X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number: 761758
Project Acronym: X5GON
Project title: X5GON: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date: 2017-09-01 to 2020-08-31
Project Duration: 36 months
Document Title: Ethical Data Management and Data Management Plan Year 2
Author(s): Davor Orlic, Bostjan Makarovic, Maja Bogataj Jančič, Abe Hsuan
Contributing partners: UCL, K4A
Date: August 31 2019
Approved by: John Shawe-Taylor
Type: Report
Status: Final Report
Contact: bostjan@aphaia.co.uk, maja@ipi.si, davor.orlic@ijs.si

Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Revision		
Date	**Lead Author(s)**	**Comments**
21.6.2018	Davor Orlic	Initial draft for both ethical management and data management parts
27.7.2018	Erik Novak	Added data types, descriptions and filled in tables for WP2 and WP4 tasks
27.7.2018	Davor Orlic	Second draft
30.7.2018	John Shawe-Taylor	First review and feedback
2.8.2018	Bostjan Makarovic	Ethics chapter added
6.8.2018	Maja Bogataj Jančič	Intellectual property chapter added
10.8.2018	Colin de la Higuera	Added data types, descriptions and filled in tables for WP3 and WP5 tasks
17.8.2018	Nils Oesting	Input on OuS tables
17.8.2018	Bostjan Makarovic	Ethical part update on second draft
20.8.2018	Abe Husan	Ethical part update and review
20.8.2018	Sahan Bulathwela	Added data types, descriptions and filled in tables for WP1 and WP9 tasks
20.8.2018	Stefan Kreitmayer	Added data types, descriptions and filled in tables for WP6 tasks
20.8.2018	Mojca.Ifko-Pinosa	Added data types, descriptions and filled in tables for Task 7.3
21.8.2018	Davor Orlic	Final review and versioning of complete deliverable
19.8.2019	Bostjan Makarovic	Addition of relation to EU Ethics Guidelines for Trustworthy AI (EGTAI) with updates in sections 1.1, 1.2 and add addition of 1.6 privacy action points to be considered
21.8.2019	Bostjan Makarovic	Meeting with John Shawe-Taylor AND Davor Orlic to define privacy scenarios
22.8.2019	Davor Orlic	Acceptation of comments with Section 2 remaining the same
29.9.2019	Maja Bogataj Jančič	IPR part general update and addition of the 2.3.
4.9.2019	Davor Orlic	Section 2 on Data Management plan moved to Appendix 1, Section 4.5 Ethics moved to Appendix 2 with minor edits
6.9.2019	Davor Orlic	Addition of 2.4, Sections 3. And 4. and circulation of draft
9.9.2019	Maja Bogataj	Final comments and edits in 2.4
10.9.2019	Davor Orlic	Final deliverable

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CC	Creative Commons
CR	Copy Right
DPA	Data Protection Impact
DSM	Digital Single Market Directive
EU	European Union
TDM	Text and Data Mining
ELT	Ethics and Legal team
EGTAI	High-Level Expert Group on Artificial Intelligence has drafted EU Ethics Guidelines for Trustworthy AI
EU	European Union
GDPR	General Data Protection Regulation
IPR	Intellectual Property Rights
ML	Machine Learning
OER	Open Educational Resources
OERDP	Open Research Data Pilot
RP	Research personnel

ABSTRACT

This deliverable is an updated and upgraded version of D9.1 which described the first thinking of the consortium around the challenge of ethical data management and data management plan.

This deliverable is comprised of the three sections, the first on Ethical data management and covers the following topics:

- Report on the privacy policies that will be used by the project and the different OER sites, these should specifically include the use of user data for research purposes;
- Ethical vs. compliance aspects of privacy, including EU Ethics Guidelines for Trustworthy AI;
- Implications of the algorithms used or foreseen for machine learning and their ethical implications, including profiling;
- Lawful bases for data processing, including in case of children;
- Proposed forms and data management policy;
- If required, compliance clearance from the relevant DPAs or existing ethics boards for the activities in different countries;
- Report on the conditions for the lawfulness of the final product under EU General Data Protection Regulation (GDPR)
- Identify data controllers and processors.

The second section is on Intellectual property focusing on:

- Directive 2019/790 of the European Parliament and the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (DSM Directive);
- New mandatory exception for text and data mining in the DSM Directive
- Effects of both policies on X5GON;
- General copyright and related rights issues;
- Definitions and procedures for obtaining and processing collected materials;
- Definitions and procedures for making available materials for public;
- Licensing of OER materials.

The third section of the report is now in the Appendix and is focused on the Data Management Plan as part of the Open Research Data Pilot and covers in details:

- The handling of research data during and after the end of the project;
- What data will be collected, processed and/or generated;
- Which methodology and standards will be applied;
- Whether data will be shared/made open access and how data will be curated and preserved (including after the end of the project).

This is accompanied by a report on the key intellectual property aspects of X5GON and a breakdown of ethics and privacy issues in relation to various aspects of the project.

1. ETHICAL DATA MANAGEMENT

1.1 INTRODUCTION

Privacy and the protection of personal data represent a crucial legal and ethical aspect of the processing of resources that include information pertaining to individuals. It is our opinion that there is a considerable overlap between legal compliance and ethics, notably because the EU General Data Protection Regulation (GDPR) that recently became applicable considerably raised the standards for personal privacy protection, which are currently in the EU at the highest level on the global scale, whilst taking into account the latest technologies such as cloud computing. In fact, the GDPR is currently the most advanced comprehensive privacy protection system of rules and principles in the world, which makes it an appropriate proxy for ethical expectations in the privacy field when it comes to business and academic conduct.

In addition, High-Level Expert Group on Artificial Intelligence has drafted EU Ethics Guidelines for Trustworthy AI (EGTAI). There is to some extent an overlap between this document and the GDPR but the EGTAI capture a wider area of ethical issues pertaining to the use of AI and should therefore be taken into account in this report as well.

We further note that the concept of data though does not necessarily have the same meaning and implications in law as it does in everyday, ethical, or scientific use. We note that, from the personal privacy (and broader human rights and ethical) perspective, only data pertaining to identified or identifiable individuals might be relevant. From the intellectual property point of view (see chapter 'Intellectual property' bellow), individual pieces of information are unlikely to matter, and it is only data organised as copyrighted works or into databases under sui generis legal protection that is likely to be relevant.

1.2 REPORT ON PRIVACY POLICIES

Privacy and Data Protection become relevant in the context of X5GON where the data processed as part of the project OER analysis constitutes personal data as defined by the GDPR i.e. any information relating to an identified or identifiable natural person (data subject), whereby *"an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*."

In the analysis of Open Educational Resources or OERs themselves, two basic scenarios can be foreseen that may result in the processing of 'crawled' personal data:

- Personal data is part of a relevant OER item.
- Personal data is part of a false positive that is deemed to be a relevant OER item. We believe the existence of false positives is inevitable if OERs are to be automatically identified online.

In addition to OERs themselves, the project further comprises OER-related 'metadata' such as (optionally) the author of the resource and further data evaluating

the OER in question, acquired through the acquisition or enrichment process e.g. through sentiment analysis of online opinions or using other third-party online sources such as Wikifier. Such analysis may include actual identifiable persons giving their opinions online and may therefore comprise personal data.

In each of these contexts, two considerations should be made:

- As to the role of the data subject appearing in crawled data i.e. the OER or in relation to it e.g. in 'metadata'
- As to the jurisdiction of the EU and its Member States when it comes to data processing.

As to the role of the data subject, a typical role of the data subject would be that of the author or the presenter of the relevant OER, appearing either with their name and other personal details such as job / academic position, on video, or voice recording.

However, provided he or she is identifiable/recognisable, other roles of a data subject can be foreseen in an OER, such as but not limited to:

- Member of a focus group
- Respondent to a research question
- Member of the audience asking a question
- Passive member of the audience.

