Horizon 2020 Grant Agreement N° 688095

SlideWiki

Large-scale pilots for collaborative OpenCourseWare authoring, multiplatform delivery and learning analytics

Topic: ICT-20-2015

Start date of project: January 1, 2016 Duration: 36 months

D12.4 Data Management Plan

Work Package 12: Project Management

Due date of deliverable: 30/06/2016
Actual submission date: 30/08/2017

Revision: Version 1.0

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS

Dissemination Level		
PU	Public	x
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)	

Author(s)	Alexandra Garatzogianni, Steffen Lohmann (Fraunhofer)
Contributor(s)	Benjamin Wulff (Fraunhofer)
Reviewer(s)	Kostis Pristouris (ATHENA)

Disclaimer:

The information in this document reflects only the authors' views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/her sole risk and liability.

Table of Contents

1 Introduction

This deliverable outlines the strategy for data management to be followed throughout the course of the SlideWiki project by formulating a Data Management Plan for the datasets used within the context of the project. Moreover, this plan includes the descriptions of dataset lifecycle, stakeholder behaviours, and best practices for data management, data management guidelines, and templates for data management used in the SlideWiki project. This plan will be updated at every milestone cycle.

Based on the Guidelines for FAIR Data management in H2020[1], Data Management in H2020[2] and Linked Data Life Cycle (LDLC)[3], we present the data management guideline for SlideWiki as follows:

1. Data Reference Name – a naming policy for datasets.

2. Dataset Content, Provenance and Value – general descriptions of a dataset, indicating whether it is aggregated or transformed from existing datasets, or original datasets from data publishers.

3. Standards and Metadata – descriptions about the format and underlying standards, under which the metadata shall be provided to enable machine-processable descriptions of dataset (supporting data transformation of Any2RDF and RDF2Any).

4. Data Access and Sharing – it is envisaged that all datasets are freely accessed under the Open Data Commons Open Database License (ODbL). Exceptions shall be stated clearly.

5. Archiving, Maintenance and Preservation – locations of physical repository of datasets shall be listed for each dataset.

This deliverable will be updated at the completion of every milestone cycle, in case significant changes have been made, aiming thus to take into account any additional decisions or newly identified best practices.

Briefly stated, the Data Management Plan (DMP) outlines the datasets that will be generated or collected during the project's lifetime highlighting the following information:

1. How datasets will be exploited/shared/licensed. For those that cannot be shared, the reasons why are explained.
2. Which standards are followed for publishing datasets.
3. Which strategies are used for curation and archiving of datasets.

[1] European Commission, [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] European Commission, [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[3] Linked Data Stack, [Online]. Available: http://stack.linkeddata.org

The first version of this deliverable outlines the a strategy for data management to be followed throughout the course of the project, in terms of data management guidelines and a template that can be instantiated for all datasets corresponding to project outputs.

This deliverable will be periodically updated to take account of additional decisions or best practices adopted during the project lifetime. At the end of the project, it will include individual Data Management Plans for the ensuing datasets (or groups of related datasets). The plan addresses a number of questions related to hosting the data (persistence), appropriately describing the data (data value, relevant audience for re-use, discoverability), access and sharing (rights, privacy, limitations) and information about the human and physical resources expected to carry out the plan.

1.1 Purpose and Scope

A Data Management Plan (DMP) is a formal document that specifies ways of managing data throughout a project, as well as after the project is completed. The purpose of DMP is to support the life cycle of data management, for all data that is/will be collected, processed or generated by the project. A DMP is not a fixed document, but evolves during the lifecycle of the project.

SlideWiki aims to increase the efficiency, effectiveness and quality of education in Europe by enabling the creation, dissemination and use of widely available, accessible, multilingual, timely, engaging and high-quality educational material (i.e., OpenCourseWare). More specifically, the open-source SlideWiki platform (available at SlideWiki.org) will enable the creation, translation and evolution of highly-structured remixable OCW can be widely shared (i.e. crowdsourced). Similarly to Wikipedia for encyclopaedic content, SlideWiki allows

(1) to collaboratively create comprehensive OCW (curricula, slide presentations, self-assessment tests, illustrations etc.) online in a crowdsourcing manner,

(2) to semi-automatically translate this content into more than 50 different languages and to improve the translations in a collaborative manner and

(3) to support engagement and social networking of educators and learners around that content. SlideWiki is already used by hundreds of educators, thousands of learners. Several hundred comprehensive course materials are available in SlideWiki in dozens of languages.

