sickness and salary (e.g. medical records, bank records, etc)
Insensitive attributes (IA)	Contains general and non-risky data that are not covered by other attributes (e.g. web site visits, number of likes of a post, etc.)

Data containing IDs, QIDs, or SAs is classified as sensitive data and needs additional transformations for publishing, while the data with IAs does not need additional data transformation (Alexandra and Brian, 2020; Personal Data Protection Commission Singapore, 2018). The result of this step is aggregated data that has been stored in a unified predetermined format together with corresponding metadata information (in the form of a metadata repository or a data dictionary), that has also been classified based on data attributes into sensitive and non-sensitive data.

3.3 Risk Assessment

In this step, the organizational risk of disclosing data and the risk of data disclosure to the individual is assessed. According to Krotova et al. (2020), four main dimensions should be produced by answering the following questions: What are the strategic risks of releasing data? What are the economic risks of releasing data? What are the legal risks of releasing data? What are the technical risks of releasing data?

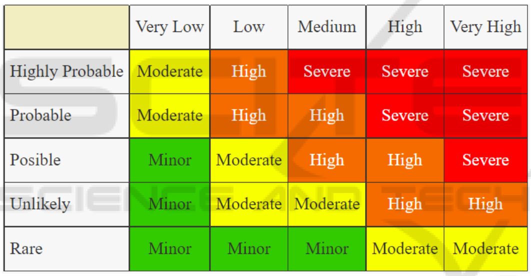
Strategic Dimension. In this step, the goal is to determine the strategic risks of releasing data. Organizations contain data of different types. Data like machine or sensor data, that do not contain any sensitive information could be freely available, without damaging the organization. However, some company data contains sensitive attributes like employee date of birth, address, personal identity numbers, and more. Table 3 describes data attribute classification based on the level of sensitive information it contains. If the data contains any ID, QID, or SA it is necessary to analyze the data from different strategic perspectives (organizational learning and growth, organizational processes, user perspectives, etc.). One such perspective is organizational reputation, for example providing data with these attributes, without any access limitations or protection, can damage the reputation of an organization or its members. Another strategic perspective is competition risk, does releasing such data put the organization at risk from competitors (e.g. abuse of methods used in an organization or user poaching) (Krotova et al., 2020; Pramanik et al., 2021).

Economic Dimension. When analyzing the data from the economic perspective, it is necessary to estimate the economic risks of sharing the data. Data produced in companies is often a byproduct or day to day workflows in organisations, which makes sharing data a low expence process. However ensuring secure usage of the correct data and data goverence can be a costly process. With the appearance of big data analytics, sharing huge data repositories free of charge, can result in a economic loss for organisations, since the produced data contains economic value (Pramanik et al., 2021; Waelbroeck, 2015). Many studies have also indicated that the development of information openness can stimulate innovative activities, the creation of innovative approaches, greater performance and greater economic benefits for organizations (Lopez-Vega et al., 2016).

Legal Dimension. Legal issues like data licensing, sensitive user information regulations (GDPR), storing and distributing regulations, and others that have legal implications have to be analyzed. The objective is to determine legal risks that arise from opening data and possible ways to mitigate them (Pramanik et al., 2021). One example is determining the usage of correct open licenses (MIT, Apache, etc.) for certain types of organizational data (primarily non-sensitive and anonymous data) and the legal risks of these licenses.

Technical and Organizational Dimension. When analyzing the data it is necessary to determine the technical and organizational implications of releasing organizational data. Organizational implications such as the question if it is necessary to invest additional staff members to maintain the dataset, how difficult is it to gather the data from different sources within the organization, what are necessary technical skills needed to release the data, are there staff members that are capable to execute the release without exposing the organization to risks (e.g. data loss and security leaks).On the other side, technical implications are how to publish the data, anonymize the data, ensure data quality, low error rate, machine readability, and continuity of access (Pramanik et al., 2021; Redman, 2022).

