future internet

MDPI

Article

Data Anonymization: An Experimental Evaluation Using Open-Source Tools

Joana Tomás [1], Deolinda Rasteiro [1] and Jorge Bernardino [1,2,*]

[1] Institute of Engineering of Coimbra—ISEC, Polytechnic of Coimbra, Rua Pedro Nunes, 3030-199 Coimbra, Portugal; a21230106@isec.pt (J.T.); dml@isec.pt (D.R.)
[2] Centre for Informatics and Systems of the University of Coimbra (CISUC), University of Coimbra, Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
* Correspondence: jorge@isec.pt

Abstract: In recent years, the use of personal data in marketing, scientific and medical investigation, and forecasting future trends has really increased. This information is used by the government, companies, and individuals, and should not contain any sensitive information that allows the identification of an individual. Therefore, data anonymization is essential nowadays. Data anonymization changes the original data to make it difficult to identify an individual. ARX Data Anonymization and Amnesia are two popular open-source tools that simplify this process. In this paper, we evaluate these tools in two ways: with the OSSpal methodology, and using a public dataset with the most recent tweets about the Pfizer and BioNTech vaccine. The assessment with the OSSpal methodology determines that ARX Data Anonymization has better results than Amnesia. In the experimental evaluation using the public dataset, it is possible to verify that Amnesia has some errors and limitations, but the anonymization process is simpler. Using ARX Data Anonymization, it is possible to upload big datasets and the tool does not show any error in the anonymization process. We concluded that ARX Data Anonymization is the one recommended to use in data anonymization.

Keywords: data anonymization; OSSpal methodology; ARX Data Anonymization tool; Amnesia

Citation: Tomás, J.; Rasteiro, D.; Bernardino, J. Data Anonymization: An Experimental Evaluation Using Open-Source Tools. *Future Internet* **2022**, *14*, 167. https://doi.org/10.3390/fi14060167

Academic Editors: Ji-Hwei Horng, Chia-Chen Lin, Ching-Chun Chang and Paolo Bellavista

Received: 26 April 2022
Accepted: 26 May 2022
Published: 30 May 2022

1. Introduction

The amount of data available online has increased in the past years. Personal data are used in marketing, in medical and scientific investigations, forecasting future trends, and in other cases by the government, companies, and individuals.

One of the problems is that these data can contain sensitive information that leads to individual identification and therefore be misused. For example, information such as address, age, bank account number, and many other things are considered sensitive information, and should not be exposed to people that are not allowed to access it. The issue is that no one wants their personal data available online without consent, where everyone can have access to it and use it, sometimes not for the best purposes.

To ensure the privacy of the individuals, their personal information needs to be hidden or modified to not be disclosed online. This prevents some attacks on individuals such as identity theft, money extraction, and others.

Data anonymization is mostly used by companies that collect and store data of individuals to be utilized for direct activities or to be released for non-direct activities (research, marketing, and public health). In these cases, the data need to be anonymized because if it has sensitive information, it may cause privacy threats if compromised [1].

Consequently, data anonymization is essential to solve these problems by changing the original data to hide or modify the sensitive information, preventing the original data from being disclosed online, and protecting the personal information while ensuring the utilization of the data.

The result is a dataset less accurate, where it is not possible to identify an individual from the anonymized data. One of the concerns with data anonymization is data utility loss. It is necessary to consider the number of used techniques and algorithms and the number of times each one of them is applied to the dataset. If techniques and algorithms are applied more than necessary, data utility is lost and, therefore, the dataset does not have the correct information to be used by the government, organizations, or individuals.

Committed legislation that helps people to keep their sensitive information secret from the companies has been created by some countries. In Europe, there exists the European General Data Protection Regulation (EU GDPR) [2], which is the regulation present in European Union to protect personal data. This regulation lays down rules regarding the protection of individuals with respect to personal data processing and their free movement. The protection of the individual can be performed using two processes: anonymization or pseudonymization. Anonymization is the "process of encrypting or removing personally identifiable data from datasets so that the person can no longer be identified directly or indirectly" [3]. When a person cannot be reidentified, the data are no longer considered personal data and the GDPR does not apply for further use since personal data are defined as "any information relating to an identified or identifiable natural person" [2]. Pseudonymization is defined by GDPR as the process of "personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable person" [2]. In short, this means that with anonymized data, it is not possible to identify a person anymore, which does not happen with pseudonymization, since by linking the pseudonymized information with additional information it is possible to identify an individual.

The difference between identifying information and sensitive information is a concept that is important to mention. Identifying information is information that allows personal identification just by itself, such as name, address, identification number, and others. Sensitive information is the information that an individual does not want to turn public, such as information about a disease, Social Security Number, salary, and other things [4].

In this work, it is possible to confirm that data anonymization is an important technique to protect personal data and to prevent attackers from having access to them. We assess open-source anonymization tools which help users to anonymize their information.

ARX Data Anonymization and Amnesia are the most popular tools in data anonymization. In this paper, the authors evaluated these tools using the OSSpal methodology and with a public dataset.

The OSSpal methodology has the purpose of evaluating open-source tools to help users and organizations to find the best solutions using a set of categories. In the context of this work, the categories used are functionality, operational software characteristics, support and services, documentation, software technology attributes, community and adaption, and development process.

Using the OSSpal methodology, it is possible to verify that ARX Data Anonymization has a higher score than Amnesia. Each one of the tools was experimentally evaluated using a public real dataset to perform the anonymization. The dataset has the most recent tweets about the Pfizer and BioNTech vaccine. After anonymizing this dataset with the referred tools, it is possible to conclude that ARX Data Anonymization is the best tool to use. Amnesia simplifies the anonymization process, but the tool has some errors, and the results are confusing because some solutions are not completed. ARX Data Anonymization does not show to the user all the steps that need to be performed to anonymize a dataset, but it does not have inaccuracies and the solutions at the end are simpler to visualize.

The main contributions of this work are the following:

- Data anonymization tool assessment using OSSpal methodology;
- Data anonymization tool experimental evaluation using one public dataset;
- Best tool to use according to dataset characteristics;
- Main weaknesses and difficulties in the practical use of each tool.

The remaining sections of this paper are structured as follows. Section 2 presents the related work. Section 3 describes the anonymization tools. Section 4 presents the OSSpal methodology. Section 5 presents the tool assessment using the OSSpal methodology, and Section 6 describes the assessment using a real dataset. Section 7 discusses the results of the experiments and suggests some recommendations. Finally, Section 8 concludes the paper and presents future work.

2. Related Work

This section presents some of the most important works related to anonymization techniques and anonymization tools.

In their work, Prasser, Eicher, Spengler, Bild, and Kuhn [5] extended the open-source ARX Data Anonymization Tool in a way that it could apply a full-domain generalization algorithm in different subsets. The idea was to implement horizontal and vertical partitioning strategies so that the tool could work with more flexible transformation models while preserving scalability. With these modifications, the ARX Data Anonymization Tool was able to support four different transformation methods: generalization, suppression, sampling, and micro-aggregation. The referred authors tested these modifications using six real-world datasets and compared the results with UTD Anonymization Toolbox, with the Mondrian algorithm, and against the Sánchez et al. algorithm [6,7]. The goal of this comparison is to prove that the ARX Data Anonymization Tool has better results for scalability and data utility when using local generalization with horizontal and vertical partitioning strategies enabled. In comparison with UTD Anonymization Toolbox, the ARX Data Anonymization Tool has better results in data utility and scalability. Comparing the new implementation with the Sánchez et al. algorithm [6,7], the ARX Data Anonymization Tool has better results in terms of data utility, but in terms of scalability, the Sánchez et al. algorithm [6,7] has better results. This could be justified by the fact that the new algorithm always guarantees identical records in input data but also the output data, so this algorithm is less flexible.

Gunawan and Mambo [8] realized that, with the usage of the existing data anonymization algorithms, when several modifications were applied to the database, data utility and data properties were considerably reduced. Therefore, they tried to create a new schema to anonymize data called subling suppression. The main goal of this anonymization approach is to reduce data utility lost and maintain the data properties, such as database size and the number of records. This schema is applied in two steps: in the first step, the records need to be grouped based on the adversary knowledge. In the second step, the items grouped before need to be replaced by a surrogated item, which is an item of the same category of the items in the adversary knowledge based on the hierarchy tree. To evaluate this new approach, they used a real-world dataset (BMS-WebView2), the Normalized Certainty Penalty (NCP) metric to measure the information loss, and the dissimilarity metric to measure the difference between the original database and the anonymized database so that they could compare the data property results. The results were positive for this new schema because the dissimilarity was zero. Thus, the number of records of the original database is the same as the number of records of the anonymized database, since this schema works in the selection and replacement of the items. They also have low values in the result for information loss, which means they did not lose data in the anonymization process. Another advantage they refer to in the paper is that data utility is preserved in data mining tasks.