The major block of these aims is the heterogenic nature of data formats used by various educational institutions, which vary extensively. Examples of the most popular formats used include CSV, XLS, XML, PDF, and RDB and presentation files, such as PowerPoint files (ppt), OpenDocumentPresentation (odp), PDF, or ePub, etc.

By applying DCAT-AP standard for dataset descriptions and making them publicly available, the SlideWiki DMP covers the 5 key aspects (dataset reference name, dataset description, standards and metadata, access, sharing, and re-use, archiving and preservation), following the guidelines on Data Management of H2020[4].

[4] European Commission, [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

While the collaborative authoring of engaging, inclusive, standard-compliant and multi-lingual OpenCourseWare content is deemed a crucial and still neglected component of advancing educational technology, there is a plethora of systems covering other stages of educational value chains, such as Learning Management (e.g. ILIAS, Moodle , OLAT), Learning Content, Learning Delivery (e.g., OpenCast, FutureLearn), Learning Analytics Systems and social networks. Instead of trying to incorporate as much functionality as possible in a single system, SlideWiki will facilitate the exchange of learning content and educational data between different systems in order to establish sustainable educational value chains and learning eco-systems. Open (learning metadata) standards such as SCORM are first steps in this direction, but truly engaging, inclusive and multi-lingual value chains can be only realized if we take the content structuring to the next level and employ techniques such as interactive HTML5 (which can be presented meanwhile on almost all devices) and fine-grained semantic structuring and annotation of OpenCourseWare. In order to implement this concept, a relational multi-versioning data structure will be employed, which is dynamically mapped to an ontology, following the MVC paradigm and exhibiting Linked Data.

1.2 Structure of the Deliverable

The rest of this deliverable is structured as follows: Section 2 presents the data life-cycle of SlideWiki, the related stakeholders and 13 best practices for data management. Section 3 describes basic information required for datasets of the Slide Wiki project, and relevant guidelines. Section 4 presents DMP templates for data management. Each dataset has a unique reference name. Each data source and each of the transformed form will be described with metadata, which includes technical descriptions about procedures and tools used for the transformation, and common-sense descriptions for external users to better understand the published data. The Open Data Commons Open Database License (ODBL) is taken as the default data access, sharing, and re-use policies of the datasets used within the context of SlideWiki. Physical location of datasets shall be provided.

2 Data Lifecycle

The SlideWiki platform is a Linked Data platform, whose data ingestion and management follow the Linked Data Life Cycle (LDLC). The LDLC describes the technical process required to create datasets and manage their quality. To ease the process, best practices are described to guide dataset contributors in the SlideWiki platform.

Formerly, data management was executed by a single person or a working group that would also take responsibility for data management. With the popularity of the Web and the widely distributed data sources, data management has shifted to a service of a large stakeholder ecosystem.

2.1 Stakeholders

For the SlideWiki platform, the stakeholders who influence the data management belong to the following categories:

1. **Data Source Publisher/Owner:** This category refers to organisations providing datasets to the SlideWiki platform. The communication between SlideWiki and DSPO is limited to two cases: SlideWiki downloads data from DSPO, and DSPO uploads data to SlideWiki.

2. **Data End-User:** This category refers to persons and organisations who use the SlideWiki platform in order to access, view and share OpenCourseWare (OCW).

3. **Data Wrangler:** This category refers to persons who integrate heterogenic datasets into the SlideWiki platform. They are able to understand both the terminology used in the datasets and the SlideWiki data model, and their role is to ensure that the data integration is semantically correct.

4. **Data Analyser:** This category refers to persons who provide query results to end-users of SlideWiki. They may need to use data mining software.

5. **System Administrator and Platform Developer:** This category refers to persons responsible for developing and maintaining the SlideWiki platform.

2.2 The Generic SlideWiki Data Value Chain

Within the context of SlideWiki, we structure the generic data value chain as follows:

1. *Discover*. An end-user query can require data to be collected from many datasets located within different entities and potentially also distributed in different countries. Datasets hence need to be located and evaluated. For SlideWiki, the evaluation of datasets results in dataset metadata, which is one of the main best practices in the Linked Data community. DCAT-AP is used as the metadata vocabulary.

