Ref. Ares(2021)2243089 - 31/03/2021

RAYUELA
a fun way to fight cybercrime

Deliverable Report

D4.9 Open Data Considerations Report

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 882828. The sole responsibility for the content of this document lies with the author and in no way reflects the views of the European Union.

Document Contributors

Deliverable No.	4.9	Work Package No.	WP 4	Task/s No.	T4.4
Work Package Title		ONLINE PRIVACY, DATA SECURITY AND ETHICS			
Linked Task/s Title					
Status		Draft	(Draft/Draft Final/Final)		
Dissemination level		PU-Public	(PU-Public, PP, RE-Restricted, CO-Confidential)		
Due date deliverable		31.03.2021	Submission date	31.03.2021	
Deliverable version		V2.0			
Deliverable responsible		TILDE			

Contributors	Organization	Reviewers	Organization
Roberts Rozis	Tilde	Stephano Cherouvis	EA
Indra Sāmīte	Tilde	Gregorio López	COMILLAS
Aivars Bērziņš	Tilde	Violeta Vázquez	ZABALA
Pieter Gryffroy	TIMELEX		

Document History

Version	Date	Comment
1.0	20.03.2021	Stable draft, ready for review
1.5	30.03.2021	Deliverable have been evaluated by experts
2.0	31.03.2021	Final version of deliverable

Table of Contents

List of Abbreviations

Abbreviation	Description
CC	Creative Commons
DMP	Data Management Plan
EU	European Union
GDPR	General Data Protection Regulation
IPR	Intellectual property rights

1. Executive Summary

This deliverable is part of the activities carried out in task 4.4. *Open Data considerations for the project's research and development framework*. The objective of this task is to identify resources developed in the scope of this project which can be delivered as Open Data to the Open Data portal. It should be reminded that the Grant Agreement of RAYUELA does not contain the general obligation to openly share all research data. Nonetheless, the RAYUELA project will endeavour to share as much data as is possible, taking into account legal and ethical restrictions.

The Consortium will ensure that the data delivered to the Open Data portal are GDPR compliant, does not contain any direct identifiers like person names, addresses etc., and are IPR free. This validation will ensure that the data developed for RAYUELA can be reused in further research.

In this deliverable, we look into key aspects that may determine when data qualify as open data. The main purpose of this deliverable is to provide general guidelines common to the opening of data and to provide a simple checklist to ensure that data generated by the project can be considered open, do not contain personal data and comply with the requirements and good practices of open data. The document is structured so that in the beginning we look at the key characteristics of open data (Section 2), in Section 3 we outline main considerations for compliance with the General Data Protection Regulation (GDPR). This section is followed by validation guidelines and checklist for open data compliance (Section 4). In the first months of the project, we conducted a survey among the partners to understand what categories of data will be processed during project implantation, Section 5 describe procedure and initial results of this survey.

2. What is Open Data

When discussing open data, one should begin with a definition. The Open knowledge foundation describes that in a single phrase "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)."[1] In essence, open means easily accessible, without any legal restrictions, and computer readable. Internationally agreed basic criteria consistent with open data are open license, accessibility, machine readable, and open format.

2.1. Open License or Status

The data must be provided under an open license. Open licence means that it allows free use of the licensed data. The license must allow redistribution of the licensed data, including sale, whether on its own or as part of a collection made from data from different sources. Derivatives of the licensed data must not be forbidden, and licence terms must allow the distribution of such derivatives under the same terms as the original license. The license must allow any part of the data to be freely used, distributed, or modified separately. The license must allow use, redistribution, modification, and compilation for any purpose. The license must not restrict

[1] http://opendefinition.org/

anyone from making use of the data in a specific field of endeavour. The license must not impose any fee arrangement, royalty, or other compensation or monetary remuneration as part of its conditions.

2.2. ACCESS

The data must be published in full, downloadable from the Internet free of charge. If data is made available in a specific format, a fee of no more than a reasonable one-time production cost may be applied. Any additional information necessary for license compliance (such as names of contributors required for compliance with attribution requirements) must also accompany the data.