Murthy, Bakar, Rahim, and Ramli [9] compared some anonymization techniques using the same dataset. The purpose of the study was to review the strengths and weaknesses of each one of the techniques. The techniques studied were generalization, suppression, distortion, swapping, and masking. They identify the best techniques that should be used for each one of the attributes of the dataset, considering the data type of the attribute. They verified that the masking and distortion techniques were the ideal for all types of data, suppression has the same results as masking but with more efficiency, and distortion makes

the data become unrecognized but can be reverted to original data by removing the noise. The conclusion is that many techniques can be applied to any data.

Liang and Samavi [10] formulated a Mixed Integer Linear Program (MILP) with a weighted objective function, which is a mathematical formulation for the k-anonymity algorithm focused on generalization of the attributes to suit different research uses. As MILP is a hard problem in general, they have also introduced two memory-efficient practical algorithms that can be applied to datasets with a larger number of records. These algorithms are based on the intuition that rows that are closer to each other in the dataset are likely to end up as equivalent rows. The MILP algorithm is implemented in Python. They evaluated the algorithms based on three metrics: optimality, performance, and scalability. In the tests, the authors used a dataset from IPUMS USA that contains 500,000 records of US Census data and the machines had eight CPUs and 7.2–8 GB of RAM. They anonymized three quasi-identifier attributes (Age, Sex, and ZipCode) using the optimal MILP model. With the tests performed by the authors, they concluded that the Greedy Search algorithm can achieve a similar utility with more efficient running time or better utility with less efficient run time. The Split and Carry algorithm shows better performance compared with other algorithms when the number of records is large and the number of attributes is small. They concluded that both algorithms are suitable tools for anonymization on large datasets and recommended the Split and Carry algorithm for large datasets with a small number of attributes. When the dataset has a large number of attributes, the Greedy Search algorithm is the recommended one.

The work conducted in this paper differs from the work of the other authors presented in this section because we are analyzing the different tools and comparing them in terms of usage, functionalities, number of algorithms, and other features, while the existing work analyses the ARX Data Anonymization Toolbox and presents some novel approaches regarding anonymization techniques and/or anonymization algorithms.

3. Open-Source Data Anonymization Tools

This section presents two popular open-source tools: ARX Data Anonymization [11] and Amnesia [12].

3.1. ARX Data Anonymization Tool

ARX Data Anonymization is a free open-source tool used in commercial big data analytics platforms, research projects, clinical trial data sharing, and training, with the aim of anonymizing sensitive personal data [11].

There are some algorithms used by ARX Data Anonymization. Those algorithms are: k-Anonymity, k-Map, l-Diversity, t-Closeness, δ-Disclosure privacy, β-Likeness, δ-Presence, and (ε, δ)-differential privacy.

This tool also supports different techniques, which are: global and local transformation schemes, random sampling, generalization, record, attribute, and cell suppression, micro-aggregation, top and bottom coding, and categorization.

Figure 1 shows the interface of the ARX Data Anonymization Tool.

To anonymize a dataset, some steps need to be performed by the user. First, it is required to create a project in the tool and then upload a dataset. The tool does not have any limitation about the dataset size. Afterwards, the type for each one of the attributes needs to be linked. The attribute types are insensitive, sensitive, quasi-identifying, and identifying. The next step is to link the privacy models to each sensitive attribute. This means that each sensitive attribute must be linked with one of the algorithms mentioned before. After setting the privacy models, hierarchies should be created for each one of the remaining attributes. After all these steps are performed, the anonymization can be performed in the original dataset. The tool presents the results in the "Explore results" tab. This exhibits to the user the different anonymization solutions, where each sensitive attribute has a different generalization level. The user needs to choose the planned solution and apply it to the original dataset. The tool displays, in the "Analysis utility" tab, the

anonymized dataset as well as some statistics related to it. The anonymized dataset can be downloaded in CSV format.

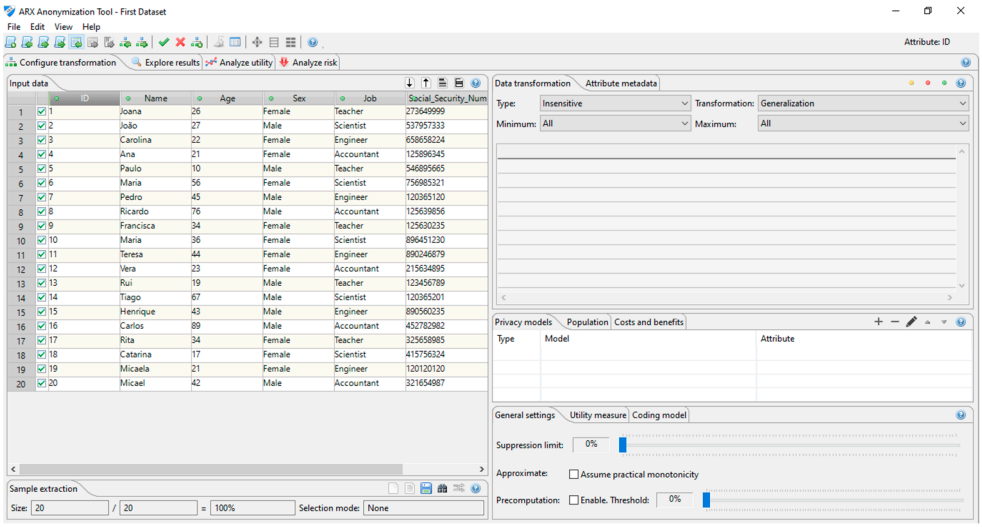

Figure 1. ARX Data Anonymization interface.

3.2. Amnesia

Amnesia is an open-source tool that follows GDPR guidelines and pseudo-anonymization. It can be used by everyone since it offers high usability and flexibility. This tool does not allow information linkage, that is, the information of the anonymized dataset is not possible to be linked to the original dataset. The tool also allows the minimal reduction of quality information, keeping the usability of the data [12].

Like ARX Data Anonymization, Amnesia uses some algorithms in the anonymization process, which are: k-anonymity, and k^m-anonymity.

Amnesia can be used online, or it can be downloaded to the desktop, and it also offers a dataset for the user to start the tool learning process.

Figure 2 shows the interface of Amnesia tool.

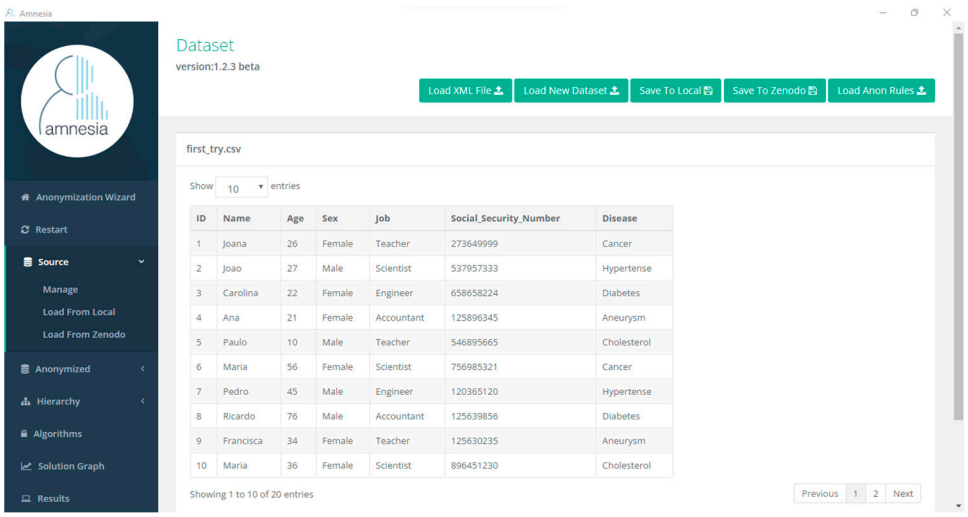

Figure 2. Amnesia interface.