2. *Ingest and make the data machine processable*. In order to realise the value creation stage (integration, analyse, and enrich), datasets in different formats are transformed into a machine processable format. In the case of SlideWiki, it is the RDF format. The conversion pipeline from heterogenic datasets into an RDF dataset is fundamental. A Data Wrangler is responsible for the conversion process. For CSV datasets, additional contextual information is required to make the semantics of the dataset explicit.

3. *Persist*. Persistence of datasets happens throughout the whole data management process. When a new dataset comes into the SlideWiki platform, the first data persistence is to backup this dataset and the ingestion result of this dataset. Later data persistence is largely determined by the data analysis process. Two strategies used in data persistence are (a) keeping local copy – copy the dataset from DSPO to the SlideWiki platform; (b) caching, to enhance data locality to increase the efficiency of data management.

4. *Integrate, analyse, enrich*. One of the data management tasks is to combine a variety of datasets and find out new insights. Data integration needs both domain knowledge and technical knowhow. This is achieved by using a Linked Data approach enriched with a shared ontology. The life cycle of Linked Data ETL process starts from the extraction of RDF triples from heterogenic datasets, and storing the extracted RDF data into a storage, that is available for SPARQL querying. The RDF storage can be manually updated. Then, the interlinking and data fusion is carried out, which use ontologies in several public Linked Data sources and creates the Web of Data. In contrast to a relational data warehouse, the Web of Data is a distributed knowledge graph. Based on Linked Data technologies, new RDF triples can be derived, and new enrichment is possible. Evaluation is necessary to control the quality of new knowledge, which further results in searching more data sources, and performing data extraction.

5. *Expose*. The result of data analysis will be exposed to end-users in a clear, salient, and simple way. The SlideWiki platform is a Linked Data platform, whose outcomes include (a) metadata description about the results; (b) a SPARQL endpoint for the metadata; (c) a SPARQL endpoint for the resulting datasets; (d) a user-friendly interface for the above results.

2.3 Best Practices

The SlideWiki platform is a Linked Data platform. Considering the best practices for publishing Linked Data, the following 13 stages are recommended in order to publish a standalone dataset, 6 of them are vital (marked as must).

1. *Provide descriptive metadata with locale parameters:* Metadata must be provided for both human users and computer applications. Metadata provides DEU with information to better understand the meaning of data. Providing metadata is a fundamental requirement when publishing data on the Web, because DSPO and DEU

may be unknown to each other. Then, it is essential to provide information that helps DEU – both human users and software systems, to understand the data, as well as other aspects of the dataset. Metadata should include the following overall features of a dataset: The title and a description of the dataset; the keywords describing the dataset; the date of publication of the dataset.; the entity responsible (publisher) for making the dataset available; the contact point of the dataset; the spatial coverage of the dataset; the temporal period that the dataset covers; the themes/categories covered by a dataset. Locale parameters metadata should include the following information: the language of the dataset; the formats used for numeric values, dates and time.

2. *Provide structural metadata:* Information about the internal structure of a distribution must be described as metadata, for this information is necessary for understanding the meaning of the data and for querying the dataset.
3. *Provide data license information:* License information is essential for DEU to assess data. Data re-use is more likely to happen, if the dataset has a clear open data license.
4. *Provide data provenance information*: Data provenance describes data origin and history. Provenance becomes particularly important when data is shared between collaborators who might not have direct contact with one another.
5. *Provide data quality information*: Data quality is commonly defined as "fitness for use" for a specific application or use case. The machine readable version of the dataset quality metadata may be provided according to the vocabulary that is being developed by the DWBP working group, i.e., the Data Quality and Granularity vocabulary.
6. *Provide versioning information*: Version information makes a dataset uniquely identifiable. The uniqueness enables data consumers to determine how data has changed over time and to identify specifically which version of a dataset they are working with.
7. *Use persistent URIs as identifiers*: Datasets must be identified by a persistent URI. Adopting a common identification system enables basic data identification and comparison processes by any stakeholder in a reliable way. They are an essential pre-condition for proper data management and re-use.
8. *Use machine-readable standardised data formats*: Data must be available in a machine-readable standardised data format that is adequate for its intended or potential use.
9. *Data Vocabulary*: Standardised terms should be used to provide metadata, Vocabularies should be clearly documented, shared in an open way, and include versioning information. Existing reference vocabularies should be re-used where possible.
10. *Data Access:* Providing easy access to data on the Web enables both humans and machines to take advantage of the benefits of sharing data using the Web infrastructure. Data should be available for bulk download. APIs for accessing data should follow REST (REpresentational State Transfer) architectural approaches. When data is produced in real-time, it should be available on the Web in real-time. Data must be available in an up-to-date manner and the update frequency made explicit. If data is made available through an API, the API itself should be versioned separately

from the data. Old versions should continue to be available.