2.3. MACHINE READABILITY

The data must be available in a machine-readable form(at) that can be easily processed by a computer and in which individual elements of the work can be accessed and modified.
Publisher of data to be shared should have a good understanding of data formats which are good and not good for reuse and machine readability. It is critical to discern between Machine Readable[2] and Digitally Accessible documents and formats. *Machine readable* is not synonymous with *digitally accessible*.[3] Many Digitally Accessible formats are good for easy access and publishing online but not good for Machine Readability and reuse. Table 1 below shows some examples of data and respective reusable and non-reusable data formats:

Table 1. Reusable and Non-Reusable Formats Examples of Misc. Data Types

Data Type	Non-Reusable Data	Reusable Data
Textual data Language Corpora	PDF. Although software can open and view contents of PDF, the contents of the PDF files are not machine-reusable.	Provide original documents, source texts or data in the Data Set in industry standard data formats (XML / CSV / TXT / RDF / JSON etc.). Include PDF as an optional secondary format.
Video with subtitles	Subtitles included as part of the video	Subtitles separated from Video in SRT or similar format allowing subtitles to be translated or reused separately from the Video, eg., searched.
Technical Projects	PDF is good for viewing the projects but not for reuse and modification	Include both PDF (for easy preview) and DGW (AutoCAD project) or similar source files for modification / reuse in the Open Data Package.
Users' Feedback / Data Summaries	Plaintext notes. Non-structured text.	Structured data tables with categorized and structured questions/responses data uniform per entire Data Set in a respective XML / CSV / database or a similar format.

When possible, rather include multiple formats of the same data – visual (PDF) and structured (XML / textual / annotated) instead of a single non-structured format in the package of the Data Set.

[2] https://en.wikipedia.org/wiki/Machine-readable_document
[3] https://en.wikipedia.org/wiki/Machine-readable_data

2.4. OPEN FORMAT

The work must be provided in an open format. An open format is a file format that is platform-independent and which places no restrictions, monetary or otherwise, upon its use and can be fully processed with at least one free/libre/open-source software tool.

3. Legal (GDPR) and Ethical Considerations

When assessing whether a Data Set can be opened for sharing, it must be assessed whether legal or ethical restrictions exist that prevent that data from being shared. Legal restrictions may exist in the terms of IP rights from a third party or the data partner providing the data, general legal limits imposed by laws, security or confidentiality obligations or data protection limits imposed by GDPR, amongst other things. Ethical constraints may exist where sharing of this data could lead to misuse or abuse of the data and research results or where this could have a negative impact on data subjects, just to name a few.
Admittedly, one of the main considerations is compliance with the General Data Protection Regulation (GDPR).
According to GDPR (Article 4), 'personal data' is any information relating to an identified or identifiable individual ('data subject'); an identifiable individual is one who can be identified, directly or indirectly, in particular by reference to an identifier as listed below or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person. An identifier that links the information to a specific person (data subject) is a key factor to measure whether data can be considered personal data.
Personal data may be:

- name and surname (e.g. John Doe, Jane Doe);
- workplace (e.g. John Doe, Jane Doe works in "X");
- position to be held (e.g. Director of "X");
- address (e.g., the declared address of John Doe, Jane Doe is the Road of Troways 15, Brussels);
- e-mail address (e.g. name.surname@company.com);
- identity number, ID card number, passport number, driving licence number;
- location data (e.g. location data function on mobile phone);
- Internet Protocol (IP) address;
- cookie identification number;
- medical records data on the patient, etc., are stored in the medical treatment institution.