The first thing to do using this tool is upload a dataset. This tool has some limitations regarding the dataset size. It only allows a dataset up to 4 MB. After uploading the dataset, the tool guides the user to hierarchy creation by clicking on a button. The hierarchies can be created for all the attributes or just for the quasi-identifier attributes.

After creating all the necessary hierarchies, the button guides the user to the algorithms. In this part, the hierarchies are linked to the attributes, and the algorithm that is going to be used for the sensitive attributes needs to be chosen, as well as the value for a k variable.

The tool, in the documentation, mentions two algorithms that can be used in the anonymization process, but it only displays one, with the name Flash. Since this is the only algorithm displayed by the tool, only the value for k can be defined.

The next step is analyzing the solution graph created by the tool. This graph displays a considerable number of solutions, with variances for the quasi-identifier and sensitive attributes, and with safe and unsafe solutions. The unsafe solutions are the ones that violate the k-anonymity application in some records. The tool mentions that these solutions can be transformed into safe solutions by clicking on the Suppress button, but every time the user tries to perform this action the tool retrieves an error. As the solution graph has so many solutions, it becomes confusing.

The next step is to choose the solution that is most suitable for the user. After choosing a solution, the anonymized dataset appears. The tool allows the download of the anonymized dataset to a CSV format.

4. OSSpal Methodology

The OSSpal methodology was created in the Business Readiness Rating (BRR) project, in 2005 [13]. This methodology aims to evaluate open-source tools using a set of categories, to help users and organizations to find the best open-source software.

The seven categories used to evaluate the software are the following [14]:

1. Functionality
2. Operational Software Characteristics
3. Support and Services
4. Documentation
5. Software Technology Attributes
6. Community and Adaption
7. Development Process

The assessment process of the OSSpal methodology for all categories, except for the functionality category, is composed of four phases [14]:

- In the **first phase**, the software components are identified and selected to be evaluated according to some criteria.
- In the **second phase**, the weights for the criteria and measures are assigned, where each criterion has a percentage. The total percentage of the sum of all criteria must be 100%. For each measure within a category, it is necessary to rank the measure following its importance and assign it.
- In the **third phase**, some data are collected to help the user to calculate the weight in a range between 1 and 5 (1—Unacceptable, 2—Poor, 3—Acceptable, 4—Very Good, 5—Excellent).
- Finally, in the **fourth phase**, the OSSpal final score is calculated.

The 'Functionality' category is calculated differently to the others. The first thing to do is choose the characteristics to evaluate, and attribute a score to them. The score of each category ranges from 1 to 3 (1—less important, 2—important, 3—most important). The categories are classified in this range. Then, the first results should be standardized on a scale from 1 to 5. This category will have the following scale:

- Under 65%—score = 1 (Unacceptable)
- 65% to 80%—score = 2 (Poor)
- 80% to 90%—score = 3 (Acceptable)
- 90% to 96%—score = 4 (Very Good)
- 96% to 100%—score = 5 (Excellent)

To assign a percentage for each category, it is important to know exactly what each one of them represents [14]:

- Functionality—this category evaluates if the software is according to the user requirements.
- Operational Software Characteristics—evaluates if the software is saved, has a reliable performance, the user interface exists or works correctly, and is easy to use, install, configure, maintain, and deploy.
- Support and Services—verifies if the software has good community or commercial support. It also evaluates if the organization gives training to help with the software usage.
- Documentation—evaluates if the existing documentation is good enough to help with the tool.
- Software Technology Attributes—with this criterion, it is verified if the software architecture is good enough, and the software is portable, extensible, open, and easy to integrate. It is also verified whether the code, design, and tests have the needed quality, if it is bug-free, and if it is complete.
- Community and Adaption—measures if the software community is active and if the component is well-accepted by the community.
- Development Process—rates how good the professionalism is at the development point and project organization.

For each category, the following percentages were defined:

- Functionality—30%
- Operational Software Characteristics—20%
- Support and Services—10%
- Documentation—10%
- Software Technology Attributes—15%
- Community and Adaption—5%
- Development Process—10%

Functionality is the category with the highest percentage (30%) because the software must comply with the user's needs. After Functionality comes Operational Software Characteristics (20%). A graphical user interface should be a mandatory requirement in a tool due to the inherent complexity of the subject. Therefore, a graphical interface is very helpful in the process of anonymizing a dataset. Furthermore, easy operation is an important requirement for each tool.

Software Technology Attributes have a value of 15% because it is important that software does not have bugs and it should be as complete as possible.

Documentation has a percentage of 10%. In these tools, especially the ones that do not have a graphical user interface, it is mandatory to have good documentation, otherwise it is difficult for the user to have the expected results from the tool.

Support and Services have the percentage of 10% because, in an open-source tool, the users do not expect much help or support from the developers and from the organization that developed the tool.

The Development Process has 10% because it is an important category in the evaluation of the tools. In Article 30 of the GDPR, some guidelines should be used by anonymization tools to agree with the regulation. A record in the tool should be maintained by a controller or its representative and should contain: the name and contacts of the controller, the controller's representative and the Data Protection Officer, the purposes of the processing, the categories of the data subject, the time limits, and the security measures [15].

Finally, Community and Adaption have the lowest percentage of all categories, with 5%. This is because we consider other categories, such as good documentation, more important.

The Functionality category needs to be classified with the help of some characteristics. Table 1 presents the characteristics and the weight that each one has for the evaluation of these tools.

The number of algorithms is important because it makes the tool more complete. The visualization of the anonymized data is also an important functionality. After anonymizing our dataset, an important thing to do is to check if the data are well-anonymized and verify some metrics after the anonymization. If the visualization of these data is not easy, it is more difficult to understand whether the results are as expected.

Table 1. Functionality Category—Characteristics and Weights.

Characteristics	Weight
Number of algorithms	2
Anonymized data visualization	1
Algorithm application	3
Anonymization process	3

The application of the algorithm to a dataset and how easy it is to anonymize a dataset are the main characteristics of these tools. The anonymization process is already too difficult, and it does not need a tool to complicate the process.

For each tool evaluation, first is given a value, between 1 and 5 for each characteristic of the Functionality category. Then, the other categories are also evaluated from 1 to 5.

In the next sections are presented the evaluation of the used tools using OSSpal methodology according to the percentages and weights specified.

5. Tool Assessment Using OSSpal

The OSSpal methodology aims to evaluate open-source tools to help users and organizations to find the best tools, using a set of categories, as defined in the previous section.

In this section, we assess the ARX Data Anonymization Tool and Amnesia using the OSSpal methodology.

5.1. ARX Data Anonymization Tool

The ARX Data Anonymization Tool was the simplest to use. It has a graphical interface that helps with the dataset upload and its anonymization. One of the biggest problems with this anonymization tool is that the help icon in some menus is not updated with the latest version of the tool and sometimes users do not have the information needed to perform some actions. It is also difficult to understand what needs to be done in some windows since the documentation is not clearly understandable.

For the Functionality criteria, the values given are represented in Table 2.

Table 2. Functionality Category—ARX Data Anonymization Tool.

Characteristics	Weight	Value
Number of algorithms	2	2
Anonymized data visualization	1	1
Algorithm application	3	3
Anonymization process	3	0

The tool has several algorithms that could be used to anonymize our dataset, but it does not allow the user to choose the algorithm s/he wants. For example, it shows a table with some algorithms linked to our sensitive attributes and other algorithms that do not have a sensitive attribute linked. If the user wants to choose one of these last algorithms, the tool does not allow it, and therefore the value 2 was given. For the anonymized data visualization, the value given was 1 because the results appear in a table side-by-side with the original dataset, so it is easy to compare both versions of the dataset (the original one and the anonymized one). The algorithm application has a value of 3 because it is easy to apply an algorithm to the sensitive attributes, and besides the hierarchy creation is not simple, the tool gives different options considering the attribute type to create the hierarchy. The anonymization process has value 0 because sometimes it is a considerable amount of information to deal with in the tool and it can be confusing instead of helping the user with the process.

Converting these values with the weight shown in Table 1, the operation that needs to be performed is as follows: $(6 \times 100)/9 = 66.7\%$, which means, in the scale for the Functionality criteria, that the value for this category is 2.