Quantitative Risk Rating (QRR). The next step is the QRR calculation for each of the previous dimensions. Based on Kaya (2018), begin by allocating a value for the Likelihood of the risk arising and the Severity of Injury for each risk dimension. The Likelihood takes the following values Highly Probable, Probable, Possible, Unlikely, Rare. The Severity of Injury can be Very Low, Low, Medium, High, or Very High. Based on the table from figure 1, each dimension has to be assigned a risk level ranging from Minor, Moderate, Major, or Severe.



	Very Low	Low	Medium	High	Very High
Highly Probable	Moderate	High	Severe	Severe	Severe
Probable	Moderate	High	High	Severe	Severe
Posible	Minor	Moderate	High	High	Severe
Unlikely	Minor	Moderate	Moderate	High	High
Rare	Minor	Minor	Minor	Moderate	Moderate

Figure 1: Risk Matrix (Kaya, 2018).

3.4 Data Transformation and Anonymization

This step focuses on determining which method to use for data anonymization and transformation. Table 4 aggregates the knowledge produced from the literature study and previous steps.

It is used to determine the correct methods (Grace et al., 2016; Personal Data Protection Commission Singapore, 2018; Van Schalkwyk and Verhulst, 2017). Based on the risk level and the level of data sensitivity, the anonymization method is determined. By applying the correct method we aim to ensure differential privacy for sensitive information and reduce the effort necessary to publish non-sensitive and low-risk information.

Table 4: Data Anonymisation Methods Matrix.

	Data Sensitivity Level	**Risk**
None	Non-Sensitive	Low
Randomization	Non-Sensitive / Sensitive	Low-Moderate
Cryptographic	Sensitive	Low-High
k-anonymity **l-diversity** **t-closeness**	Sensitive	Low-High
Remove	Non-Sensitive / Sensitive	High-Severe

3.5 Evaluation

After the data transformation or data anonymization methods, it is necessary to evaluate the resulting dataset. The main question to answer in this step is: Does the new dataset mitigate risks and fulfills the purpose and scope defined earlier? According to Carrara et al. (2018) for the evaluation of the data, it is necessary to review the following: check the dataset on quality, check the data on timeliness and consistency, check the dataset on the use of standards, and check the dataset on technical openness.If the resulting dataset does not mitigate risks and fulfills the purpose and scope defined earlier, it is necessary to iterate back to a previous step. This can be a return to selecting a new method for data transformation, establishing a new purpose or/and scope for the data, or reevaluating the risk. Otherwise, the data is ready for the publishing step (Kučera et al., 2015; Personal Data Protection Commission Singapore, 2018).

3.6 Publishing

Before publishing, it is necessary to prepare descriptive information about the data. It should contain a comprehensive description (e.g., sources, entities, metadata information), a self-explanatory title, privacy declaration, contact information and the information related to the scope and purpose defined in 3.1. Depending on 3.2 and 3.3, the Open Source Licence[3] needs to be be carefully selected. When adding metadata to the dataset and data attributes, the use of standardized metadata schema by public bodies such as the W3C Data Catalog Vocabulary (DCAT) is recommended to ensure the dataset respects the FAIR principles. The next step consists in determining the correct ODR for publishing and preparing data governance rules. The final step is publishing the data with all meta-information to a ODR.

[3] https://opensource.org/licenses/category

4 USE CASE STUDY

This section focuses on the application and evaluation of the framework. The goal of this chapter is to answer the following research questions: What are the benefits and drawbacks of the proposed framework for organizations? What is the value of the data before and after applying anonymization methods? The framework was applied to the CERN Mattermost Dataset. It contains information about teams, channels, message threads, user messages, user reactions, and basic user information, together with information about user connections to other users, teams, and channels.