Company registration numbers, general email addresses such as info@company.com, or anonymized data are not considered personal data.
Many of the above items are direct identifiers, i.e. pieces of personal data that may directly point towards a specific identifiable person. However, personal data as a concept is broader and also when no direct identifiers are present, the information may still be personal data. This would for example be the case when

the record of an interview does not contain any contact details (name, surname, email, address) or any other information that may directly help identify the person (workplace, position, ID number, etc.). Still, the content of the information clearly relates to one single person and given sufficient additional information may lead to that person being identified. In RAYUELA as a project, removing direct identifiers through pseudonymization will generally be used as a minimum measure to protect the privacy of the data subjects. However, when sharing data openly additional measures may be needed as the data is leaving the protected research context of RAYUELA and being openly shared with a larger public, where the RAYUELA partners no longer have control over what happens with the data. It is therefore expected that anonymization will be required in most (if not all) cases. If anonymization is not possible, it may lead to the data being unable to being shared.

Whenever openly sharing data in RAYUELA, Data Sets that contain personal data will always be processed to remove any direct identifiers and to pseudonymize the Data Sets completely. This means that only RAYUELA researchers will be holding the data that allows to directly identify the person. However, such data is still personal data and hence it will need to be considered whether such data can be shared under GDPR. In addition, GDPR imposes general limits of data minimisation and supports the processing for research purposes which does not permit the identification of the data subject. Most likely, therefore, the pseudonymized Data Set, if able to be shared at all under GDPR, will have to undergo further measures of reduction of unnecessary data items (for future research purposes) and further implementation of safeguards to prevent identification of data subjects by any third party which may use the open data in the future. One may also consider that that also third parties with malicious intent or third parties with other purposes than research purposes may access the open data. Taking the example of the interviews, it is generally not expected to be compliant to share a full record, because even absent direct identifiers, identification of the research subject may still be possible, if sufficient additional data is collected or held by the third party acquiring this data as open data. In such a case, it might be more appropriate to aggregate the data before sharing or only share high-level insights rather than the interview records themselves. Full anonymization of the records might be an option, depending on whether this can practically be done while retaining some usefulness. This shows the balance between wanting to share as much data openly while taking the rights of the research subjects duly into consideration. If there is conflict between the two, the rights and interests of the data subject must take preference. Before a Data Set which contains or might contain personal data can be shared, the compliance must be verified with legal partner **Timelex**.

Whenever GDPR rules prevent data from being shared or indicate serious concerns despite certain measures (de-identification, aggregation, etc.) already being implemented, the RAYUELA partners will assess to what extent data can be processed to become anonymous. It can be expected that this will be the case for most (or all) of the Data Sets initially containing personal data considered for being shared openly. Since anonymous data does no longer qualify as personal data, GDPR does no longer apply. In that way anonymous data would pass the GDPR check and could openly be shared.

Moreover, ethical considerations may also plead for limiting the sharing of Data Sets and for processing them to include less information or to anonymize them completely. However, even anonymous data may still raise ethical concerns in theory, so Data Sets must pass both the legal check, including GDPR, as well as the ethical check, before being able to be shared.

4. Validation Guidelines

Upon publishing of a Data Set as Open Data, the critical areas of validation are

- Quality technical aspects of the Data Set;
- Legal and technical considerations specific to Open Data;
- Metadata relating to the Data Set.

This section lists the Validation Guidelines for checking and ensuring that the dataset conforms with Open Data requirements. The Guidelines consist of Data Set validation, Critical Requirements and additional optional yet strongly recommended specific additional conditions for re-use to be taken into consideration. If a dataset fails against a specific criterion of the Guidelines, the specific criterion will suggest actions to prevent the cause of the failure of the validation of the dataset against this criterion.

4.1. Validation of the Data Set

Series of technical validations to be performed to the data itself to ensure that the data quality is up to the appropriate standard.

- Contents of the Data Set – data files – meaningful file names in ASCII character set;
- Standard compression methods to be used and checked if used;
- Conformance of the Data Set to an open and machine readable format which the domain and Data Set type matches the best;
- Simple Q/A of the data specific to the type of the data:
  - Textual data - should not contain typos, spell checked and proofed with natural language processing tools;
  - Images – they should be of decent quality and of a meaningful size to host it in a repository;
  - Geospatial data – data should be accurate and verified thoroughly;
  - Audio – clear sound, trimmed, normalized, no background noise etc.;
  - Statistical data – data tables should go in pair with matching and included comments and charts.
- Presence of a 'Read Me' file describing the contents and the background of the Data Set, containing the license attached to the Data Set (see 4.2.2 License for more details) and all additional references and information.