- Functionality—2 × 30% = 0.6
- Operational Software Characteristics—5 × 20% = 1
- Support and Services—4 × 10% = 0.4
- Documentation—3 × 10% = 0.3
- Software Technology Attributes—4 × 15% = 0.6
- Community and Adaption—4 × 5% = 0.2
- Development Process—0 × 10% = 0

The score given to Operational Software Characteristics is 5 due to the great graphical user interface which significantly helps users to anonymize their data. The tool is also easy to install, easy to configure, and easy to use. Support and Services and Community and Adaption, respectively, have a score of 4, because there are many articles and much information on the Internet about this tool. Some of these articles are mentioned on their website. These last points were also important to the Documentation criterion, but the way that the documentation is used in the tool is not good and it makes it difficult for users to have a valuable experience with it, which is why we define a score of 3. Software Technology Attributes have a score of 4 because the architecture seems good, and the tool does not have bugs. Finally, the Development Process has a value of 0, because the tool does not maintain any records about the controller.

The result of all criteria values is: 0.6 + 1 + 0.4 + 0.3 + 0.6 + 0.2 + 0 = 3.1.

Consequently, the result given for the ARX Data Anonymization Tool is 3.1 out of 5, between Acceptable and Very Good.

5.2. Evaluating Amnesia with OSSpal

Amnesia was not difficult to use. The bigger problem with this anonymization tool is that sometimes it takes too long to perform an action (for example, to anonymize a small and simple dataset).

For Amnesia, the values given for the Functionality category characteristics are represented in Table 3.

Table 3. Functionality Category—Amnesia.

Characteristics	Weight	Value
Number of algorithms	2	0
Anonymized data visualization	1	1
Algorithm application	3	0
Anonymization process	3	3

The tool has not many anonymization algorithms for the user to choose the best one to apply, this is way the characteristic for the number of algorithms has the value 0. For the anonymized data visualization, the value is 1 since, after conducting a study with seven attributes, the solution graph with the anonymization results was huge and it was very difficult to verify the best result for the dataset used. However, after choosing a solution, it is possible to visualize the anonymized dataset. The algorithm application has a value of 0 because of the offer given by the tool for the algorithms to apply for the anonymization process. The only step where the user could have doubts is in the hierarchy creation, but the rest of the process is easy to perform. The anonymization process has a value of 3, because sometimes it could take more time than expected, but the results are retrieved.

To calculate the value for the Functionality criteria, we need to perform the following operation: (4 × 100)/9 = 44.4%, which means, in the scale for the Functionality criteria, that the value for this category is 1.

Converting this for a 1–5 value, the result is: 0.8 + 1.2 + 2.4 + 1.8 = 6.2, and 6.2 × 5/10 = 3.1. Therefore, in a 1–5 scale, for the Functionality criteria, the given value is 3.1.

- Functionality—1 × 30% = 0.3
- Operational Software Characteristics—4 × 20% = 0.8
- Support and Services—4 × 10% = 0.4

- Documentation—3 × 10% = 0.3
- Software Technology Attributes—3 × 15% = 0.45
- Community and Adaption—4 × 5% = 0.2
- Development Process—0 × 10% = 0

The Operational Software Characteristics have a score of 4, because in this case, the tool has a graphical user interface, and the visual solution perception helps considerably in the tool usage. This tool is easy to install, easy to configure, and easy to use. The only issue is when it takes a little longer to anonymize the dataset. Support and Services, and Community and Adaption, respectively, have a score of 4, because the website has some different spaces where the user can reach the organization and it also has a Twitter account where the users could make some comments on the tool, which means that the organization want to be close to the users. The Documentation for the tool is trivial, indicating all the things that the user can make at each step in the anonymization process. It could be more helpful in the creation of hierarchy files, for example, which justifies the score of 3 on the Documentation criteria. Software Technology Attributes have a score of 3, because the architecture seems good, and the tool has a small number of bugs. However, it could be more complete with additional anonymization algorithms. Finally, the Development Process has a value of 0, because the tool does not save any record about the controller.

The result of all criteria values is: 0.3 + 0.8 + 0.4 + 0.3 + 0.45 + 0.2 + 0 = 2.45.

Therefore, the result given for Amnesia is 2.45 out of 5.

5.3. Assessment Summary

In the previous subsections, an evaluation of each open-source tool was presented. As it can be verified in each subsection, the results for each tool are the following:

- ARX Data Anonymization Tool: 3.1 out of 5
- Amnesia: 2.45 out of 5

In Table 4 are shown the values given for each category and also the final result of all values.

Furthermore, Figure 3 presents the representation of a Kiviat chart with the results for each category.

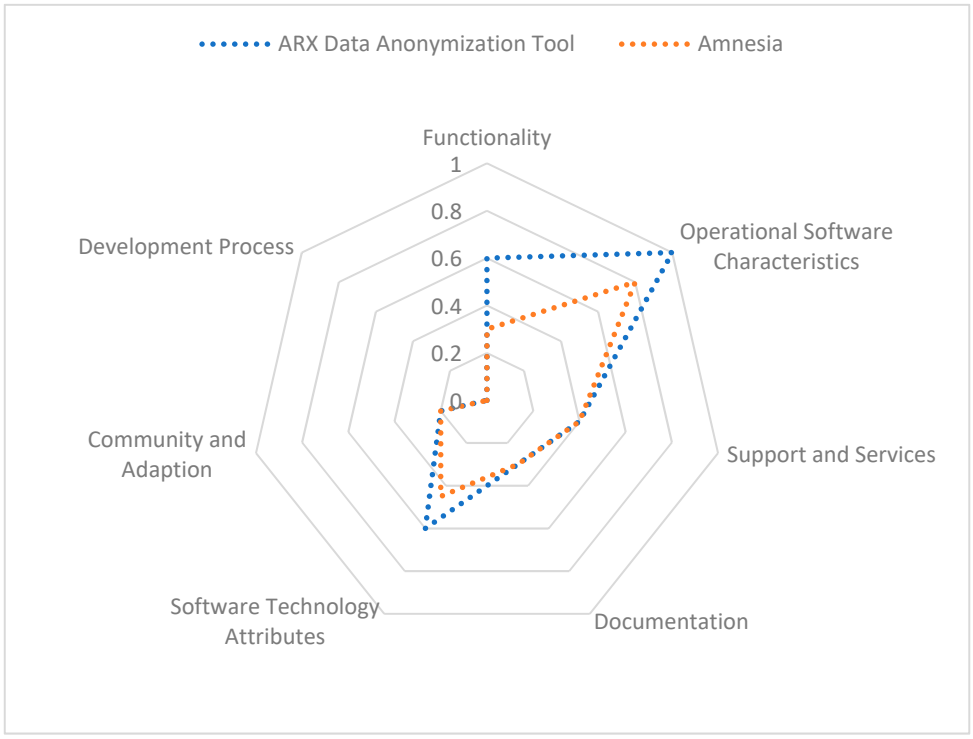

Figure 3. Kiviat chart with OSSpal results.

Table 4. Category Results.

Categories	ARX Data Anonymization	Amnesia
Functionality	0.6	0.3
Operational Software Characteristics	1	0.8
Support and Services	0.4	0.4
Documentation	0.3	0.3
Software Technology Attributes	0.6	0.45
Community and Adaption	0.2	0.2
Development Process	0	0
Total	***3.1***	*2.45*

This means that the ARX Data Anonymization Tool is the one that has the best evaluation. One thing that proves this is the fact that it is one of the most complete and most used tools. There exist a considerable number of papers and discussions all over the internet regarding this tool.

6. Tool Assessment Using a Real Dataset

In this section, we assess the ARX Data Anonymization Tool and Amnesia using one real dataset available online: Pfizer Vaccine Tweets Dataset [16].

6.1. Pfizer Vaccine Tweets Dataset

This is the most recent dataset that contains tweets about Pfizer and BioNTech vaccine. These tweets were collected using a Python package that accesses Twitter API and collects the data [16].

This dataset is composed of 16 attributes, which are the following: id, username, user-location, userdescription, usercreated, userfollowers, userfriends, userfavourites, userverified, date, text, hashtags, source, retweets, favorits, and isretweet.

The size of the dataset is 3,488,606 bytes and it has 11,021 registers. The attribute types set for this dataset are represented in Table 5.

Table 5. Pfizer Vaccine Tweets Dataset Attribute Types.