As to the jurisdictional aspects of privacy and data protection, one should first recognise that the EU GDPR includes several rules of establishing its jurisdiction, most vital for EU-based operations such as X5GON being that it applies "to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not."

The main consequence of this rule in the context of OERs published or otherwise processed outside the EU yet analysed in the context of X5GON could therefore be that a lower privacy and data protection standard might apply in the context of their original publication or other processing than in the context of X5GON processing. This means that X5GON operations could not be based on the assumption that the original publication or other processing of the OER imply a blanket permission for further processing in the context of X5GON. Accordingly, such processing will require lawful basis even if the OER resources relate to the data subjects or OER publications outside the EU.

A further aspect of personal data processing as part of X5GON relates to the X5GON platform / product users themselves, including their 'implicit feedback' (D1.3 – Initial Content Representations), and their 'social network' whereby Learning Analytics Engine includes: "a map of users: a user can be seen as a distribution of skills."

These skills are built from topic maps, annotations, forum discussions, student interactions, results from games and quizzes will be recovered and organised for this purpose. For example, if a user is watching a particular video and stops it, can we guess why she has stopped it? Has she heard something she can't understand? Was it too difficult or too easy? Is there a need to learn other resources first before proceeding with this one? Is she wanting to hear the same thing in another language? The discovery of intent has been identified as a crucial component. Whereas these operations on user data constitute profiling, they are likely to require

user consent and would be subject to the users' 'right to object to profiling'. We address this issue in relation to machine learning below. We note that the user's age group may also be used in this context (D3.2 – Learning Analytics). We further note that X5GON does not import individual user data from non-affiliated repositories (D4.5 – Prototype of cross-site recommendation engine).

It is therefore critical for privacy-related policies in the context of X5GON to capture all of the above aspects of personal information, recognise every context in which personal data may be processed as part of the project or the final product, and consider the jurisdictional privacy aspects that may in the context of X5GON differ from those in relation to the original publication or presentation of the relevant OER.

1.3 LAWFUL BASES FOR DATA PROCESSING, INCLUDING WHERE THE USERS ARE CHILDREN

We initially note that big data processing operations make it extremely difficult to seek and obtain consent for data processing in relation to the information contained in OER and OER 'metadata'. Accordingly, until reliable global universal machine-readable consent mechanisms are developed and adopted, X5GON operations would largely have to rely on legitimate interest as the appropriate data processing ground. This however requires the weighing of the legitimate interest of the data controller against rights and freedoms of the affected data subjects, which means that X5GON personal data processing operations must be considered with great care.

In case of commercial exploitation of the data gathered, one has to bear in mind that legitimate interest test i.e. weighing exercise of the legitimate interest of the controller compared to data subjects' rights and freedoms may yield different results as opposed to purely academic / educational / non-commercial exploitation.

Platform users, on the other hand, actively use the X5GON platform / product, which makes it possible to obtain their consent or justify the processing of their data by means of the contract into which they have entered in order to use X5GON. Where the users are children, one will have to take into consideration Article 8 of the GDPR. Whereas X5GON platform and the products based on it are most likely going to be offered as information society services (i.e. online interactive service) directly to a child, typically a student, the processing of the personal data of the child shall be lawful where the child is at least 16 years old. Where the child is below the age of 16 years, such processing shall be lawful only if and to the extent that consent is given or authorised by the holder of parental responsibility over the child. The controller must make reasonable efforts to verify in such cases that consent is given or authorised by the holder of parental responsibility over the child, taking into consideration available technology. We note that Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years, the latter being the case in the UK.

1.4 ETHICAL AND COMPLIANCE IMPLICATIONS OF MACHINE LEARNING ALGORITHMS

Machine learning algorithms predict, as foreseen in X5GON, the preferences of the platform/product user. However, they could also be used e.g. to predict the quality of the OER resources based on its author or presented. In both cases, the use of such algorithms would amount to 'profiling', which is under GDPR "any form of automated

processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements". X5GON is expected to be able to go as far as capturing, 'implicit feedback', 'the mood' of the user or 'topic maps' predicting the questions of the user, building probablistic relational model of the resources the user may be interested in, the user's topic vector, learning path (D3.2 – Learning Analytics). Information suggesting the user's current knowledge and goals is further used to inform X5GON's Learner Engagement Model (D1.3 – Initial Content Representations).

There are two major GDPR compliance considerations in relation to the use of such algorithms: transparency of such algorithms and the right to object. With a view to EGTAI, one needs to add human agency and oversight, diversity, non-discrimination and fairness, societal and environmental well-being, plus accountability.

Full transparency of the algorithms used is not expected, notably due to the need to safeguard intellectual property. However, the GDPR does in relation to the mandatory algorithms refer to "meaningful information about the logic involved," a standard that could be used as ethical guidance for providing information in all cases where profiling is used.

Machine learning used in the context of X5GON is based on the attempt to deliver the most appropriate OERs to the right users at the right time. Its learning algorithms are explained in X5GON documentation and their results are overseen by the X5GON expert team. Based on our analysis of the documentation, it would not appear that the algorithms are used in a way that could discriminate against- or adversely affect any group of people based on their personal circumstances apart from the interest they express in specific OERs, their approach to consuming OERs in accordance with their prior knowledge, and basic information such as language or age, which should not be considered discriminatory in the context of learning activities. People with disabilities should be able to access X5GON using standard computer accessibility interfaces. Extended and fair access to the relevant OERs as enabled by X5GON applications is expected to be beneficial to societal and environmental well-being, particularly compared to traditional commercial search engines that are designed primarily to drive more traffic regardless of wider societal benefits or harms. Accountability is further ensured by means of H2020 reporting mechanism, plus our recommendation to explain the basics of the algorithms in a publicly available privacy policy, in line with GDPR. Trade-offs are reflected in the balancing of user privacy with the need to deliver the most relevant OER recommendations.

The right to object would apply both where the algorithms are used to process OER data and 'metadata', and where they are used on user data. Once the objection is raised, "the controller shall no longer process the personal data unless the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject." That said, such grounds are more likely to be present when it comes to publicly available OER information than in case of user profiling. It may therefore be worth considering the use of the algorithms and machine learning in relation to a particular user as an optional feature.

This may be further relevant in relation to the plan "to develop architecture that supports creation and management of user models for capturing and modelling behavioural information from the OER Social Network that takes advantage of the platform". We note that additional safeguards such as pseudonymisation may be

needed where algorithms are developed on actual persons' data. Since cross-lingual / cultural profiling is expected be used in relation to the social network, we also note some data such as 'ethnicity' may only be processed based on express consent.

1.5 DATA CONTROLLERS AND PROCESSORS

Initially, consortium members would be expected to be in the role of (joint) data controllers with regard to the data gathered from OERs and in the related 'metadata'. However, subject to future commercial arrangements for the backend technologies exploitation, we have foreseen that it might be possible for consortium members / X5GON to act as data processors for their clients. In such cases, the responsibilities towards data subjects would primarily fall on the clients. However, under Article 28 of the GDPR, X5GON / its consortium members would still need to assist the clients in fulfilling their obligations under GDPR towards data subjects whose data would be processed by the system.

That said, considering X5GON's machine learning development, it might be challenging to reconcile a scenario whereby X5GON would continuously learn based on its use by all of its users with the role of a data processor. This is because the learning element would be expected to constitute 'determining purpose and means' of data processing according to GDPR. Accordingly, it is only under a scenario where a separate user information databased would be set up for the client that X5GON consortium members could be deemed processors. Otherwise, they would more likely be considered 'joint controllers' together with their client.

1.6 PRIVACY ACTION POINTS TO BE CONSIDERED

We recommend the writing up and the publication of a privacy policy for X5GON that would include all the key information mandated by the GDPR, including basic information about the algorithms (e.g. how user models are built/affected by accessing different OERs, see D4.2 – Final prototype of user modelling architecture, or basic information on Bayesian approach to predict engageability - see D1.3 – Initial Content Representations) used discussed above and data subject rights. Where the right to erasure or to object to processing is exercised, X5GON should have a feature in place to prevent repeated adding of the relevant resource.