4.2. Critical Requirements

4.2.1. GDPR Compliance

As described in Section 3 Legal (GDPR) and Ethical Considerations, the dataset must be assessed in terms of GDPR, legal and ethical compliance.

The owner of the dataset may need to check the data and its usage in / out of context and may need to consult an expert in case of doubt. Before moving to the next steps, we strongly recommend that the owner's

Data Protection Officer confirms that the datasets are compliant with the GDPR and any other applicable EU member state data protection provisions. Data owners also have to verify this compliance with legal partner Timelex.

4.2.2. License

As described in Section 2.1 Open License or Status, upon making the dataset public and sharing it, a license must be attached to it either by the compiler of the data or the owner specifying the conditions of data use and reuse. Any license or custom license grant statement is possible, however for open data the more relaxed the license the better. Creative Commons licenses are a good option. Public Domain, CC-0 will mean free data. CC-BY is the other best alternative, and similar CC- derived licenses are additional alternatives. The license assigned should not hinder data use and reuse possibilities. To choose the most appropriate CC licence you can use the License Chooser, a tool developed by Creative Commons, a non-profit organization that helps overcome legal obstacles to the sharing of knowledge and creativity. The tool can be accessed through following link: Choose a License (creativecommons.org)[4]

4.2.3. Machine Readable Format of Reusable Data

As stated above, as an owner or provider of the data, you should have fair enough knowledge about the standards which your Data Set should conform to. It should be digital and machine readable and reusable data in a digital and structured file format, encoding and industry standard which the potential user of the data anticipates to receive it.