11. *Data Preservation:* Data depositors willing to send a data dump for long term preservation must use a well-established serialisation. Preserved datasets should be linked with their "live" counterparts.
12. *Feedback*: Data publishers should provide a means for consumers to offer feedback.
13. *Data Enrichment*: Data should be enriched whenever possible, generating richer metadata to represent and describe it.

3 Data Management Plan Guidelines

In this section, we describe guidelines of the DMP of SlideWiki. In order to enable the export of SlideWiki content on Data Web, as a proof -of-concept the RDB2RDF mapping tool Triplify[5] is employed in order to map SlideWiki content to RDF and publish the resulting data on the Data Web. The SlideWiki Triplify Linked Data interface will soon be available. With regard to Social Networking, at the current stage, SlideWiki supports limited social networking activities. In the future, it is envisaged that SlideWiki users will be able follow other users, slides and decks, they can discuss and comment on slides and decks, login/register to system using their Facebook account and share slides/decks on popular social networks (e.g. Facebook, LinkedIn, G+, Twitter).

3.1 Privacy and Security

It is a fact that educational data mining needs to cope with large unstructured (live) data which needs to be handled, transferred and translated into interpretable structured datasets[6]. Analog to other data sensitive domains there is the critical question of privacy and (learning) data protection. Also the irresolution of which data is important from a pedagogical/technical point of view is still a complex intent and open question, taking the complex and individual learning process into account.

It is mandatory for the collection of such data that the involved learner is using campus tools and platforms that support tracking of learning action. These analytics remain an immature field that has yet to be implemented broadly across a range of institutional types, student populations and learning technologies. So-called Learning Record Stores are the next generation tracking and reporting repositories that support ideas like the Tin Can protocol and the successor xAPI. Open analytic solutions as provided by the Open Academic Analytics Initiative[7] (OAAI), which is already fostering the collection of and meaningful interpretation of data across learning institutes.

Given that the central aim of this consortium is to provide benefit to the European community, the project will prefer open data and free, open tools and provide the resources developed in the project under the Creative Commons Attribution 4.0 License (CC-BY)[8]. This license allows the learning material to be shared and adapted for any purpose, even commercially. The only restriction is attribution: linking to the source and indicating the changes made. Released in November 2013, CC-BY 4.0 improves its predecessor CC-BY 3.0, as it is an international license and includes databases.

In order to prevent data loss and to ensure SlideWiki users' privacy, a sophisticated backup and archiving strategy, guaranteeing data security will be implemented and developed within the context of WP1. Within the context of T1.3 Privacy, Data Security, Backup and Archiving, all OpenCourseWare content stored in SlideWiki, be it slides, presentations,

[5] http://triplify.org

[6] Horizon Report: 2013 Higher Education, L. Johnson, S. Adams Becker, M. Cummins, V. Estrada, A. Freeman, und H. Ludgate. The New Media Consortium, Austin: Texas (2013)

[7] Open Academic Analytics Initiative https://confluence.sakaiproject.org/x/8aWCB

[8] https://creativecommons.org/licenses/by/4.0/

questionnaires, diagrams, images, user data etc.), is regularly backed-up and archived. In SlideWiki all content (also versioning histories and prior revisions) will be made available via Linked Data, SPARQL interfaces, APIs and data dumps. Incremental updates will be published, so that the interested parties (e.g. large universities, school authorities) can host their own synchronized SlideWiki mirrors (similar to services such as arXiv.org or DBLP), while ensuring that all privacy and data security regulations are enforced.

In D1.8, privacy and data security report (M28), it is outlined how SlideWiki implements all relevant privacy and data security regulations and best practices.

Moreover, at the SlideWiki website[9], the Statement of Data Protection Conditions can be accessed, and it provides the following info with regard to personal data:

"The Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. (Fraunhofer-Gesellschaft) takes the protection of your personal data very seriously. When we process the personal data that is collected during your visits to our Web site, we always observe the rules laid down in the applicable data protection laws. Your data will not be disclosed publicly by us, nor transferred to any third parties without your consent."