We further note that the API and the models used would have to ensure that the full analytics data attributable to an individual user can only be accessible to that particular user, not all the API users. To the extent the user model for a particular user is used to inform recommendations/searches pertaining to other users or building models, including as part of X5GON's product sold to another entity, such data must be anonymised. It is our understanding that this can be done by adding the user data in an anonymous form to the relevant OER.

2. INTELLECTUAL PROPERTY MANAGEMENT

2.1 CATEGORIES OF PROTECTED SUBJECT MATTER

The materials collected and made available in the X5GON project can be subject to copyright or sui generis database rights.

A Copyrighted works

The materials are subject to copyright, if they fulfil the requirements for copyrighted works according to the governing legislation. Most copyright legislation of EU member states protects works, if they are an "individual intellectual creation".

1 In case of materials, collected and made available in the X5GON project (educational and scientific content, encyclopaedia content, social media posts, etc.) it is necessary to consider, that these works are likely to be eligible for copyright protection and in some cases also subject of neighbouring rights (i.e. performers rights);
2 If the materials collected and made available in the X5GON project represent a database or a part thereof, which fulfils the criteria of copyrighted works (i. e. if they are "individual intellectual creations"), such databases or their parts will be eligible for copyright protection.

B Sui generis databases

If the materials collected and made available in the X5GON project represent a database or a part thereof, such materials can be protected by rights, similar to copyright, even if they do not fulfil the criteria for copyright protection, if they are eligible for sui generis database protection. The maker of a database has exclusive rights to the database, if the obtaining, verification or presentation of its contents demands a qualitatively or quantitatively substantial investment. Protection applies to the entire contents of a database, every qualitatively or quantitatively substantial part of its contents and even insubstantial parts of its contents, when they are used repeatedly and systematically, which conflicts with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database. Due to the nature of the X5GON project it is worth noticing, that even the alphabetical listing of words in a dictionary (even without the translation or explanation of the words) can be subject to sui generis rights, depending on the national case law.

2.2 RIGHTS CLEARANCE AND MANAGEMENT

For the X5GON project it is necessary to address copyright and related rights issues in two phases of the project:

a) Obtaining and processing the collected materials.
b) Making materials available for the public.

A Rights clearance for obtaining and processing the collected materials

For the collection and processing (that is transformation, arrangement, translation, modification, systematization, etc.) of materials copyrights and sui generis database

rights will be adequately cleared with rights holders (originators of educational and non-educational material (instructors, professors, educators, teachers), crowdsourcing workers, the X5GON project research personnel, other authors and makers of databases, publishers and other rights holders). Rights holders will be informed in detail about their participation in the project and called upon to sign Rights Transfer Agreements, which transfer the relevant copyrights and allow for specific uses of the materials for the purposes of the X5GON project. It is crucial to obtain above all the right of reproduction and right of making available, as well as transformation rights, that allow for modification, translation, adaptation and other changes of the material, necessary for the execution if the project.

Copyrights will be observed even in cases, where materials have been made available under open access and similar terms (like under specific terms of Creative Commons or similar licenses), to avoid rights infringement in cases, where certain uses of the materials made available are restricted (for instance prohibiting the making of derivative works, non-commercial uses only, further sharing materials under original restrictions, etc.).

Users of the X5GON project results will be obliged to sign Rights Transfer Agreements for all instances where their activities generate copyrighted material and/or sui generis databases. Through such agreements users will explicitly agree to further use, transformation and making available of their works.

B Making available the collected materials

We identify two possible ways of making the collected materials available. The first is that users of the outcomes or results of the project will need to sign a Right Transfer Agreements before they will be able to use the outcomes. However, as this might not be correct or beneficial, we list an option where an open license could be chosen for the project results.

1 For the purposes of the X5GON project Terms & Conditions will be drafted, which must be abided by all entities, who will make the materials available (through their platforms, apps, etc.) as well as by all users. Appropriate Rights Transfer Agreements will be concluded with platform providers, allowing them the making available of any copyrighted content as well as all uses (including modifications) by the users of the platforms.

2 For the purposes of the X5GON project all copyrighted materials will be made available under an appropriate Creative Commons or similar license (like the CC-BY-SA license - attribution-share alike), which allows for use and making available of any copyrighted content by platforms as well as all uses (including modifications) of the materials by the users of the platforms.

2.3 TEXT AND DATA MINING

During the past 12 months the development of the project and significant changes in the legal framework in which the project operates, especially the regulation of copyright in the European Union, have led to the clarification of certain issues concerning intellectual property that have to be addressed. This is important development for the X5GON project specially for the partners who come from jurisdictions that currently do not have TDM exception in their legislation. The new legislation when implemented will provide new legal certainty for text and datamining for scientific purposes.

2.3.1 The DSM Directive

The Directive 2019/790 of the European Parliament and the Council of 17 April 2019 on copyright and related rights in the Digital Single Market[1] (DSM Directive) will have an important effect on the functioning of participants on the internet as it will reshape the rules relating to copyright to which these participants have to abide. The Directive came into force on 7 June 2019 and the Member States of the EU now have until 7 June 2021 to transpose those rules into national legislations.

The provisions of the DSM Directive that are most relevant for the X5GON project are Articles 3 and 4 which introduce a new mandatory exception for text and data mining (TDM). It should be noted that some Member States have already introduced a TDM exception for instance, UK, France, Germany, Estonia and Ireland.

TDM is defined in the DSM Directive as the automated analysis of a large amount of information in digital form, such as text, sounds, images and data, in order to generate information which includes but is not limited to patterns, trends and correlations. TDM can involve acts protected by copyright, by the sui generis database right or by both. Before the introduction of this exception (except in rare cases where jurisdictions regulated TDM) research organisations were confronted with the legal uncertainty as to the extent to which they can perform TDM. However, with the adoption of the DSM Directive these rules will now be much clearer.

According to Article 3 of the DSM Directive, research organizations and cultural heritage institutions mining for the purposes of scientific research (the beneficiaries) will be allowed to use TDM on content and data, to which they can legally accede, without being obliged to ask permission of the authors and without having to pay any special compensation. Lawful access means access to content based on open licenses and contractual agreements with copyright holders. Apart from allowing the mining as such, the exception also permits the data to be stored for the time needed to provide the verifiability of the researches.

According to Article 4 of the DSM Directive, subjects other than the beneficiaries will also benefit from the TDM exception, but only under the condition that the right holders have not expressly reserved the use of the work or other subject matter. The collected data would be stored only for as long as that is necessary for the purposes of TDM.

2.3.2 DSM Directive action points to be considered

In the X5GON project, different actions that can be classified as TDM are performed in relation to copyrighted content and database. For instance, as described in the “D1.3 – Initial Content Representations” report, data mining is used on the text of educational resources to automatically extract units of learnable concepts contained therein.

Before the adoption of the DSM Directive, it was not clear which of the actions would require the authorisation of the copyright holder in order to be performed. While reading published content and data was always permitted, reproductions and extractions might have been considered an infringement. With the introduction of the mandatory TDM exception in the DSM Directive and its transposal into national legislations, TDM will be permitted for research organisations involved in the X5GON project mining for scientific research purposes.

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790&from=EN

2.4 INITIAL LICENSING CONSIDERATIONS

For OER, the most widely used open licenses are the CC licenses, which make it possible for educators to freely and legally share their work. OER is licensed under different licenses and the difference among these licenses should be a special concern of the X5GON project. It should be made clear that only certain open licenses are adequate, presumably the one allowing the widest possible portfolio of reuse, namely the CC-BY licenses. The CC-BY license allows the reproduction and redistribution of the material in any medium or format and allows to transform (also translate) remix, and build upon the material for any purpose, even commercially.