[4] https://creativecommons.org/choose/

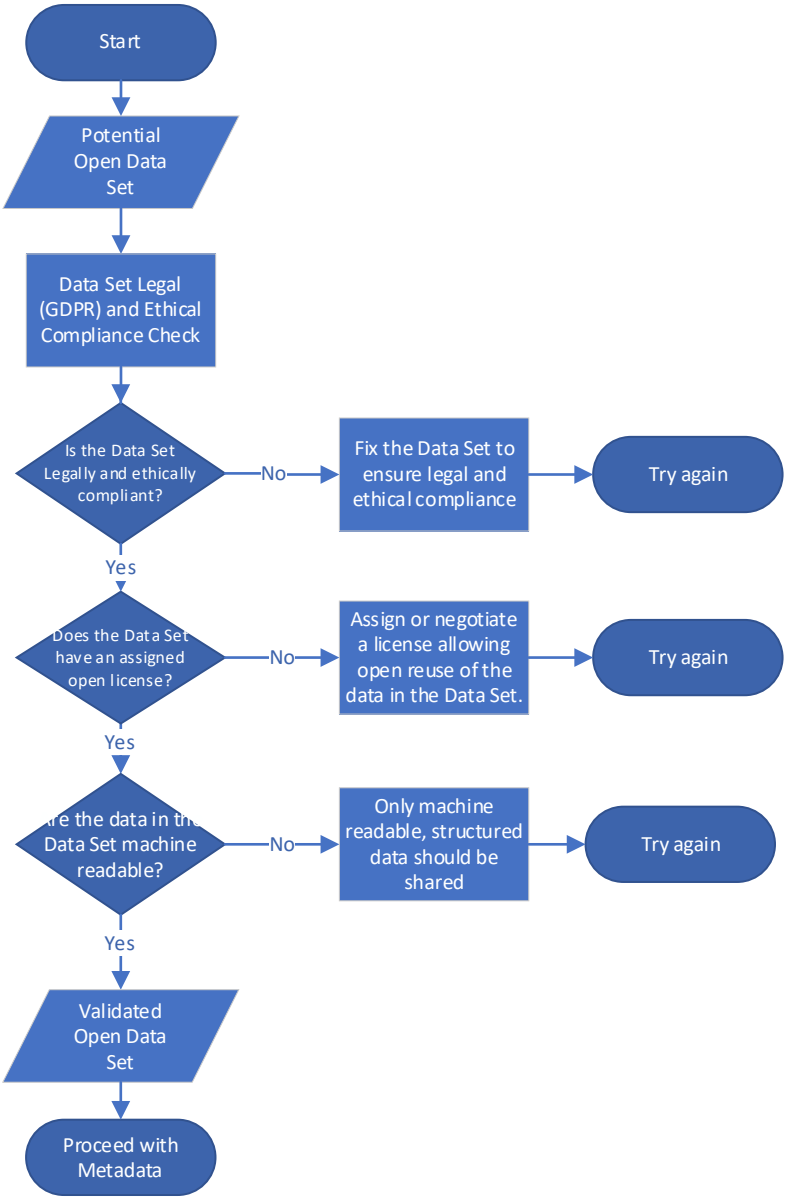

Figure 1. Open Data Critical Criteria

If you were able to answer 'Yes' to the three-qualifying question regarding your Data Set, it means you have a legally (including GDPR) and ethically compliant Data Set with an open license assigned in a machine-readable data format:

- ☒ Legal (including GDPR) and Ethical Compliance
- ☒ License Assignment
- ☒ Machine Readable & Reusable Data

Next, we proceed with assigning, defining, and adding metadata to the Data Set.

4.3. Data Set Metadata – Recommended Fields

The FAIR Guiding Principles[5] – FAIR—Findable, Accessible, Interoperable, Reusable – best describe the properties of Open Data. Applying these principles to the Data Set in focus will ensure excellent description and presentation of your Data Set from various criteria of metadata.
It is best practice to describe Open Data Sets with Metadata keeping the potential data recipient in mind. Assigning metadata correctly often ensures filtering and search facilities, make your data findable and accessible, possibly specific to the repository chosen. The checklist below contains most typical metadata categories to be included and defined:

- ☐ Assignment of a Globally Unique and Persistent Identifier
- ☐ Domain Assignment, domain classification scheme
- ☐ Data Set Title – version in English present
- ☐ Data Set Description – version in English present
  Describe the background of the Data Set, how the data originated
- ☐ Data set Type (Audio / Geospatial / Statistical / Text / Video / other)
- ☐ Data set MIME Type
- ☐ Keywords

- ☐ Data set License Information
- ☐ Data set Author Information
- ☐ Data set IPR holder
  Name / Institution / Address / Email / URL
- ☐ Data set Compiler / Producer Information
- ☐ Data set Contact Person Information – name and email

- ☐ Data set size – quantity and unit. Multiple versions possible
- ☐ Data Set Data Collection
  Date / Location
- ☐ Date of Data Set Creation
- ☐ Data Set Funding sources
- ☐ Version of Data Set

- ☐ Metadata Assignment in relation and specific to the area of the Data Set[6]
- ☐ Information About Creation / Processing / Compilation of the Data Set

[5] https://www.go-fair.org/fair-principles/
[6] http://rd-alliance.github.io/metadata-directory/standards/

- ☐ Quick Data Set Content double-check
- ☐ Readability of files
- ☐ Data Set Files and Metadata Files included

Now, once we have fulfilled the critical requirements towards Open Data and described our Data Set with metadata and added it to the Data Set, we are ready to publish it to an Open Data Repository.

You should study the options to select a repository which matches your data best. It may be a single repository for an entire project producing a multitude of open datasets for example Zenodo[7]. You could consider national open data repositories where data comes in from miscellaneous national institutions, and is annotated with the respective metadata of the organization and the area of its activity and the nature of the data. There also exist academic repositories related to academic programmes or areas of content. CLARIN is a very typical representative of such a repository. Alternately there are other open data repositories driven by the community of open data enthusiasts which often contain data collected by data enthusiasts-researchers collected, processed and published for the common good. You can search for data repositories at: re3data.org[8]. National open data repositories in Europe get scanned, and the published data aggregated and republished in the EU Open Data Repository. If you put your data in your national open data repository then go and check if you can also find your data in the EU Open Data Repository.