Attribute	Attribute Type
id	Sensitive
username	Identifier
userlocation	Quasi-Identifier
userdescription	Quasi-Identifier
usercreated	Non-sensitive
userfollowers	Non-sensitive
userfriends	Non-sensitive
userfavourites	Non-sensitive
userverified	Non-sensitive
date	Non-sensitive
text	Non-sensitive
hashtags	Non-sensitive
source	Non-sensitive
retweets	Non-sensitive
favorits	Non-sensitive
isretweet	Non-sensitive

6.2. Evaluating Pfizer Vaccine Tweets Dataset with ARX Data Anonymization

In this section, the Pfizer Vaccine Tweets dataset is inserted into the ARX Data Anonymization Tool. After creating the project, the dataset is inserted into the tool. The file format chosen is CSV and the delimiter is semicolon. After verifying that every data type is correct, the dataset is uploaded, and it is displayed as shown in Figure 4.

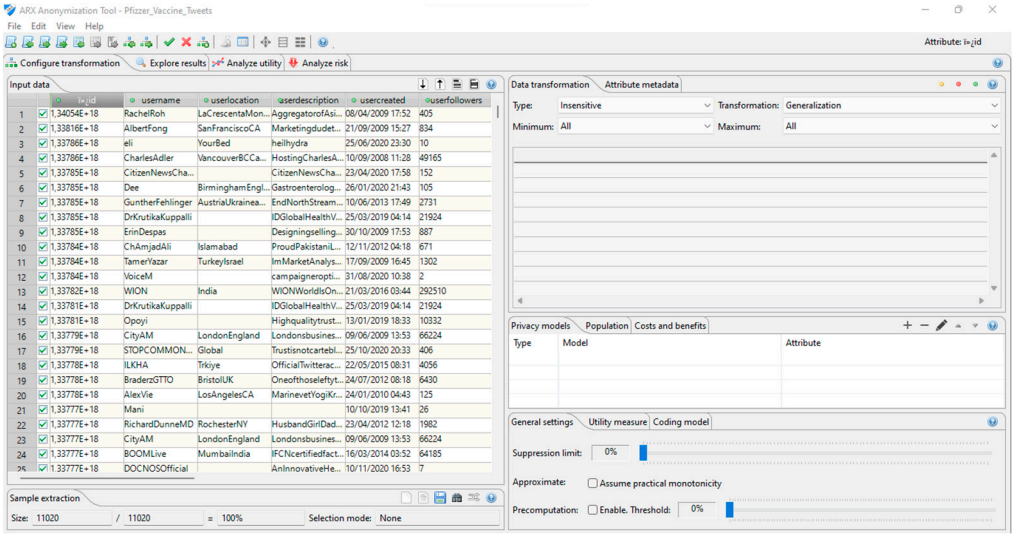

Figure 4. Dataset uploaded in ARX Data Anonymization.

Afterwards, the attribute types need to be selected according to the information in Table 4. The privacy model needs to be set for the sensitive attributes. The tool selects the privacy models l-Diversity, t-Closeness, δ-Disclosure privacy, and β-Likeness. The one chosen is the l-Diversity with an l value of 5, as shown in Figure 5. The value 5 is chosen because if the value is too high data utility is lost, and considering the size of the dataset, 5 is the appropriate value.

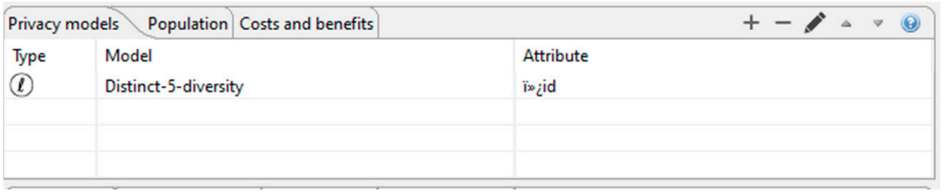

Figure 5. Privacy Model.

After that, it is necessary to create the hierarchies for the identifier and quasi-identifier attributes. For the username attribute, the hierarchy type chosen is Masking with the masking character *. The Masking type was chosen because the username attribute is a text attribute and as the purpose of this hierarchy is to hide the information, this is the correct type to use since it is going to replace the attribute values with the masking character. The Max characters value chosen is 6 because in Amnesia this is the value used. In Amnesia, this size is the number of characters that all the values for this attribute have in the hierarchy. In ARX, this is not the case, but we use the same value, and it is also used when there is a cell without a value in the dataset. The result for this hierarchy can be seen in Figure 6, where it is possible to verify that as we go up in the level, a masking character is added to the attribute value, which means that this is going to have the same number of levels as the number of characters of the biggest attribute value.

The special character asterisk (*) is the masking character chosen, which means that in each level of the hierarchy, a character of the original value is going to be replaced by this special character. As it is possible to verify in Figure 6, in the Level-1 was added a masking character, in Level-2 was added two masking characters, and it goes like this until the end of the hierarchy. The hierarchy has as many levels as the number of characters of the largest original value in the dataset for this attribute. At the last level of the hierarchy, the largest value needs to be replaced entirely with asterisks.

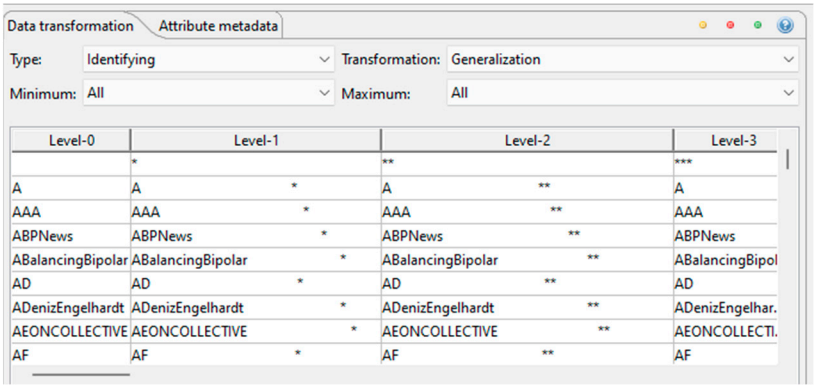

Figure 6. Hierarchy username.

The next hierarchy created is for userlocation attribute. As the type of the attribute is the same as the one for username attribute, the hierarchy was created with the same values, and the result can be seen in Figure 7.

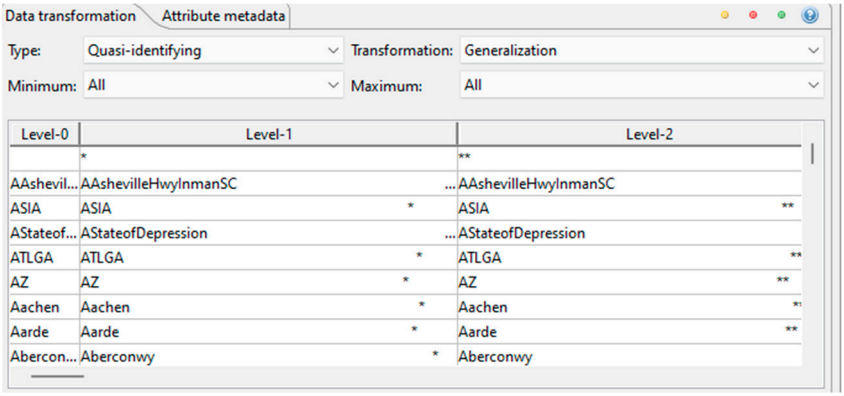

Figure 7. Hierarchy userlocation.

The special character asterisk (*) is the masking character chosen, as in username hierarchy, so, a masking character replaces also all the original characters of the values for this attribute. As it is possible to verify in Figure 7, in the Level-1 was added a masking character, in Level-2 was added two masking characters, and it goes like this until the end of the hierarchy. The hierarchy as many levels as the number of characters of the biggest original value in the dataset for this attribute. In the last level of the hierarchy, the biggest value needs to be totally replaced by asterisks.

The userdescription attribute is created as the previous ones, because it is also from the same attribute type. The result of the hierarchy is shown in Figure 8.