Since the exploitation routes have been identified, we suggest that in Year 3 of the project the differentiation of the licensing condition of the harvested material is a special concern and a special task of all involved. i.e. only licenses that allow derivative works (and commercial use if this is necessary for the project) should be determined adequate and not all open licenses and consequently all OER. It is also worth noting that X5GON merely points to materials and doesn't use them or store them. At his point materials are only used for ML modelling via the case studies.

We have also determined the various key players in the X5GON project and the relations between the partners, and have been able to transparently determine their obligations relating to the clearance of rights and rights transfer chains for harvesting the material, processing the material (transformation, adaptation, translation, systematisation, etc.), storing the material and making it available to public.

The sources from where the material is harvested influences the rights clearance. Harvested material can be available online under different conditions – some of the material is available under open licenses as explained above, some of materials can be used for specific purposes only (like non-commercial, research purposes, teaching, etc.), some of the material could have other non CR restrictions: i.e, that enable read-only and prevent further use, etc. others. It is important to stress that some material can be also in public domain which means that the CR has already expired or it is not subject of CR protection at all.

For the right clearance it is important to employ a definition of the meaning of the term *data* (which can be personal or non-personal) and a such cannot be copyright (CR) protected and *material* (content), which is usually subject of the CR protection. We furthermore make clear, that CR rights clearance refers only to material (content), which is the subject of copyright or sui generis rights (it does not refer for example to the harvesting, processing and making available of "personal data", within the meaning of the GDPR).

Additionally, we have an initial understanding to what extent can the project be considered a research and what parts or results of the X5GON project will be exploited commercially (to be further developed in D8.4), as some material could be free to use for non-commercial purposes but restricted for commercial exploitation.

Finally, conditions under which the material will be made available to users are also defined. The material can either be made available by concluding transfer agreements with platforms (if the material will be made available by a third party platform and not the data harvester himself) and by imposing binding terms and conditions on the users of the platforms. Such rights clearance is carried out more or less on a case by case basis for each individual entity, that makes the material available to the public. The

other option is for the harvester to make the material available under a Creative Commons or similar license, which allows everyone the use of material under conditions, as applied by the harvester. In this case the rights to the material must be previously cleared with rights holders to allow the further sharing of the material under a chosen open license.

3. CONCLUSION

For the purpose of this deliverable we engaged with the ELT (Ethics and Legal) team, comprised of project personnel and three lawyers with expertise in privacy, IP and copyright and licensing. The team members have helped the X5GON consortium partners to understand the optimal ways in which to engage with ethics, data protection and IP and copyright issues in the currently changing legal landscape in the EU.

This deliverable is comprised of three parts. The first part describes and contains updates on the Ethical procedures around data, the second reports on the Intellectual Property Rights aspects of the project and highlights important aspects of the new text and data mining legislation, the third describes the projects data strategy. Both part 1 and 2 pave the way for further research in ethical questions of data management and copyright clearance in its further iteration (M36) will be researching on how to support our WP8 exploitation decisions and strategies in D8.4 Final business plan (M36).

The Ethical Data Management presents updates on the nature of the projects data in light of the GDPR and EGTAI and researches further on the privacy policies that will be used by the project towards the different OER sites, users and content in its final Year 3. It reflects on ethical and compliance aspects of privacy and presents the implications of the artificial intelligence algorithms and as a result reports on the conditions for the lawfulness of the final product under the GDPR regulation.

The Intellectual property rights section presents the implications of the new DSM copyright directive in the European Union for X5GON, and specifically Articles 3 and 4 which introduce a new mandatory exception for text and data mining TDM that will need to be implemented by EU member states. The section also presents the notion of OER being licensed under different licenses or a portfolio of licences, resulting that only certain open licenses are adequate for the X5GON project, assuming the final platform is used specifically for OER.

The Data Management Plan remains unchanged and provides a short and general overview of the partner's policy for data management.

In terms of future plans, the ELT team has identified a new list of challenges on including further discussions on the right to be forgotten, content management, right clearance and derivative works, and how parts or results of the X5GON project will be reflected in the event of commercial exploitation. Finally, the team plans to release the final deliverable with an additional section as recommendation to the ML and AI community in terms of guidelines for applications for educational purposes.

4. FUTURE STEPS

- Update the privacy policy for X5GON which already includes key information mandated by the GDPR, to include basic information about the algorithms;
- Include a section to exercise the right to erasure or to object to processing, and present and design a feature in place to prevent repeated adding of the relevant resource;
- Reassure that the used API models have the full analytics data attributable to an individual user, can only be accessible to that particular user, not all the API users;
- Make sure that in case of a commercial and exploitation activity, the user model for a particular user is used to inform recommendations/searches pertaining to other users or building models is anonymised.
- Understand issues about material being free to use for non-commercial but restricted for transformation
- Understand issues about material being free to use for non-commercial purposes but restricted for commercial exploitation.

5. APPENDIX I - DATA MANAGEMENT PLAN

5.1 INTRODUCTION

The Data Management Plan presents the management of the project's data and its intellectual property rights. This document should be considered in combination with Section 8, specifically Articles 8 and 9, Attachment 1 and 2 of the Consortium Agreement and Chapter 3, specifically Articles 23 - 31 of the Grant Agreement No. 761758. The format of the plan follows the Horizon 2020 template[2] and its content has been guided by the advice from the Ethics and Legal Team established as a monitoring board on legal requirements concerning data.

Here we also clarify that our definition of data (Section 2.4) states that data is not protected by CR. Herewith (Section 5) the term data is used for both data (non CR protectable) and for material (content, CR protectable).

X5GON data will be made publicly available after the complete setting-up of the X5GON platform, i.e. after M22 (see Task 2.2). This will include all available, collected and created data used. As the project is participating in the Open Research Data Pilot (OERDP), we will make our data findable, accessible, interoperable and reusable (FAIR). Therefore, it is planned to make the data widely open-access, free, and downloadable for verification purposes via the project website[3].

The project and the project partners are currently in the process of data collection and will continue to do so in cycles with each new upgrade of the platform until M36. More precisely we will be collecting content and user data from the case studies (JSI, UPV, OUS), partners with newly established OER repositories (NA, UCL), and entities outside of the project consortium.

[2] Guidelines on Data Management in Horizon 2020, (http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-datamgt_en.pdf), accessed 7/12/2018

[3] Website, https://platform.X5GON.org/, accessed 7/12/2018

The project will involve data collection for modelling and evaluation purposes. All the content (material) data at OER sites already conform to one of the licensees that enable the free distribution of an otherwise copyrighted work. In most cases these are Creative Commons licenses. All the work that will be done on the level of content processing and translation, will fall under the same license and be available for public free to use. The created learning behaviour models on the anonymous groups of individuals will also be distributable among OER sites.

This part of this two-fold deliverable focuses on the data types or data genre being used, collected and generated as promised in the Description of Action (DoA). Further, it will provide a general description of the data infrastructure as well as the legal and ethical measures addressed by the consortium (see Section 1). This is part of the impact, dissemination and exploitation strategy and specifically in synergy with Task 8.2 Impact assessment.

The final set of data collected in X5GON will be provided in the form of a database, designed as a combination of existing and new data from the OER domain which is automatically gathered from existing (online) open educational resources or created during the project. Finally, we present the design of the workflow for data collection, creation, analysis, and storage in Figure 1 during the lifetime of the project and after its completion. In the first level we collect and analyse the data, whereas on the second we create a metadata record for each dataset with data attributes and publish it in the X5GON research data repository. Below we present (Figure 1) the data management workflow, however at this stage we do not make a distinction between data and material (presented in D2.2), but we do take into consideration that we are addressing personal data in connection to privacy issues.