In the following sections, we explain what types of data we record when you visit our Web site, and precisely how they are used:

3.1.1 Recording and processing of data in connection with access over the Internet

When you visit our Web site, our Web server makes a temporary record of each access and stores it in a log file. The following data are recorded, and stored until an automatic deletion date:

1. IP address of the requesting processor
2. Date and time of access
3. Name and URL of the downloaded file
4. Volume of data transmitted
5. Indication whether download was successful
6. Data identifying the browser software and operating system
7. Web site from which our site was accessed
8. Name of your Internet service provider

The purpose of recording these data is to allow use of the Web site (connection setup), for system security, for technical administration of the network infrastructure and in order to optimize our Internet service. The IP address is only evaluated in the event of fraudulent access to the network infrastructure of the Fraunhofer-Gesellschaft.

Apart from the special cases cited above, we do not process personal data without first obtaining your explicit consent to do so. Pseudonymous user profiles can be created as stated under web analysis (see below).

[9] https://slidewiki.org/dataprotection

3.1.2 Orders

If you order information material or other goods via our website, we will use the address data provided only for the purpose of processing your order.

3.1.3 Use and transfer of personal data

All use of your personal data is confined to the purposes stated above, and is only undertaken to the extent necessary for these purposes. Your data is not disclosed to third parties. Personal data will not be transferred to government bodies or public authorities except in order to comply with mandatory national legislation or if the transfer of such data should be necessary in order to take legal action in cases of fraudulent access to our network infrastructure. Personal data will not be transferred for any other purpose.

3.1.4 Consent to use data in other contexts

The use of certain services on our website, such as newsletters or discussion forums, may require prior registration and involves a more substantial processing of personal data, such as longer-term storage of email addresses, user IDs and passwords. We use such data only insofar as it has been sent to us by you in person and you have given us your express prior consent for this use. For example, we request your consent separately in the following cases:

3.1.4.1 Newsletters and press distribution

In order to register for a newsletter service provided by the Fraunhofer-Gesellschaft, we need at least your e-mail address so that we know where to send the newsletter. All other information you supply is on a voluntary basis, and will be only if you give your consent, for example to contact you directly or clear up questions concerning your e-mail address. If you request delivery by post, we need your postal address. If you ask to be included on a press distribution list, we need to know which publication you work for, to allow us to check whether specific publications are actually receiving our press material.

As a general rule, we employ the double opt-in method for the registration. In other words, after you have registered for the service and informed us of your e-mail address, you will receive an e-mail in return from us, containing a link that you must use to confirm your registration. Your registration and confirmation will be recorded. The newsletter will not be sent until this has been done. This procedure is used to ensure that only you yourself can register with the newsletter service under the specified e-mail address. You must confirm your registration as soon as possible after receiving our e-mail, otherwise your registration and email address will be erased from our database. Until we receive your confirmation, our newsletter service will refuse to accept any other registration requests using this e-mail address.

You can cancel subscriptions to our newsletters at any time. To do so, either send us an e-mail or follow the link at the end of the newsletter.

3.1.4.2 Visitors' books and forums

If you wish to sign up for an Internet forum run by the Fraunhofer-Gesellschaft, we need at least a user ID, a password, and your e-mail address. For your own protection, the registration procedure for this type of service, like that for the newsletters, involves you confirming your request using the link contained in the e-mail we send you and you giving your consent to the use of further personal data where this is necessary to use the forum.

You can cancel your registration for this type of service at any time, by sending us an e-mail via the Web page offering the service.

As a general rule, the content of visitors' books and forums is not subject to any form of monitoring by the Fraunhofer-Gesellschaft. Nevertheless, we reserve the right to delete posted contributions and to prohibit users from further use of the service at our own discretion, especially in cases where posted content contravenes the law or is deemed incompatible with the objectives of the Fraunhofer-Gesellschaft.

3.1.5 Cookies

We do not normally use cookies on our Web site, but in certain exceptional cases we may use cookies which place technical session-control data in your browser's memory. These data are automatically erased at the latest when you close your browser. If, exceptionally, one of our applications requires the storage of personal data in a cookie, for instance a user ID, we will point out you to it.