As part of the evaluation, six expert interviews were conducted. The participants were young professionals, between the age of 25 and 29, working at CERN in the IT department as Full Stack Developers. The interviews were done in person, the participants needed to first read and analyze all steps of the framework. Then they were asked to evaluate each step by stating issues that they found, positive comments, and general feedback. They had to additionally rate how understandable each step was, with one of the following values: very understandable, understandable, neither understandable nor confusing, confusing, or very confusing. After the questionnaire, the participants were presented with how the framework was applied to the CERN dataset. Each step was discussed with the participants where they had to evaluate and express the drawbacks and benefits. At the end of the evaluation, the participants were asked to provide general feedback and statements about the framework and its use within organizations. The participants did not have disagreements regarding the attribute classification. Since all the participants were in a similar age range and profession, the results of the study could lean more to open data principles, then if the study was conduced with individuals from different professions and/or age ranges.

4.1 Define the Purpose and Scope of Data

The list bellow contains answers to questions from the first step mentioned in section 3.1.

Q1 Answer: The selected data should be used only for research purposes, mostly implementation of ML algorithms for user-channel recommendations and community dynamics analysis. Commercial application of the datasets should be forbidden.

Q2 Answer: The necessary data for processing will be generated from Mattermost. It will consist of user data such as User-Channel, User-Organisation, User-Building information. Besides the user information, channel information (creation date, number of messages, etc.) will be included in the dataset. The selected Mattermost data entities are Channel Information, Channel Member, Channel Member History, Post, Threads, Teams, Team Members, and User.

Q3 The data will be distributed to the public. The aim demographics are machine learning and recommendation systems researchers.

Q4 Answer:The data will be stored indefinitely on a Open Data Repository but the data governance will be taken care of by CERN

Q5 Answer: No details needed

Interview participants expressed that this step was understandable and provided a solid start to the process of open-sourcing data. The defined questions are clear and objective while allowing space for generalization. The participants also stated that some points made, seem more related and focused on why the organization is gathering data than on open-sourcing the data.

4.2 Data Classification

The selected entities for publishing are Channel Information, Channel Member, Channel Member History, Post, Threads, Teams, Team Members, and User. Data attributes of these entities were classified by data sensitivity level. Attributes such as Channel or Team Ids are classified as ID attributes, DisplayNames or Names attributes are classified as SA, while attributes such as TotalMsgCounts, and MentionCounts are classified as IA. Besides this classification, the data attributes will be defined by the standard W3C metadata convention. Jakovljevic et al. (2022) describes documents with the result of data attribute classification. It was also decided to store the data in JSON format since it adheres to the policies of the organization. Even though the participants stated that this step provides understandable and clear introductions on how to classify and identify private/public information, concerns were identified. Concerns like specification on what seems to be possible subjectiveness. Private and sensitive seems very tied with common sense that might vary between organizations and people that might try to apply this framework.

4.3 Risk Assessment

Based on 3.3 and examples given, the four risk dimensions have been analysed and described.

Strategic Dimension - Possible damage to the organisations reputation in case that sensitive information is leaked or user behavioral patterns are linked to individuals. Poor data documentation might also lead to extraction of organizational processes that should not be shared with the public.
Economic Dimension - In the case that identifiable information is released the organisation is financially liable for breaking privacy rules imposed by Swiss and EU regulations. In the case, that identifiable information about individuals is released, the organization might need to pay out compensation to the individuals. Depending on the licence unforeseen costs might arise
Legal Dimension - Since one of the initial constrains is that the published data cannot be used for commercial purposes, it is necessary to determine a adequate licence for the data. Because the data contains sensitive information it is necessary to publish the data according to organizational and governmental legislation, otherwise legal repercussions might incur.
Technical Dimension - To publish the data to the correct data repository it is necessary to coordinate technical collaboration with multiple organizational units and share data data between them in a secure way.
Risk Estimate - Based on the risks mentioned in the previous section and the matrix displayed in figure 1 the risk dimensions have been assigned a value for severity and probability. The average risk of disclosure of data is High.