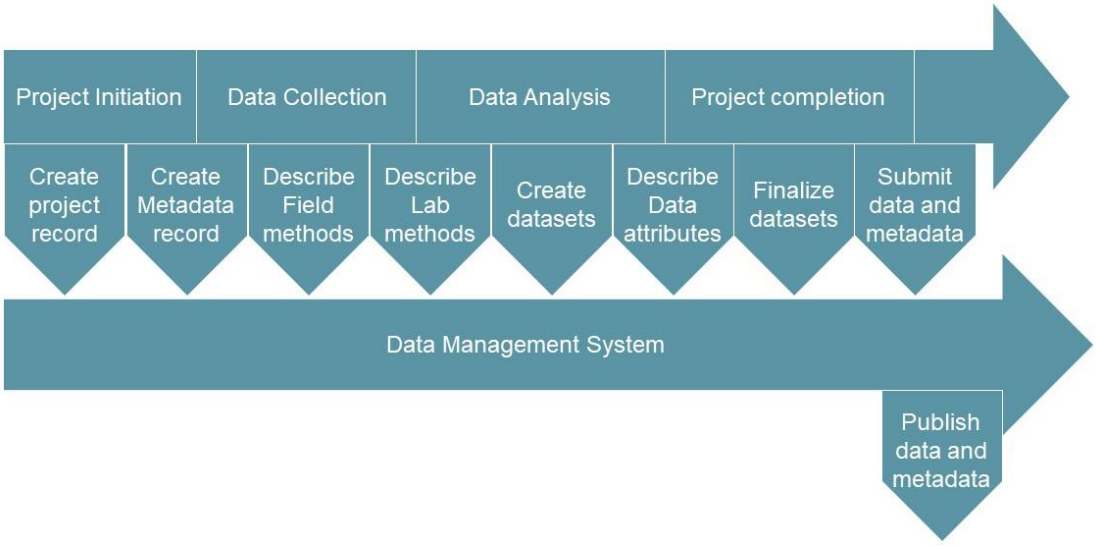

Figure 1: *X5GON data management workflow*

5.2 DATA DESCRIPTION AND DATA TYPES

The data collected, generated, and used in the project will include the following data types:

1. **Crawled data** is mainly educational and scientific OER repositories and OER web pages, LMS and CMS online information environments, etc. Generally,

the data collected via crawling will be multilingual and used for training the recommendation models and learning engines, respectively in WP2, WP3 and WP4.

2. **User data** (data in its traditional definition) includes data collected and indexed in OER resources, track data of users and their progress and use that will drive an analytics engine driven by state-of-the-art machine learning that can improve recommendations through better understanding of users, their progress and goals, and hence their match with knowledge resources of all types.

3. **Source code** is developed for the purpose of analysing crawled and user data, enriched metadata and for service development in WP1, WP2, WP3 and WP4. The source code will include both external services and services available within the consortium. External services will be carefully picked to follow

4. **Reports** include written reports, mostly produced in the form of deliverables, which describe the work progress in the project. Other types of documentation such as written blogs, social media bites, questionnaires, interviews, think-a-loud protocols, general surveys, self-assessment surveys, guidelines, meeting minutes, presentations, posters, promotional materials, etc. also fall into this data type. The Grant and Consortium Agreements regulate the level of access to the reports produced.

5. **Scientific publications** include relevant scientific journals, books, and conferences which report on the work in the project. All project related publications will contain an explicit acknowledgment to the project including the name and EU grant number.

The data included in types 1) to 3) will be collected and processed on site by JSI, UCL, NA, UPV, the partners which are responsible for the technological and data collecting activities according to the DoA. All partners will be involved in the production of data types 4) to 5).

Apart from the research teams in the consortium, the data employed in the project will be useful to other research groups working on Open Educational Resources, machine learning, machine translation, cross-site, cross-domain, cross-modal, cross-language, cross-cultural, cross-social, adaptive learning, policy making, educational data, open education in general, and OER platforms and ICT studies in general.

We plan to make our datasets, tools, and resources together with dissemination materials, scholarly journals, and open access publications publicly available. This is aimed at facilitating and encouraging data reuse residing in the X5GON research data repository. Below we provide a data preservation strategy and procedures for the sharing infrastructure.

5.3 Data Formats and Standards

In terms of data sharing the main challenge is the variety of formats in which the data to be used is available and the size of data types 1) to 3). Therefore, when making project data publicly available, we will seek to employ unified, widely accepted data

formats and standards. Below, we give a general description of the data format for each of the five data types identified in the previous section.

2.3.1 Crawled and User Data

All crawled and user data will be available in a JSON format and stored in a PostgresQL database. Each data type will be also exported in a single file with each JSON record in a single line. While crawled data will be publicly available through a public REST API developed in WP2, user data will be available only to project partners for analysis and will not be publicly exposed. In the future we may follow existing metadata sharing formats to support metadata sharing protocols such as Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH), but remain open to new developments.

2.3.2 Source Code

All source code will be available on public Github repositories. It will contain components written in different programming languages, such as C++, javascript and python, which will analyse and generate new data. The components will also communicate and share data with each other. Developers will follow code conventions relevant for the programming language they are working with. Unit tests will be provided where applicable.

2.3.3 Reports and Scientific Publications

Generally, all data which falls under the Reports category will be available in PDF format. There is a unified format for all this data created in WP7. Scientific publications will follow the format required by the conferences or journals in which these publications will appear. Whenever possible, we will try to provide a stand-alone PDF version of the publication.

2.3.4 Metadata

Crawled data will also be accompanied by metadata acquired through the use of enrichment services, such as Wikifier, Enrycher and XLing. This data includes Wikipedia concepts, named entities extraction, topic and keyword detection and other enrichments which will be used for understanding the data, and improving models developed in WP1, WP2, WP3 and WP4. This data will be stored in the crawled data dataset under the material metadata field and will follow the JSON format.

5.4 Data Storage, Access, Sharing and Reuse

2.4.1 Internal Data Storage and Sharing

PS, JSI and K4A have set up a secure data repository running on a protected server cloud. This presents the main internal storage and sharing mechanism and it is used by all partners. Every partner as well as the EU project officer has equal access to the repository. The repository is backed-up on a daily basis which minimizes the risk of crucial data loss. It has a modern, comfortable web interface in English making it easy to use. Partners involved in the preparation of the first version of the X5GON platform have already provided positive feedback on the repository.

Furthermore, all deliverables and other types of reports are uploaded in the project's Intranet. The same is valid for all project related publications and promotional materials. All templates used within the project (e.g., for deliverables, presentations, etc.) are also available for download in the Intranet. The Intranet is a collaborative tool which facilitates the sharing of data among the partners.

2.4.2 Data sharing and Reuse

The partners will make all data collected, generated, and used in the project publicly available as X5GON participates in the Horizon 2020 Open Research Data Pilot. The Pilot guidelines for easy access and open access will be put into action before the 1st platform version becomes available (D2.2 Final Server Side Platform at M24). The partners committed to this milestone as the majority or almost all project data will have been created by that time.

Assuming any requests are made, X5GON data will be made available for sharing to qualified and certified parties before M26. This will be done by contacting the project coordinator. The sharing will be considering intellectual property interests, publication dates, privacy and confidentiality. The dataset will be shared via a publicly accessible disciplinary repository using descriptive metadata as required/provided by that repository. The access tools for accessing and using the X5GON data files will be common such as a desktop PC, the Windows 7 operating system, and Adobe Reader 9 software as well as open source tools.

An ftp client with specification will be setup by PS as the technical mechanism with detailed protocols for accessing the data and its reuse. This will be prepared before the 1st platform version becomes available (M24). An access analysis will be performed and a track record of data sharing will be kept. Data access will be open to our target groups (described in D8.1) and the general public and not restricted to specific research groups unless otherwise noted or required by the Grant or Consortium Agreements.

Finally, the data will be available to the public in an infrastructure with coordinated efforts between WP7 and WP8. We plan on adding resources and tools which are currently not publicly available but which all partners can access after the end of the project.