Of course, it is perfectly possible to consult our Web site without the use of cookies. Please note, however, that most browsers are programmed to accept cookies in their default configuration. You can prevent this by changing the appropriate setting in the browser options. If you set the browser to refuse all cookies, this may restrict your use of certain functions on our Web site.

3.1.6 Security

The Fraunhofer-Gesellschaft implements technical and organizational security measures to safeguard stored personal data against inadvertent or deliberate manipulation, loss or destruction and against access by unauthorized persons. Our security measures are continuously improved in line with technological progress.

3.1.7 Links to Web sites operated by other providers

Our Web pages may contain links to other providers' Web pages. We would like to point out that this statement of data protection conditions applies exclusively to the Web pages managed by the Fraunhofer-Gesellschaft. We have no way of influencing the practices of other providers with respect to data protection, nor do we carry out any checks to ensure that they conform to the relevant legislation.

3.1.8 Right to information and contact data

You have a legal right to inspect any stored data concerning your person, and also the right to demand their correction or deletion, and to withdraw your consent for their further use.

In some cases, if you are a registered user of certain services provided by the Fraunhofer-Gesellschaft, we offer you the possibility of inspecting these data online, and even of deleting or modifying the data yourself, via a user account.

3.1.9 Acceptance, validity and modification of data protection conditions

By using our Web site, you implicitly agree to accept the use of your personal data as specified above. This present statement of data protection conditions came into effect on October 1st, 2013. As our Web site evolves, and new technologies come into use, it may become necessary to amend the statement of data protection conditions. The Fraunhofer-Gesellschaft reserves the right to modify its data protection conditions at any time, with effect as of a future date. We recommend that you re-read the latest version from time to time.

3.2 Dataset Content, Provenance and Value

3.2.1 What dataset will be collected or created?

i. Used Datasets:
   a. Continuously generated Web Server logs and (Google) analytics of project's website access;
   b. Continuously generated Social Media engagement data.

ii. Produced Datasets:
   a. Aggregated analytics of the courses developed within the framework;
   b. Aggregated statistics of networking and engagement data produced as part of D10.4 and D10.5 reporting, usage statistics of the framework.

3.2.2 What is its value for others?

This will ensure flexibility, adaptability and an improved user experience (UX).

3.3 Standards and Metadata

3.3.1 Which data standards will the data conform to?

SlideWiki aims to be a long-lasting open-standard based incubator for the collaborative creation of OpenCourseWare in Europe. The evaluation of the quality of use, based on standards like ISO 25010, will also produce recommendations to improve SlideWiki user interfaces, novel interaction paradigms, information architecture components, etc.

Furthermore, the distribution of the learning material to the following educational platforms

- Massive Open Online Courses (MOOCs)
- Learning Management Systems (LMSs)
- Interactive eBooks
- Social Networks

will be facilitated by SlideWiki's standard compliant HTML5, RDF/Linked Data and SCORM compatible content model, within the context of WP5. In the context of WP6, Secondary Education Trial, gold standards for the reconciliation of different open data sources of the city and of external organisations, such as the Spanish National Library or DBpedia will be generated. This activity is being also transferred into other cities, such as Madrid, and a similar expansion such as that of CoderDojo activities is expected.

3.3.2 What documentation and metadata will accompany the data?

Following the best practices for data on the web, the technical and user documentation of the platform will be constantly updated (MS3). In T1.4, Semantic search, an intuitive search facility for content, structure, metadata, provenance and revision history of the educational material will be designed and implemented.

Within the context of the Semantic representation of SlideWiki (D2.1), existing ontologies and vocabularies for semantic representation of OpenCourseWare material and enhancement of these for capturing SlideWiki representations will be reviewed and the resulting vocabulary will support representation of content, structure, metadata, provenance, and revision history.

For D4.2, SlideWiki SEO plans and appropriate strategies such as integrating embedded and structured metadata into SlideWiki pages as well as using smart URLs will be implemented in order to increase the visibility of SlideWiki content among popular search engines.

With regard to T4.4, Search engine optimization, embedded and structured metadata will be integrated into SlideWiki pages following vocabularies recognized by the main search engines like Schema.org so that its visibility in search engines and results pages of SlideWiki is improved. Another strategy is providing mechanisms like RDF2HTML converter for SEO in RDFaware search engines. Smart URLs will be implemented as more user-friendly URLS (e.g., using http://slidewiki.org/semantic-web/ to refer to a deck about Semantic Web). SlideWiki uses Ajax for client-side interactions and one problem we are dealing with is how to facilitate indexing of dynamic Ajax pages by search engines. To resolve this issue we will define suitable URL patterns and SEO strategies for making dynamically loaded content fragments more visible to search engines.