4.4. Validation After Publishing

- ☐ Metadata of the Data Set must be indexed and searchable
- ☐ Data Set can be found by intended means of lookup and search
- ☐ Data Set is accessible for free download to anonymous user

Make sure you fill in all the properties the hosting site offers. That will make sure your data will be findable provided your descriptions are matching the interest of the data seeker.

5. Consortium Data Activities

At the beginning of the project, we conducted a data survey to understand what data project partners plan to collect. The information was collected through an online survey tool. The project partners were asked to provide answers to following data related questions:

- What data will you gather or produce in this WP/tasks?
- What are the sources of this data?
- Are there research participants involved in gathering this data?

[7] https://www.zenodo.org/
[8] https://www.re3data.org/

- In what manner will this data be gathered?
- In what format will this data kept?
- With whom do you plan to share this data (inside and outside the consortium)?
- What existing data will you use for this WP/task?
- What are the sources of this data?
- In what format is this data kept?
- With whom do you plan to share this data (inside and outside the consortium)?
- Is there other data you may use or generate that is not yet certain?
- What is the size of the dataset(s) (approximate)?
- How are you securing the data?
- Will this WP/task contain personal data, if yes, please specify. ('personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or
- If there is personal data, do you plan to anonymize it or pseudonymize it? If yes, how?
- Full name
- Email
- Organization

The answers to the questions provide an overview of the planned activities, which include data processing and creation aspects. At the same time, it makes it possible to identify in a timely manner which Data Sets could be processed without having to worry about the presence of personal data and would therefore allow such Data Sets to be opened more efficiently. It also makes it possible to identify in a timely manner those datasets that need special attention and which may require large investments of resources to make them open.

Table 2 lists the main data groups and sources from which the data will be derived. It is clear from the initial assessment that certain groups of data will include personal data. Particular attention will be paid to these data groups to ensure that the data is processed in accordance with the DMP plan and that the requirements of GDPR are met. If it is decided that Data Sets are important to publish and need to be made open then the Data Sets will be processed for this purpose and mostly likely fully anonymized. Additional analysis will be performed to define the requirements for the anonymization tool, for example, an analysis of the data structure will be performed to determine whether the solutions should be able to work with structured data or unstructured data. At the same time, the language and domain of the original data will be assessed to identify solutions that are appropriate for that language and domain.

Table 2 Assessment of data collections

Data Group	WP	Data Description (Data that will be used)	Existing Data Re-used	Create new data	Source	Data Utility and Purposes
Dissemination materials	WP 7	Dissemination and training reports	Yes	No	Internal deliverables from previous tasks	Open access To disseminate project outcomes