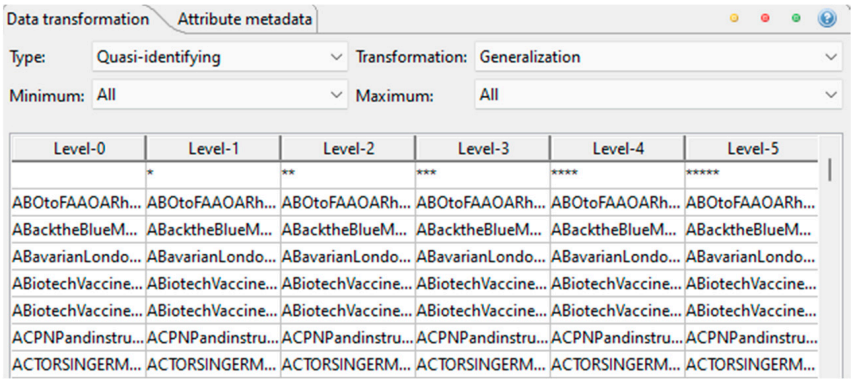

Figure 8. Hierarchy userdescription.

The special character asterisk (*) is the masking character chosen, which means that in each level of the hierarchy, a character of the original value is going to be replaced by this special character.

With all the hierarchies created and the privacy model defined, the anonymization can be performed on the original dataset. ARX only returns one set in the solution graph, as it is possible to verify in Figure 9. The values 85 and 146 correspond to the levels of the hierarchy to which the attributes were anonymized and, in this case, are the higher values in the tree since in the anonymized dataset all the characters were replaced by the masking character, as it is possible to verify in Figure 10. Thus, the userlocation attribute was anonymized to level 85 and the userdescription attribute was anonymized to level 146.

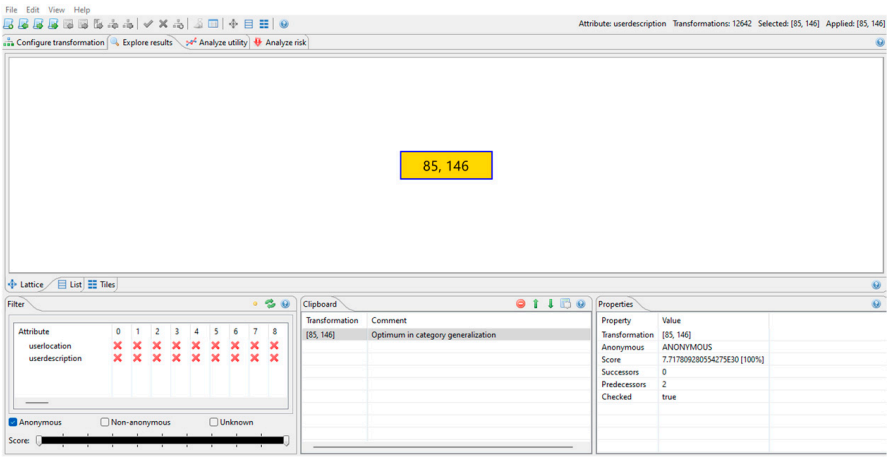

Figure 9. Anonymization graph solutions.

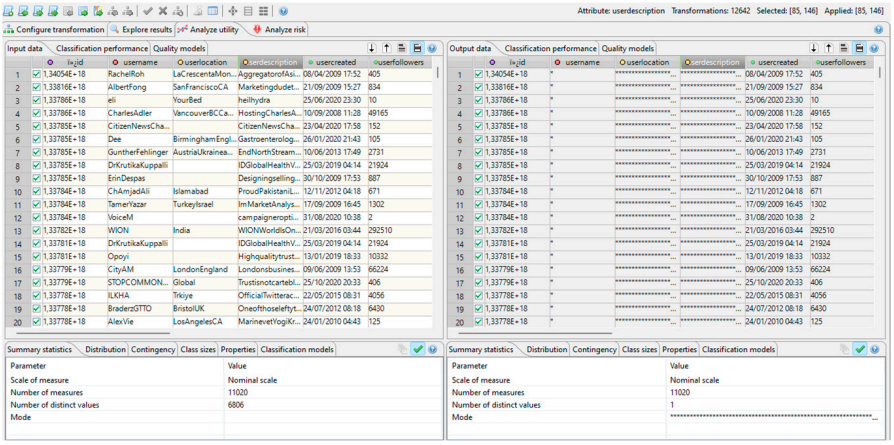

Figure 10. Anonymization Results.

Only the quasi-identifier attributes are represented in this figure, and they are 100% anonymized. After applying this information, it is possible to verify, in the Analyze utility tab, the difference between the original and the anonymized dataset. All the values from the quasi-identifier attributes were anonymized to the biggest levels of the hierarchy, as shown in Figure 10.

It should be mentioned that the anonymized dataset can be downloaded as a CSV file.

6.3. Evaluating Pfizer Vaccine Tweets Dataset with Amnesia

In this section, we explain the anonymization process for Pfizer Vaccine Tweets Dataset in Amnesia tool. To upload the CSV file, the delimiter semicolon was chosen. The display of the uploaded dataset can be seen in Figure 11.

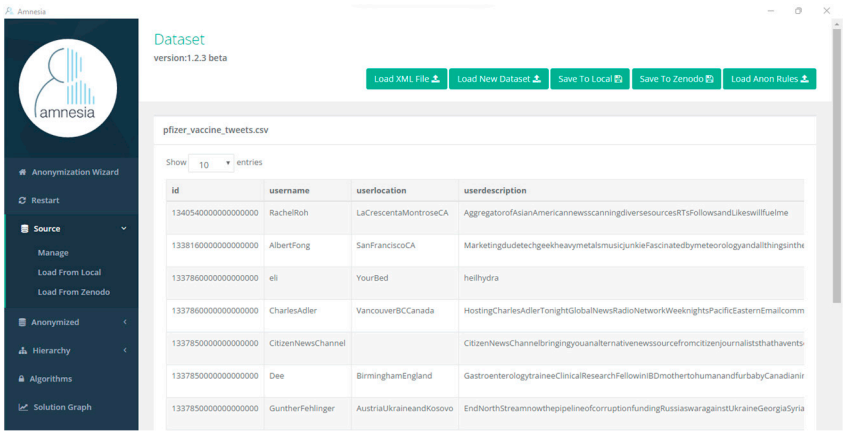

Figure 11. Dataset uploaded in Amnesia.

With the set uploaded, the hierarchies for the identifier, quasi-identifier, and sensitive attributes must be created. For the username attribute, the hierarchy type chosen is Masking Based and the characters are going to be replaced by the character '*'. The length chosen is 6, which means that all the values have 6 characters. The Masking Based type was chosen because this attribute has the string type string, and we want to hide information.

This hierarchy type allows replacement of the characters of the attribute value with the masking character, hiding the information of the attribute. A small set of this hierarchy can be seen in Figure 12.

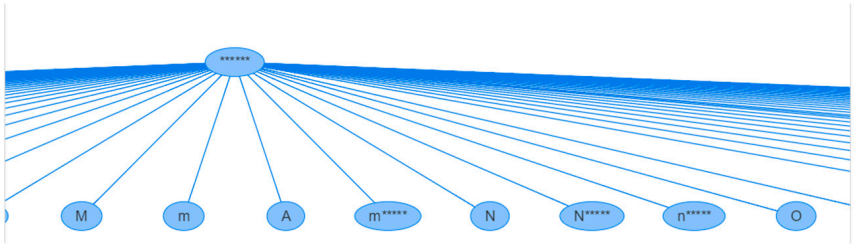

Figure 12. Hierarchy username.

As the hierarchy type chosen is the Masking Based, the characters of the original value are going to be replaced by a special character, in this case, the asterisk (*). The tool always leaves the first character unchanged, but the rest are replaced by the special character and as the length chosen is 6, when the values are modified, they get the size 6, this is why, some of the values have the first letter unchanged plus five special characters. As the username hierarchy was so large, to be able to see some of the values of the hierarchy, the Figure 12 is not complete.

For the sensitive attribute, as it is an integer, the type chosen is Range, because we want to distinguish the values divided into categorized subsets. The Domain start end limit variable is set with the interval 1,337,730,000,000,000,000–1,463,240,000,000,000,000 because these values correspond to the smaller and highest values for this attribute, respectively. The step defined is 10,000,000,000,000,000, so that there are not a huge number of values in the second level of the hierarchy tree. The result of this hierarchy can be seen in Figure 13.

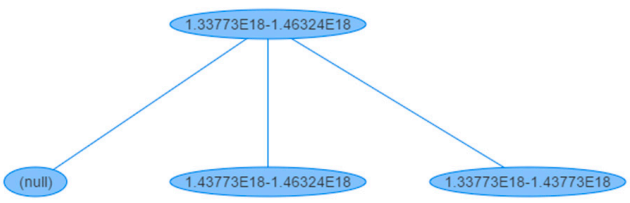

Figure 13. Hierarchy id.