According to the participants, this step contains understandable descriptions of all relevant dimensions required for risk assessment, with sufficient descriptions to understand the goal. It was also stated that this step was the least understandable, but also the most complex. There are whole departments dedicated to risk assessment. It was also stated that companies/organizations find grey areas with risk assessments and without strict and careful guidelines this can lead to data abuse.

4.4 Data Transformation and Anonymization

Taking into consideration the values in table 4 and the results from the previous two steps, different methods for anonymization have been used. For sensitive information such as user first names, user last names, and post message text, the removal of attributes method has been used since these attributes are at the same time sensitive and have a severe risk level. For channel display name, team name, user building name, user organizational unit name, and all sensitive ID values cryptographic methods have been used. Elements that did not contain sensitive information and that were not dates were left as is. To eliminate the possibility to detect smaller clusters of individuals and reidentify users, channels, organizational units, or buildings k-anonymity was used to ensure that these clusters are removed from the dataset.The participants identified that the risk matrix and method selection matrix provides a simple but effective way to determine which data and how to anonymize. They also recognized that other methods could be also used for anonymization and that specific cases could not be covered with the use of the previously mentioned tools.

4.5 Evaluation

After the application of anonymization methods, the dataset was evaluated to determine if there was an impact on the quality of the data and consistency compared to the original dataset. Besides these evaluations, it was also determined that the applied methods mitigate risks without eroding the purpose of releasing the dataset. Even though the participants rated this step as crucial and highly beneficial, it was suggested, to avoid any sort of bias, that some evaluation parameters should be defined during the first step. The evaluation parameters should be clearly defined and presented as a checklist or a set of specific questions to answer.

4.6 Publishing

Following the guidelines from 3.6 a data description document was created with general information about the dataset, together with metadata information about the data and data attributes. Due to the restrictions from the initial step, it was necessary to select an open license that prohibits the usage of the dataset for commercial purposes. It was decided to select the CC BY-NC-ND (Attribution NonCommercial NoDerivatives) Licence. As the ODR, it was decided to store all the data on Zenodo, since it is an integral part of CERN infrastructure and it enables easy data governance (Jakovljevic et al., 2022). The participants stated that this step contains very clear points on how to publish the data, from elements (description, title, etc.) to references of a vocabulary. They also stated that having a source to consult makes it easier to follow the whole process of open-sourcing data.

5 CONCLUSION AND FUTURE WORK

In conclusion, this research investigates various guidelines for compiling data into open-source data with a focus on organizational data, data transformation, and anonymization methods. Relevant tasks have been identified, adapted, and synthesized into a framework to transform organizational data into open and sharable data. To evaluate the newly created framework, it was applied and evaluated on CERN data. The value of the data before and after the application of the framework has been discussed. Even though creating a framework that encompasses all necessary steps needed to convert sensitive organizational information into open data is a hard task, this framework takes advantage of a diverse set of national and organizational frameworks. It provides a generic framework that can be adapted for organizational use cases easily and provides the initial solution for the generalization of the process for compiling data to open data. Based on the evaluation of the framework, more detailed descriptions of individual steps, improved methods for anonymization can be a way to improve the initial framework in the future.

REFERENCES

Adar, E. (2007). User 4xxxxx9: Anonymizing query logs.

Aggarwal, C. C. and Yu, P. S. (2008). *A General Survey of Privacy-Preserving Data Mining Models and Algorithms*, pages 11–52. Springer US, Boston, MA.

Alexandra, S. and Brian, K. (2020). Data anonymisation: legal, ethical, and strategic considerations.

Antony, S. and Salian, D. (2021). *Usability of Open Data Datasets*, pages 410–422.

Carrara, W., Enzerink, S., Oudkerk, F., Radu, C., and van Steenbergen, E. (2018). Open data goldbook for data managers and data holders - europa.