3.4 Data Access and Sharing

3.4.1 Which data is open, re-usable and what licenses are applicable?

The SlideWiki project aims at creating widely available, accessible, multilingual, timely, engaging and high-quality educational material (i.e., OpenCourseWare).

In particular, the Open Data Commons Open Database License (ODbL) to open datasets is adopted as a project's best practice. Suitable applicable licenses (such as ODBL), anonymization of personal data, possibility and suitability for reuse, and the long term management of the data resources in compliance with the LOD lifecycle and best practices will be implemented where applicable.

Overall only 28 out of the 100 courses have a truly open license, the vast majority (i.e. 57 out of 100) are restricting reuse to non-commercial scenarios (i.e. CC-BY-NC-SA), which is not open licensing according to the Open Definition. Often, for example, if courses are offered with a fee or the training organization is a for-profit organization, the non-commercial restriction thus prevents reuse. With regard to content acquisition, an inventory of existing material (e.g. PowerPoint presentations, PDFs, images etc.), which can be used for the creation of new OCW will be created. Particular attention will be given to license clearance, so that the content can be published under the least restrictive conditions.

Furthermore, new opportunities are emerging in online education (technology-enhanced learning), largely driven by the availability of high quality online learning materials, also known as Open Educational Resources (OERs). OERs can be described as teaching, learning and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use or repurposing by others depending on which Creative Commons license is used[10].

SlideWiki aims to provide solutions for the very limited OCW availability, the fragmented educational content, the restrictive licenses (e.g. non-commercial) and the lack of inclusiveness or accessibility of educational content. This will be achieved by establishing an Open Educational Content and an educational ecosystem with focus on accessibility, which will be further supported by multilingualism. The created content itself will be published in an open manner without usage restrictions or license costs. However the content itself shall keep records with regard to authorship, modifications and possibly also its usage.

The SlideWiki consortium aims to benefit the European community. Therefore, open data and free, open tools, such as the Creative Commons Attribution 4.0 License (CC-BY)[11] will be preferred. The Creative Commons Attribution 4.0 License will allow the learning material to be shared and adapted for any purpose, even commercially. The only restriction is attribution: linking to the source and indicating the changes made. Released in November 2013, CC-BY 4.0 improves its predecessor CC-BY 3.0, as it is an international license and includes databases.

With regard to open data, where possible, the project will make use of existing open source libraries and make its efforts highly visible and open to external input aiming thus to attract collaboration rather than competition. During the trials, innovative approaches will be implemented, such as the use of crowdsourcing techniques among the participants and the collaboration with key stakeholders, such as university researchers, in generating gold standards for the reconciliation of different open data sources. The course material will be completely open to the community that is represented by the organisation, with the

[10] Atkins, D. E., Brown, J. S. & Hammond, A. L. (2007) A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities. The William and Flora Hewlett Foundation

[11] https://creativecommons.org/licenses/by/4.0/

intention of incorporating additional materials from additional potential users that are not members of the organisation, and with the objective of making the course materials a reference for such domain.

3.4.2 How will open data be accessible and how will such access be maintained?

Data should be available for bulk download. APIs for accessing data should follow REST architectural approaches. Real-time data should be available on the Web in real-time. Data must be available in an up-to-date manner, with explicitly demonstrated update frequency. For data available through an API, the API itself should be versioned separately from the data. Old versions should continue to be available.

3.5 Data Archiving, Maintenance and Preservation

3.5.1 Where will each dataset be physically stored?

Datasets will be initially stored in a repository hosted by the SlideWiki server, or one of participating consortium partners. Depending on its nature, a dataset may be moved to an external repository, e.g. European Open Data Portal, or the LOD2 project's PublicData.eu.

3.5.2 Where will the data be processed?

Datasets will be processed locally at the project partners. Later, datasets will be processed on the SlideWiki server, using cloud services.

3.5.3 What physical resources are required to carry out the plan?

Hosting, persistence, and access will be managed by the SlideWiki project partners. They will identify virtual machines, cloud services for long term maintenance of the datasets and data processing clusters.

3.5.4 What are the physical security protection features?

For open accessible datasets, security will be taken to ensure that the datasets are protected from any unwanted tampering, to guarantee the validity.