Data Group	WP	Data Description (Data that will be used)	Existing Data Re-used	Create new data	Source	Data Utility and Purposes
Dissemination materials	WP 7	Multimedia dissemination materials	Yes	No	**1.** Internal deliverables from previous tasks **2.** Publicly available multimedia data	Open access To disseminate project outcomes
Feedback from Users	WP 3	User's feedback about game	No	Yes	Players' responses	WP3 game validation and improvement
Feedback from Users	WP 5	User's feedback about game during pilots of the game in schools	No	Yes	Players' responses	WP3 game validation and improvement
Game data	WP 3	Data provided by the player (e.g., age, gender, avatar characteristics, etc.)	No	Yes	Players' responses	WP6 Profiling and Data analysis
Game data	WP 3	In-game data (e.g., decisions made in the game, time playing, number of playing sessions)	No	Yes	Playing sessions	WP6 Profiling and Data analysis
Game data	WP 5	Data provided by the player (e.g., age, gender, etc.) during pilots of the game in schools	No	Yes	Players' responses	WP3 game improvement
Game data	WP 5	In-game data (e.g., decisions made in the game, time playing, number of playing sessions) during pilots of the game in schools	No	Yes	Playing sessions	WP3 game improvement
Game data (Summary Information)	WP 6	Player profiles	Yes	No	**1.** Publicly available data at region level to enrich the profile analysis (such as GDP, religiosity, HDI, etc.) **2.** The rest of the data belonging to the *Game data* group	WP7 to create policies and guidelines, and for academic publications
Game data	WP 6	Synthetic data	Yes	Yes	Algorithms fed with the rest of the data belonging to the *Game data* group	WP6 for research purposes Academic publications and open access datasets
Interviews	WP 1	State of the Art interviews (youths, victims, offenders, experts, etc.)	No	Yes	interviewees' responses	WP1 for research purposes, and the whole consortium for following tasks
Interviews	WP 5	Questionnaires to the involved youths/teachers	No	Yes	Youths/teachers' responses	WP3 game improvement, and the whole consortium for following tasks
Language data	WP 3	Collection of Language data (including terminology and parallel corpora)	Yes	Yes	**1.** Publicly available datasets for translation tasks **2.** Proprietary data accessible to TILDE	WP3 Game translations and open access translation datasets
Open content	WP 1	Sentencing reports	Yes	No	Open access Internet sites.	WP1 for research purposes, and the whole consortium for following tasks

Data Group	WP	Data Description (Data that will be used)	Existing Data Re-used	Create new data	Source	Data Utility and Purposes
Open content	WP 1 WP 2 WP 3 WP 4 WP 5 WP 6 WP 7	Academic documents, reports and general open access literature	Yes	No	Open research repositories and open access Internet sites.	The whole consortium for research purposes, and for following tasks
Open content	WP 1	Open access information about the offenders' cases	Yes	No	Open access Internet sites.	WP1 for research purposes, and the whole consortium for following tasks
Open content	WP 3 WP 4 WP 7	Collection of legal documents and regulations	Yes	No	Open access Internet sites.	The whole consortium for research purposes, and for following tasks
Open content	WP 7	Awareness campaigns materials	Yes	No	Open access Internet sites.	WP7 for dissemination activities

6. Conclusions

In this deliverable, we have summarized the key steps that are essential for open data considerations. The main purpose of this deliverable is to provide Practical guidance to data holders on how to Open data collected generated during project implementation. Data management and data management principles are set out in deliverable D8.6 and not the purpose of this deliverable.

One of the most important aspects to consider is the legal and Ethical Considerations. One of the main considerations is compliance with the General Data Protection Regulation (GDPR). As described in section 3 it is important to understand the meaning of “personal data” and presence of personal data in Data Sets. The presence of personal data cannot necessarily mean that Data Sets are close datasets and cannot be opened. It only indicates that such Data Sets need special attention and special data processing techniques must be applied (e.g. anonymisation or masking) before opening such Data Sets. Before a Data Set which contains or might contain personal data can be shared, the compliance will be verified with project legal partner **Timelex**. Only after “Green light” data holder can proceed to metadata activities and validation steps.

We are confident that following validation steps described in Section 4 data holder will be self-assured that there are no critical elements which may have a negative effect on the institution which opened the data and Data Sets corresponding to the main characteristics of Open Data – Open License, freely accessible, machine readable and open format.

7. References

- The Open Definition, http://opendefinition.org/
- Wikipedia, https://en.wikipedia.org/wiki/Machine-readable_document
- Wikipedia, https://en.wikipedia.org/wiki/Machine-readable_data
- Creative Commons, https://creativecommons.org/choose/
- FAIR data principles, https://www.go-fair.org/fair-principles/
- Research Data Alliance Metadata Standards Directory, http://rd-alliance.github.io/metadata-directory/standards/
- General-purpose open-access repository, https://www.zenodo.org/
- Registry of Research Data Repositories. https://doi.org/10.17616/R3D, https://www.re3data.org/