2.4.3 Long term preservation and Archiving

Following the guidelines of the Horizon 2020 Pilot on Open Research Data, all project data will be put in place for long-term preservation for a period of at least 10 years after the end of the project. The digital data will be preserved at the data centre at the PO. It will provide a high degree of robustness, security, GDPR compliance, and overall necessary infrastructure (communications, power, UPS replacement power supply, fire prevention system, alarm system, etc.) and professionally trained staff for technical responses.

Finally, backup and copies of the digital data will be preserved and finally backup will be checked at intervals of two weeks. All other data types 4) to 5) such as deliverables, etc., will be secured by JSI and K4A on the same server setup and on its own server setup. If requested, access to the data will be provided via contact with K4A's contact personnel. As the project continues it will be easier to project the volume of the data and the associated costs for preservation.

6. APPENIX II - ETHICS AND PRIVACY

We provide a detailed description of four identified groups of entities that will provide data during the project, here we give a description of each group according to case study, target audience and ethical issues with strategies for solving these issues. The tables will be revised in following report (M36) to reflect changes in data approaches towards these groups.

Role	To provide educational material to be translated in different language(s), and data mined.
Ethics and Privacy implications	No processing of personal or sensitive data related to Authors/Teachers will take place during the project, whereas any personal data included would be expected to be covered by legitimate interest. However, unconsented (manual or automatic) translation of teaching material constitutes copyright infringement.
Strategy	1. Authors/Teachers in pilots will be informed in detail about their participation, and will be called to sign a Copyright Transfer Agreement. 2. A copyright transfer agreement between VideoLectures.Net & X5GON will be drafted, provided that teachers have transferred proprietary rights to VideoLectures.Net.

Table 1: VideoLectures.Net authors of educational material (instructors, professors and educators (Teachers)

Role	To provide non-educational material (captions, forum text, tweets) to be translated and/or processed for according to the proposed research agenda (i.e., for automatic quality assurance, learning analytics, personalisation analysis etc.).
Ethics and Privacy implications	Videos, text and data produced in VideoLectures.Net users, such as fora, tweets etc. are considered to be copyright protected, in case they are original. Published material will normally pass legitimate interest test registered users' data may be processed based on contract, but opt-out (right to object) as to individual

	profiling may need to be enabled. In addition, content produced by Users may contain or be linked to personal data and/or other identifying information.
Strategy	All users entering personal data and other data related to translated material will be called to agree to a Terms Of Use document, where they will be called to explicitly consent to further re-processing of their data, according to the project's specifications.

Table 2: VideoLectures.Net users of educational material

Role	To provide instructional materials of any form.
Ethics and Privacy implications	Copyrighted materials shall not be used within the project.
Strategy	Authors are asked to specify the license for their material.

Table 3: Polimedia authors of educational material

Role	To provide feedback and usage statistics concerning OER recommended by the X5GON software within the local learning management system.
Ethics and Privacy implications	All feedback and usage statistics must be provided on a voluntary basis.
Strategy	Participation in the pilot project is voluntary. This will be documented by declarations of informed consent.

Table 4: Polimedia users of educational material

Role	To provide instructional materials of any form.
Ethics and Privacy implications	Copyrighted materials shall not be used within the project.
Strategy	Authors are asked to specify the license for their material. Only OER will be added to the repositories.

Table 5: virtUOS authors of educational material

Role	To provide feedback and usage statistics concerning OER recommended by the X5GON software within the local learning management system.
Ethics and Privacy implications	All feedback and usage statistics must be provided on a voluntary basis.

Strategy	Participation in the pilot project is voluntary. This will be documented by declarations of informed consent.

Table 6: virtUOS users of educational material

Role	X5GON will crawl educational and scientific OER repositories, websites and governmental infrastructures.
Ethics and Privacy implications	All data and knowledge collected during this phase are considered as Public Domain data. Part of the collected data may contain private data. Legitimate interest or contracts with such users may serve as the basis for personal data processing.
Strategy	• Any personal data and/or other identifying information obtained during this phase will be processed according to the GDPR and national legislation • The project will crawl content and data licensed under Creative Commons, allowing it to be freely processed to serve the projects objectives.

Table 7: Other users outside the consortium

Role	To provide various services and data during the project e.g., translations, feedback, computer-related services etc.
Ethics and Privacy implications	All data and knowledge that will be produced by RP, needs to be shared and re-used according to the specifications in this deliverable. TraMOOC will process personal data obtained from RP according to GDPR rules and DP legislation in the Member State where data processing will be carried out). Employee data is generally processed based on legitimate interest.
Strategy	• The project will follow the Consortium and Grant agreements as legal basis for specific activities that RP personnel will undertake during the project, specific data and knowledge that RP will produce. • Concerning students or teachers in focus groups that might undertake platform testing tasks, their OER documents will constitute their own intellectual property. We will setup a copyright agreement in order for the project to use/process this work.

Table 8: X5GON Research personnel (RP)

6.1 DATA TYPES AND PROJECT TASKS

Below, we provide a table which shows the Data Management Plan by project Task defined with type(s) of data expected to be generated by each task specified in the DoA. As mentioned above, the missing information will be specified in the next iterations of this deliverable in M24 and M36.

Task1.1. Quality Assurance Models (M1-M12, Leader UCL)	
Data Description	We will deliver a report on the derivation of several quality assurance models based on a thorough literature survey, data and tools available.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	Deliverable D1.1 Report on Quality assurance models. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are “Public” and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report. The publication schedule is outlined in DoA section 1.3.2. WT2 list of deliverables.
Data Preservation and Archiving	Preserved according to project specification.
Task 1.2. Initial Content Representation (M6-M24, Leader UCL)	
Data Description	We start experimenting with current user and OER resource data. The models-built focus on representing the subject matter contained in resources. We also attempt to use quality related features for personalization. This task is at its initial stage and we will refine and update the deliverables in the future updates of this deliverable.
Data Types	crawled data, user data, source code
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON and CSV format. Platform source code is documented and available through a Github repository open to all project members– quality is controlled by developers who follow coding conventions specified within the developers' group.

	Additionally, unit testing and timely review of source code are enforced to preserve quality of code.
Data Access and Intellectual Property	Data is stored in UCL data centres in Distributed JSON and CSV format where only project developers have access. Private data is anonymized and follow GDP. As the representation work is at its early stages, source code is currently open to all members of the consortium. We plan to make the source code publicly available as soon as we develop the initial working models.
Data Sharing and Reuse	Quality services will be publicly available through a secure public REST API. User data and private data will be anonymised. Representations will be available only to the project partners through a secure private REST API.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers
Task 1.3. Evaluation of representations (M6-M36, Leader UCL)	
Data Description	We will deliver a report on the evaluation of several quality assurance models based on multiple evaluation metrics outlining how the final model is selected. We will further explain additional diagnostics run on the final model to sanity check the model.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	Deliverable D1.2 Report on selected and evaluated quality assurance models. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are “Public” and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report. The publication schedule is outlined in DoA section 1.3.2. WT2 list of deliverables.
Data Preservation and Archiving	Preserved according to project specification.
Task 1.4. Advanced content representations (M12-M36, Leader UCL)	

Data Description	We start experimenting with current user and OER resource data. The models built focus on representing the subject matter contained in resources. We also attempt to use quality related features for personalization. Additionally, we explore how to update user models based on what resources learners consume. This task is at its initial stage and we will refine and update the deliverables in the future updates of this deliverable.
Data Types	Crawled data, Topic enriched data, source code
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON and CSV format. Platform source code is documented and available through a Github repository open to all project members– quality is controlled by developers who follow coding conventions specified within the developers' group. Additionally, unit testing and timely review of source code are enforced to preserve quality of code.
Data Access and Intellectual Property	Data is stored in UCL data centres in Distributed JSON and CSV format where only project developers have access. Private data is anonymized and follow GDPR. As the representation work is at its early stages, source code is currently open to all members of the consortium. We plan to make the source code publicly available as soon as we develop the initial working models.
Data Sharing and Reuse	Quality services will be publicly available through a secure public REST API. User data will be anonymised. Advance Representations will be available only to the project partners through a secure private REST API.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.