Costa, P., Cordeiro, A., and OliveiraJr, E. (2021). Comparing open data repositories. pages 60–69.

Data.Gov.IE (2015). Open data technical framework.

De Bie, T., De Raedt, L., Hernández-Orallo, J., Hoos, H. H., Smyth, P., and Williams, C. K. I. (2022). Automating data science. *Commun. ACM*, 65(3):76–87.

Ergüner Özkoç, E. (2021). *Privacy Preserving Data Mining*.

European Commission (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016.

European Commission and Directorate-General for the Information Society and Media (2002). *Commercial exploitation of Europe's public sector information : executive summary*. Publications Office.

Grace, P., Patsakis, C., Zigomitros, A., Papageorgiou, A., and Pocs, M. (2016). Operando.

Jakovljevic, I., Wagner, A., and Christia, G. (2022). Cern anonymized mattermost data.

Jakovljevic, I., Wagner, A., and Gütl, C. (2020). Open search use cases for improving information discovery and information retrieval in large and highly connected organizations.

Kaya, G. (2018). *Good risk assessment practice in hospitals*. PhD thesis.

Krotova, A., Mertens, A., and Scheufen, M. (2020). Open data and data sharing.

Kučera, J., Chlapek, D., Klmek, J., and Nečaský, M. (2015). Methodologies and best practices for open data publication. *CEUR Workshop Proceedings*, 1343:52–64.

Lee, D. (2021). Open data publication guidelines.

Lee, D., Cyganiak, R., and Decker, S. (2014). Open data ireland: Best practice handbook.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115.

Loenen, B., Welle Donker, F., Eijk, A., Tutic, D., and Alexopoulos, C. (2020). Towards an open data research ecosystem in croatia. pages 59–70.

Lopez-Vega, H., Tell, F., and Vanhaverbeke, W. (2016). Where and how to search? search paths in open innovation. *Research Policy*, 45:125–136.

Narayanan, A. and Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *CoRR*, abs/cs/0610105.

Navarro-Arribas, G., Torra, V., Erola, A., and Castellà-Roca, J. (2012). User k-anonymity for privacy preserving data mining of query logs. *Inf. Process. Manag.*, 48(3):476–487.

Ontario Human Rights Commission (2022). What is involved in collecting data – six steps to success.

Open Knowledge, editor (2015). *The open data handbook*. Open Knowledge.

Personal Data Protection Commission Singapore (2018). Guide to basic data anonymisation techniques.

Pramanik, I., Lau, R., Hossain, M., Rahoman, M., Debnath, S., Rashed, M. G., and Uddin, M. (2021). Privacy preserving big data analytics: A critical analysis of state-of-the-art. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11.

Redman, T. C. (2022). Seizing opportunity in data quality.

Runeson, P., Olsson, T., and Linåker, J. (2021). Open data ecosystems — an empirical investigation into an emerging industry collaboration concept. *Journal of Systems and Software*, 182:111088.

Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Sousa, S., Guetl, C., and Kern, R. (2021). Privacy in open search: A review of challenges and solutions.

The Federal Assembly of the Swiss Confederation (2019). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016.

Van Herreweghe, N. (2021). Open data manual practice-oriented manual for the publication and management of open data using the flemish open data platform.

Van Schalkwyk, F. and Verhulst, S. (2017). *The state of open data and open data research*.

Waelbroeck, P. (2015). The economic value of personal data: An introduction. *SSRN Electronic Journal*.

Yates, D., Keller, J., Wilson, R., and Dodds, . L. (2018). The uk's geospatial data infrastructure: Challenges and opportunities. open data institute. https://theodi.org/article/geospatial-data-infrastructure-report/.

Zhang, J., Wang, Y., Yuan, Z., and Jin, Q. (2020). Personalized real-time movie recommendation system: Practical prototype and evaluation. *Tsinghua Science and Technology*, 25:180–191.