The hierarchies for the userlocation and username attributes are created in the same way since the attributes have the same type. A small part of the result is shown in Figure 14.

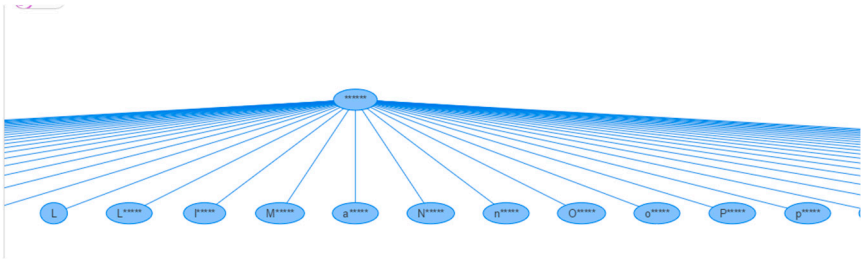

Figure 14. Hierarchy userlocation.

As the hierarchy type chosen is the Masking Based, the characters of the original value are going to be replaced by a special character, in this case, the asterisk (*). The tool always leaves the first character unchanged, but the rest are replaced by the special character and as the chosen length is 6, when the values are modified, they get the size 6, this is why some of the values have the first letter unchanged plus five special characters. Because userlocation hierarchy was so large, to be able to see some values of the hierarchy, Figure 14 is not complete.

The last hierarchy that needs to be created is for the userdescription attribute. This attribute has the same type as the userlocation and username, so the values chosen in the hierarchy creation are the same as the ones chosen in userdescription hierarchy creation. A small example of the hierarchy can be seen in Figure 15.

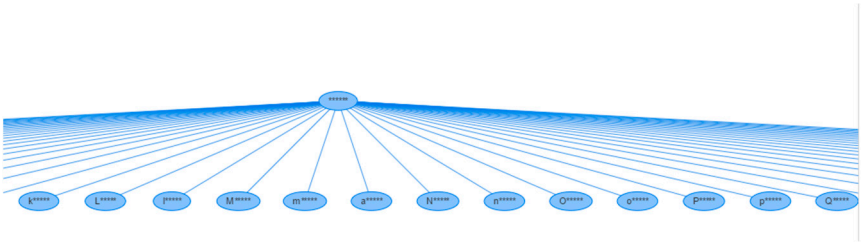

Figure 15. Hierarchy userdescription.

As the hierarchy type chosen is the Masking Based, the characters of the original value are going to be replaced by a special character, in this case, the asterisk. The tool always leaves the first character unchanged, but the remaining are replaced by the special character and as the length chosen is 6, when the values are modified, they get the size 6, this is why some of the values have the first letter unchanged plus five special characters. Because the userdescription hierarchy was so large, to be able to see some values of the hierarchy, Figure 15 is not complete.

With all the hierarchies created, the tool conducts the user to the algorithms page. In this step, the hierarchies are linked to each attribute, and the algorithm and the k value are chosen to anonymize the dataset. As was mentioned before, the only algorithm presented by the tool is "Flash", instead of the algorithms mentioned in the documentation. As the value 5 was chosen for the l value in ARX Data Anonymization Tool, this is also the value that is going to be set for the k value in Amnesia. The solution graph retrieved by the tool for these values can be seen in Figure 16.

As may be observed in Figure 16, the total solution graph is very confusing. Nevertheless, Figure 17 shows in detail the safe and the unsafe solutions retrieved by the tool for this dataset.

As can be seen in Figure 17, simply three of the solutions are safe solutions (blue ones). The first solution in the graph has the values [3, 6, 6, 6], which means that attribute id was generalized to level 3. Then, username, userlocation, and userdescription attributes were generalized to level 6.

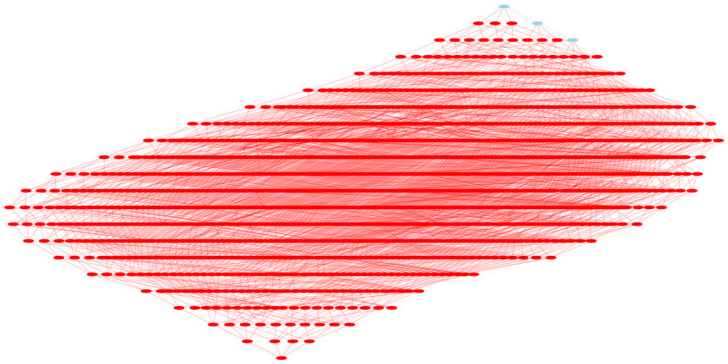

Figure 16. Solution Graph.

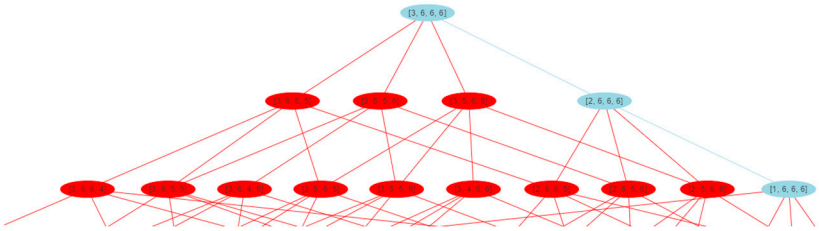

Figure 17. Safe Solutions (Blue).

The safe solution from the second level of the graph has the values [2, 6, 6, 6], which means that id was generalized to level 2, and username, userlocation, and userdescription attributes were generalized to level 6.

The safe solution with the values [1, 6, 6, 6] is the answer from the third level of the solution graph. This means that the id attribute was generalized to level 1, and username, userlocation, and userdescription attributes were generalized to level 6.

With these three solutions, it is possible to conclude that only the attribute id is generalized to distinct levels of the hierarchy, and this is the reason why Amnesia presents different safe solutions while ARX Data Anonymization only represents one solution. The preview of the anonymized dataset for the first solution can be seen in Figure 18.

Anonymized Dataset

Show 10 entries

id	username	userlocation	userdescription	usercreated	userfollowers	userfriends	userfavourites
1.33773E18-1.46324E18	******	******	******	08/04/2009 17:52	405	1692	3247
1.33773E18-1.46324E18	******	******	******	21/09/2009 15:27	834	666	178
1.33773E18-1.46324E18	******	******	******	25/06/2020 23:30	10	88	155
1.33773E18-1.46324E18	******	******	******	10/09/2008 11:28	49165	3933	21853
1.33773E18-1.46324E18	******	******	******	23/04/2020 17:58	152	580	1473
1.33773E18-1.46324E18	******	******	******	26/01/2020 21:43	105	108	106
1.33773E18-1.46324E18	******	******	******	10/06/2013 17:49	2731	5001	69344
1.33773E18-1.46324E18	******	******	******	25/03/2019 04:14	21924	593	7815
1.33773E18-1.46324E18	******	******	******	30/10/2009 17:53	887	1515	9639
1.33773E18-1.46324E18	******	******	******	12/11/2012 04:18	671	2368	20469

Showing 1 to 10 of 11,020 entries Previous 1 2 3 4 5 ... 1102

Figure 18. Anonymized Dataset.

In Figure 17 the red nodes are the unsafe solutions. An unsafe solution is a solution that violates the k-anonymity guarantee in a few records. These solutions can be transformed into safe solutions when clicking in the "Suppression" button that appears when a solution like this is chosen [12]. However, this button always retrieves an error when it could be used.

The statistic for this solution shows that the percentage of the displayed dataset is 100%, which means that all of the dataset was anonymized.

6.4. Discussion of the Results Using Pfizer Vaccine Tweets Dataset

It should be noted that the Pfizer Vaccine Tweets Dataset has a considerable size. Consequently, the Amnesia tool takes almost 5 min to anonymize the dataset and more than 3 min to display the solution graph. The time that it takes to display the solution graph is always the same every time this page is left to visualize the anonymized dataset. Compared to the ARX Anonymization Tool, Amnesia took longer because the anonymization result was displayed immediately in the first tool. The Amnesia tool also showed different results for the dataset anonymization, while ARX just retrieved one solution. This is because Amnesia retrieves a solution for each level of each sensitive, identifier, and quasi-identifier attribute, while ARX only retrieves a solution for the quasi-identifier attributes. As it is possible to verify in Amnesia results, the identifier and quasi-identifier attributes were always anonymized for the same hierarchy level in the safe solutions, so the quasi-identifier attributes only have one safe solution. The ARX Data Anonymization Tool only retrieves this solution. The variation in the Amnesia, in the safe solution, is for the id attribute, a sensitive attribute which is unchanged in the anonymization result.