Table 9: *WP1 detailed data description per task*

Task 2.1 Requirements & Architecture of the platform (M1-M6, Leader JSI)	
Data Description	We deliver a report on the platform development requirements and architecture. This will serve as a roadmap for developing services in WP1, WP2, WP3 and WP4.
Data Types	Report

Data standards and description - Organisation, Documentation and Metadata	Deliverable D2.1 Requirements & Architecture of the platform. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are “Public” and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues for the report as it is a public deliverable.
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report. The publication schedule is outlined in DoA section 1.3.2. WT2 list of deliverables.
Data Preservation and Archiving	Preserved according to project specification.
Task 2.2 Setting-up the platform (M6-M22, Leader JSI)	
Data Description	We setup the platform responsible for acquiring, enriching, processing and storing the data.
Data Types	Crawled data, user data, source code
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through PostgresQL database. Platform source code is documented and available through a public Github repository – quality is controlled by developers who follow coding conventions specified within the developers group.
Data Access and Intellectual Property	Data will be stored on PO cloud in a PostgresQL database where only project developers have access. User data will be anonymized following lone with GDPR. Source code has no access or IPR issues – code is publicly available.
Data Sharing and Reuse	Crawled data will be publicly available through a secure public REST API. User data will be anonymized and available only to the project partners through a secure private REST API.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.
Task 2.3 Integration of the newly developed components (M6-M24, Leader JSI)	

Data Description	We include newly developed components into the existing platform architecture. Newly developed components would run as services on the PO cloud. Communication between the services and the platform would occur through a secure REST API which would proxy requests.
Data Types	Crawled data, user data, source code
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through PostgresQL database. Mechanism for integrating developed components will be documented and available in a public Github repository - quality is controlled by developers who follow coding conventions specified within the developers group.
Data Access and Intellectual Property	Data will be stored on PO cloud in a PostgresQL database where only project developers have access. User data will be anonymized following line with GDPR. Source code has no access or IPR issues – code will be publicly available.
Data Sharing and Reuse	Data will be shared between components through a secure REST API.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.
Task 2.4 Visualization and User Interface (M6-M24, Leader JSI)	
Data Description	We plan to develop a user interface for analysing and monitoring the platform. Analysis will be described using different visualization components.
Data Types	Crawled data, user data, source code
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through PostgresQL database. Visualization methods will be documented and available in a public Github repository – quality is controlled by developers who follow coding conventions specified within the developers group.
Data Access and Intellectual Property	Data will be stored on PO cloud where only project developers have access. For source code there is no IPR issues – code is publicly available.
Data Sharing and Reuse	Visualizations will be provided within the user interface developed in this task.

Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.
Task 2.5 Final product and services development (M18-M36, Leader JSI)	
Data Description	Within this task we will finalize and continually monitor processes of the platform.
Data Types	Crawled data, user data, source code
Data standards and description - Organisation, Documentation and Metadata	Data will be stored on PO cloud in a PostgresQL database where only project developers have access. User data will be anonymized in line with GDPR. Source code has no access or IPR issues – code will be publicly available.
Data Access and Intellectual Property	Data will be stored on PO cloud where only project developers have access. For source code there is no IPR issues – code is publicly available.
Data Sharing and Reuse	Data will be shared between components through a secure REST API.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.
Task 2.6 Evaluation (M12-M36, Leader JSI)	
Data Description	We will deliver a report on the platform evaluation in terms of analytics capabilities, scalability and usability.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	Deliverable D2.4 Final Evaluation Report. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are “Public” and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report. The publication schedule is outlined in DoA section 1.3.2. WT2 list of deliverables.
Data Preservation and Archiving	Preserved according to project specification.

Table 10: *WP2 detailed data description per task*

Task 3.1 Architecture and data for Learning Analytics Engine (M1-12, Leader NA)	
Data Description	Development of the learning analytics engine with analytics built from user navigations and the OERs
Data Types	Crawled data, user data, transcriptions of OER
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through mySQL database. Learning Analytics Engine source code will be documented and available in a public Github repository – quality is controlled by developers who follow coding conventions specified within the developers group.
Data Access and Intellectual Property	Data will be stored on PO cloud where only project developers have access. For source code there are no IPR issues – code is publicly available.
Data Sharing and Reuse	Data will be shared with other X5-GON Workpackages through a secure REST API.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.
Task 3.2 Learning Analytics Engine (M6-24, Leader NA)	
Data Description	Development of the learning analytics engine with analytics built from user navigations and the OERs
Data Types	Crawled data, user data, transcriptions of OER
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through mySQL database. Learning Analytics Engine source code will be documented and available in a public Github repository – quality is controlled by developers who follow coding conventions specified within the developers group.
Data Access and Intellectual Property	Data will be stored on PO cloud where only project developers have access. For source code there are no IPR issues – code is publicly available.
Data Sharing and Reuse	Data will be shared with other X5-GON Workpackages through a secure REST API.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.

Task 3.3 Cross-lingual issues in Learning Analytics (M12-30, Leader UPV)	
Data Description	N/A
Data Types	N/A
Data standards and description - Organisation, Documentation and Metadata	N/A
Data Access and Intellectual Property	N/A
Data Sharing and Reuse	N/A
Data Preservation and Archiving	N/A
Task 3.4 Privacy and user oriented guidelines (M12-30, Leader NA)	
Data Description	A set of guidelines for the questions of privacy
Data Types	No data expected
Data standards and description - Organisation, Documentation and Metadata	No data expected apart from limited examples for demonstration
Data Access and Intellectual Property	Guidelines will be shared publically
Data Sharing and Reuse	No data expected
Data Preservation and Archiving	No data expected

Table 11: *WP3 detailed data description per task*

Task 4.1 Architecture for real-time cross-site and cross-lingual user models (M1-M18, Leader JSI)	
Data Description	Development of user models for capturing and modelling behavioural information
Data Types	Crawled data, user data
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through PostgresQL database. User modelling architecture source code will be documented and available in a public Github repository – quality is controlled by developers who follow coding conventions specified within the developers group.
Data Access and Intellectual Property	Data will be stored on PO cloud where only project developers have access. For source code there is no IPR issues – code is publicly available.
Data Sharing and Reuse	Data will be shared between the platform and the user modelling service through a secure REST API.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.
Task 4.2 Recommendation engine for online learning materials (M6-M24, Leader JSI)	
Data Description	Development of recommendation models for online learning materials.
Data Types	Crawled data, user data

Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through PostgresQL database. Recommendation engine source code will be documented and available in a public Github repository – quality is controlled by developers who follow coding conventions specified within the developers group.
Data Access and Intellectual Property	Data will be stored on PO cloud where only project developers have access. For source code there is no IPR issues – code is publicly available.
Data Sharing and Reuse	Data will be shared through secure REST API. Additionally, a generated HTML containing recommendation data will be provided for embedding in external portals.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.
Task 4.3 Cross-site and cross-lingual recommendation (M6-M30, Leader JSI)	
Data Description	Extend the recommendation models developed in T4.2 with cross language and cross site support.
Data Types	Crawled data, user data
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through PostgresQL database. Source code developed in this task will be added to the code developed in Task 4.2.
Data Access and Intellectual Property	Data will be stored on PO cloud where only project developers have access. For source code there is no IPR issues – code is publicly available.
Data Sharing and Reuse	Data will be shared through secure REST API. Additionally, a generated HTML containing recommendation data will be provided for embedding in external portals.
Data Preservation and Archiving	Data preserved according to GDPR and project specification. Source code preserved according coding conventions specified by project developers.