The solution in Amnesia is very confused, contrary to what happens in ARX. In ARX, the solution is clearer, and it is simpler to access the result of the anonymization because it only has the results for the quasi-identifier attributes. The fact that Amnesia retrieves the solutions that violate the k-anonymity process makes it very difficult to choose a solution.

7. Discussion of the Experiments and Recommendations

The recommended tool to anonymize a dataset is ARX Data Anonymization so that the user does not need to be concerned about the dataset size, the elements in the dataset, or the complexity of the solution.

We recommend to Amnesia programmers the correction of the errors pointed out and suggest the tool's enhancement with other algorithms in order to provide more solutions to the users. Another beneficial improvement that could be performed in the tool is the addition of information loss metrics. If this were implemented in a simple way, it could bring great improvement to the tool.

In summary, the main conclusions of the tools assessment are the following:

- Applying the OSSpal methodology, it is possible to verify that ARX Data Anonymization achieves 3.1 points and Amnesia 2.45 points, where the maximum is 5. Consequently, the ARX Data Anonymization is the best tool to use according to the OSSpal methodology.
- Using the public dataset with the most recent tweets about the Pfizer and BioNTech vaccine to evaluate the tools, it is also inferred that ARX Data Anonymization was the best tool. This tool was faster in the anonymization process and the results were not ambiguous.
- According to dataset size, if it has a size bigger than 4 MB or has symbols in the text fields, Amnesia does not support the dataset. Furthermore, for datasets where the size is near 4 MB, the anonymization process is also slower in Amnesia. Additionally, the results are more difficult to understand in Amnesia compared to the ARX Data Anonymization Tool.
- The main weaknesses and difficulties in using ARX Data Anonymization are that the tool does not mention in detail the steps that need to be performed to anonymize a dataset and the user can not choose a different privacy model from the ones that are previously selected by the tool.

- For Amnesia, the main weaknesses and difficulties are the fact that the solution graph is very confusing since it has several solutions and also that it is not possible to transform unsafe solutions into safe solutions.

8. Conclusions and Future Work

In this work, we used the OSSpal methodology to assess open-source data anonymization tools. This methodology has revealed a score of 3.1 points to ARX Data Anonymization and 2.45 points to Amnesia tool, out of a total of 5 points.

We also assess the tools using a real public dataset. In Amnesia, it is simpler to follow all the steps to anonymize the data since the tool guides the user through the different views, but the tool stops various times along the process, which makes the anonymization process somewhat slow. The results obtained in the solution graph are confusing even for a small dataset, because this solution graph retrieves safe and unsafe solutions. The unsafe solutions are the ones that violate k-anonymity definition, but the tool informs the user that it is possible to transform the unsafe solutions into safe solutions. However, the Suppress button retrieves an error every time the user tries to perform this action.

The ARX Data Anonymization tool was simple to use, and it has more algorithms than Amnesia. The results are simpler to understand since the solutions retrieved are just for the quasi-identifier attributes and not for all types of attributes (quasi-identifier, identifier, and sensitive, all in the same solution). The process is not as simple to follow as in Amnesia, but every time a step is missing, the tool retrieves a warning for the user, mentioning what is missing.

In the case of the dataset size, Amnesia is very slow when the size of the dataset increases, and the anonymization process is time-consuming. In addition, it only supports datasets up to 4 MB. When uploading a dataset, the dataset cannot contain symbols in the strings, otherwise, the tool retrieves an error and does not upload the dataset. The anonymization graph has numerous solutions, which makes it very difficult to understand. Moreover, the unsafe solutions cannot be transformed into safe solutions, as mentioned in the documentation, because the tool retrieves an error. Using ARX Data Anonymization, the data are anonymized immediately.

As future work, we intend to study the ARX Data Anonymization tool with other algorithms to verify the differences in the results. Another topic that can be addressed in the future is the comparison of the information loss between the results from each tool using specific metrics that can help with these values. All of these studies will help in the decision of the best tool to use, but with the present analyses, we are able to conclude that the most complete tool to anonymize a dataset is ARX Data Anonymization.

Author Contributions: Conceptualization, J.B. and D.R.; Methodology, J.T. and J.B.; Software, J.T.; Validation, J.T., D.R. and J.B.; Formal analysis, J.T., D.R. and J.B.; Investigation, J.T.; Resources, J.T.; Data curation, J.T.; Writing—original draft preparation, J.T.; Writing—review and editing, J.T., J.B. and D.R.; Supervision, J.B. and D.R.; Project administration, J.B. and D.R.; Funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Saeed, R.; Rauf, A. Anatomization through generalization (AG): A hybrid privacy-preserving approach to prevent membership, identity and semantic similarity disclosure attacks. In Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 3–4 March 2018; pp. 1–7. [CrossRef]
2. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. Office Journal of the European Union. 2016. Available online: https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1 (accessed on 4 January 2022).

3. Anonymization and GDPR Compliance; an Overview. Available online: https://www.gdprsummary.com/anonymization-and-gdpr/ (accessed on 13 May 2022).
4. Sharma, S.; Choudhary, N.; Jain, K. A Study on Models and Techniques of Anonymization in Data Publishing. *Int. J. Sci. Res. Sci. Eng. Technol. IJSRSET* **2019**, *6*, 84–90. [CrossRef]
5. Prasser, F.; Eicher, J.; Spengler, H.; Bild, R.; Kuhn, K.A. Flexible data anonymization using ARX—Current status and challenges ahead. *Softw. Pract. Exp.* **2020**, *50*, 1277–1304. [CrossRef]
6. Sánchez, D.; Martínez, S.; Domingo-Ferrer, J. Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". *Science* **2016**, *351*, 1274. [CrossRef] [PubMed]
7. Sánchez, D.; Martínez, S.; Domingo-Ferrer, J. Supplementary materials for "How to avoid reidentification with proper anonymization"—comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". *arXiv* **2015**, arXiv:1511.05957v22015.
8. Gunawan, D.; Mambo, M. Set-valued Data Anonymization Maintaining Data Utility and Data Property. In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, Langkawi, Malaysia, 5–7 January 2018; pp. 1–8. [CrossRef]
9. Murthy, S.; Bakar, A.A.; Rahim, F.A.; Ramli, R. A Comparative Study of Data Anonymization Techniques. In Proceedings of the 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), Washington, DC, USA, 27–29 May 2019; pp. 306–309. [CrossRef]
10. Liang, Y.; Samavi, R. Optimization-based k-anonymity algorithms. *Comput. Secur.* **2020**, *93*, 101753. [CrossRef]
11. ARX—Data Anonymization Tool. Available online: https://arx.deidentifier.org/ (accessed on 15 March 2022).
12. Amnesia. Available online: https://amnesia.openaire.eu/Scenarios/AmnesiaKMAnonymityTutorial.pdf (accessed on 15 March 2022).
13. Marinheiro, A.; Bernardino, J. Experimental Evaluation of Open Source Business Intelligence Suites using OpenBRR. *IEEE Lat. Am. Trans.* **2015**, *13*, 810–817. [CrossRef]
14. Pereira, A.K.; Sousa, A.P.; Santos, J.R.; Bernardino, J. Open Source Data Mining Tools Evaluation using OSSpal Methodology. In Proceedings of the 13th International Conference on Software Technologies, Porto, Portugal, 26–28 July 2018; pp. 672–678. [CrossRef]
15. Cantiello, P.; Mastroianni, M.; Rak, M. A Conceptual Model for the General Data Protection Regulation. In Proceedings of the Computational Science and Its Applications—ICCSA, Cagliari, Italy, 13–16 September 2021; Gervasi, O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D., Apduhan, B.O., Rocha, A.M.A.C., Tarantino, E., Torre, C.M., Eds.; Springer International Publishing: Cham, Switzerland, 2021; Volume 12956, pp. 60–77. [CrossRef]
16. Pfizer Vaccine Tweets. Available online: https://www.kaggle.com/gpreda/pfizer-vaccine-tweets (accessed on 16 March 2022).