Table 12: *WP4 detailed data description per task*

Task 5.1 Piloting on individual components (M6-M24, Leader UPV)	
Data Description	In-house assessment of components developed in WP1 to WP4
Data Types	Crawled data, metadata
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through PostgresQL database and an SQL-based relational database.
Data Access and Intellectual Property	Crawled data and metadata will be accessible through different services

	developed in WP1 to WP5 (eg. X5GON-TTP editor/player, OER recommendation module, etc).
Data Sharing and Reuse	Crawled data and metadata will be shared through API which feed the user interfaces.
Data Preservation and Archiving	Data preserved according to DGDR and project specifications.
Task 5.2 Piloting on integrated components (M13-M36, Leader UCL)	
Data Description	Larger educational networks assessing integrated components developed in WP1 to WP4
Data Types	Crawled data, metadata
Data standards and description - Organisation, Documentation and Metadata	Data stored in a JSON format – accessible through PostgresQL database.
Data Access and Intellectual Property	Crawled data and metadata will be accessible through different services developed in WP1 to WP5, as well as the X5GON dashboard developed in WP6.
Data Sharing and Reuse	Crawled data and metadata will be shared through a secure API which feed the user interfaces.
Data Preservation and Archiving	Data preserved according to DGDR and project specifications.

Table 13: *WP5 detailed data description per task*

Task 6.1 Evaluation of the initial OER Network (M1-M12, Lead: UCL)	
Data Description	We will deliver a report on the learner-centric user research that was conducted in year 1. The reported studies comprise evaluations of initial designs of inclusive OER interfaces for learners and teachers. Empirical evidence (anonymised) shows how OER and recommender systems can support cross-cultural learning journeys. Furthermore, the report summarises UCL's ongoing collaboration with partners to enable OER users and providers (including pilot sites) to take part in the initial OER Network.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	Deliverable D6.1 Report of the OER network model and interface design evaluation (M12). The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are "Public" and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.

Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report.
Data Preservation and Archiving	Preserved according to project specification.
Task 6.2 In-the-wild studies of the use and experiences with the initial X5GON services (M13-M24, Lead: UCL)	
Data Description	We will deliver a report on human-centred studies based around the initial X5GON services.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	D6.2 Report of in-the-wild studies investigating performance and usability of the initial services for virtual and real-world adaptive learning. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are “Public” and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report.
Data Preservation and Archiving	Preserved according to project specification.
Task 6.3 In-the-wild Studies of OER user Experience, Engagement and Enjoyment (M25-M36, Lead: K4A)	
Data Description	We will deliver a report on studies of OER user Experience, Engagement and Enjoyment. The report will contain follow-up studies on previous findings, as well as explorations of novel approaches to using OER.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	D6.3 Report of in-the-wild study of OER user Experience, Engagement and Enjoyment. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are “Public” and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.

Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report.
Data Preservation and Archiving	Preserved according to project specification.

Table 14: *WP6 detailed data description per task*

Task 7.1 Creation and sustainability of project online presence (M1-M36, Leader: K4A)	
Data Description	This task is concerned with establishing the project's visual identity, materials and website
Data Types	Website
Data standards and description - Organisation, Documentation and Metadata	No data generated
Data Access and Intellectual Property	No data access or IPR issues as the website is public.
Data Sharing and Reuse	No data generated
Data Preservation and Archiving	No data generated
Task 7.2 Communication with the public and networking (M1-M36, Leader: K4A)	
Data Description	We deliver a report on dissemination activities concerning the project's results which will be iterated throughout the life of the project M12, M24, M36.
Data Types	Report / dissemination materials
Data standards and description - Organisation, Documentation and Metadata	Deliverables *D7.2 First, Interim and Final real-world and online learning community engagement plan*. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are "Public" and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report. The publication schedule is outlined in DoA section 1.3.2. WT2 list of deliverables
Data Preservation and Archiving	Preserved according to project specification.
Task 7.3 Communication with governments and disabled communities (M1-M36, Leader: MIZS)	

Data Description	Input into reports on dissemination activities in accordance with project's activities.
Data Types	Report / dissemination materials
Data standards and description - Organisation, Documentation and Metadata	The Task Leader, the Work Package Leader and the Project Coordinator control the quality. Documents are public and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or intellectual property issues.
Data Sharing and Reuse	There are no data sharing issues.
Data Preservation and Archiving	Preserved according to project specification.
Task 7.4 Communication with businesses and exploitation communities (M1-M36, Leader: PS)	
Data Description	We deliver a report on dissemination activities concerning the project's results which will be iterated throughout the life of the project M12, M24, M36.
Data Types	Report / dissemination materials
Data standards and description - Organisation, Documentation and Metadata	Deliverables *D7.2 First, Interim and Final real-world and online learning community engagement plan*. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The documents are "Public" and will be featured on the X5GON website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report. The publication schedule is outlined in DoA section 1.3.2. WT2 list of deliverables
Data Preservation and Archiving	Preserved according to project specification.

Table 15: *WP7 detailed data description per task*

Task 8.1 Market analysis (M1-M6, Lead: PS)	
Data Description	This is an exploitation task concerning the market size and business potential of the project.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	Deliverables *D8.1 Market Analysis* are for internal purposes only. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The task deliverable is "Non-

	Public" and will not be featured on the project website.
Data Access and Intellectual Property	The reports will be accessible only to authorised project participants and the EU project officer.
Data Sharing and Reuse	As those are "Restricted" deliverables, they will not be made publicly available.
Data Preservation and Archiving	Preserved according to project specification.
Task 8.2 Impact assessment (M12-M36, Lead: PS)	
Data Description	This is an exploitation task concerning the impact of the project.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	Deliverables *D8.2 First, Interim* are for internal purposes only, and *Final Impact assessment report*. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The task deliverable is "Public" and will be featured on the project website.
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable.
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report. The publication schedule is outlined in DoA section 1.3.2. WT2 list of deliverables
Data Preservation and Archiving	Preserved according to project specification.
Task 8.3 Development of a business plan (M12-M36, Lead: PS)	
Data Description	This is an exploitation task concerning the industrial impact of the project.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	Deliverables *D8.3 Business plan draft and D8.4 Final business plan*. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The task deliverables are Confidential, only for members of the consortium (including the Commission Services).
Data Access and Intellectual Property	The reports will be accessible only to authorised project participants and the EU project officer.
Data Sharing and Reuse	As those are "Restricted" deliverables, they will not be made publicly available.
Data Preservation and Archiving	Preserved according to project specification.

Table 16: *WP8 detailed data description per task*

Task 9.1 Project administration (M1-M36, Lead: UCL)	
Data Description	This is a project management task with a report.
Data Types	No data is generated in this task.
Data standards and description - Organisation, Documentation and Metadata	N/A
Data Access and Intellectual Property	N/A
Data Sharing and Reuse	N/A
Data Preservation and Archiving	N/A
Task 9.2 Quality and Risk Management (M1-M36, Lead: UCL)	
Data Description	This is an project quality management task.
Data Types	Report
Data standards and description - Organisation, Documentation and Metadata	Deliverables *D9.1 Ethical Data Management and Data Management Plan: year 1 (M12), D9.2 (M24), D9.4: First year report (12), D9.5: Second year report (M24), D9.6: Final report (36) and D9.3 (M36), D9.7: Project Handbook (M36)*. The quality is controlled by the Task Leader, the Work Package Leader and the Project Coordinator. The task deliverables are Confidential, only for members of the consortium (including the Commission Services).
Data Access and Intellectual Property	No data access or IPR issues as the report as it is a public deliverable
Data Sharing and Reuse	There are no data sharing issues. The results will be made publicly available. Publications or any other use of this output should reference the X5GON project in accordance to EC Grant rules and the respective deliverable – report. The publication schedule is outlined in DoA section 1.3.2. WT2 list of deliverables
Data Preservation and Archiving	Preserved according to project specification.

Table 17: *WP9 detailed data description per task*