3.5.5 How will each dataset be preserved to ensure long-term value?

Since the SlideWiki datasets will follow Linked Data principles, the consortium will follow the best practices for supporting the life cycle of Linked Data, as defined in the EU-FP7 LOD2 project. This includes curation, reparation, and evolution.

3.5.6 Who is responsible for the delivery of the plan?

Members of each WP should enrich this plan from their own aspect.

4 Data Management Plan Template

The following template will be used to establish plans for each dataset aggregated or produced during the project.

4.1 Data Reference Name

A data reference name is an identifier for the dataset to be produced [1].

Description	A dataset should have a standard name within SlideWiki, which can reveal its content, provenance, format, related stakeholders, etc.
Metadata	Interpretation, guideline, and software tools shall be given, provided, or indicated for generating, interpreting data reference names.

Table 1 - Template for Data Reference Name

4.2 Dataset Content, Provenance and Value

When completing this section, please refer to questions and answers 1-2 in Section 3.1

Description	A general description of the dataset, indicating whether it has been: aggregated from existing source(s) created from scratch transformed from existing data in other formats generated via (a series of) other operations on existing dataset The description should include reasons leading to the dataset, information about its nature and size and links to scientific reports or publications that refer to the dataset.
Provenance	Links and credits to original data sources
Operations performed	If the dataset is a result of transformation or other operations (including queries, inference, etc.) over existing datasets, this information will be retained.
Value in Reuse	Information about the perceived value and potential candidates for exploiting and reusing the dataset. Including references to datasets that can be integrated for added value.

Table 2 - Template for Dataset Content, Provenance and Value

4.3 Standards and Metadata

When completing this section, please refer to questions and answers 3-4 in Section 3.2

Format	Identification of the format used and underlying standards. In case the DMP refers to a collection of related datasets, indicate all of them.
Metadata	Specify what metadata has been provided to enable machine-processable descriptions of dataset. Include a link if a DCAT-AP representation for the dataset has been published.

Table 3 - Template for Standards and Metadata

4.4 Data Access and Sharing

When completing this section, please refer to questions and answers 5-6 in Section 2.3

Data Access and Sharing Policy	It is envisaged that all datasets in the SlideWiki project should be freely accessed, in particular, under the Open Data Commons Open Database License (OdbL). When an access is restricted, justifications will be cited (ethical, personal data, intellectual property, commercial, privacy-related, security-related)
Copyright and IPR	Where relevant, specific information regarding copyrights and intellectual property should be provided.
Access Procedures	To specify how and in which manner can the data be accessed, retrieved, queried, visualised, etc.
Dissemination and reuse Procedures	To outline technical mechanisms for dissemination and re-use, including special software, services, APIs, or other tools.

Table 4 - Template for Data Access and Sharing

4.5 Archiving, Maintenance and Preservation

When completing this section, please refer to questions and answers 6-12 in Section 3.4

Storage	Physical repository where data will be stored and made available for access (if relevant) and indication of type: ● SlideWiki partner owned ● societal challenge domain repository ● open repository ● other
Preservation	Procedures for guaranteed long-term data preservation and backup. Target length of preservation.
Physical Resources	Resources and infrastructures required to carry out the plan, especially regarding long-term access and persistence. Information about access mechanism including physical security features.
Expected Costs	Approximate hosting, access, maintenance costs for the expected end volume, and a strategy to cover them.
Responsibilities	Individual and/or entities are responsible for ensuring that the DMP is adhered to the data resource.

Table 5 - Template for Archiving, Maintenance and Preservation

5 Storage of the Datasets

All project-related datasets are stored on either GitHub or our SlideWiki servers for public access. The so-called Learning Record Stores are the next generation tracking and reporting repositories that support ideas like the Tin Can protocol and the successor xAPI. Open analytic solutions as provided by the Open Academic Analytics Initiative (OAAI) already fostering the collection of and meaningful interpretation of data across learning institutes. The Learning Locker data repository stores learning activity data generated by xAPI-compliant (Tin Can) learning activities. A Learning Activity Database and the mechanism for logging and collecting all activity data in the platform will be developed. It will seamlessly track user actions, associate them with semantically rich events of the activity data model and store them in the Learning Activity Database. The mechanism will be implemented on top of state-of-the-art open source big data logging and ingestion tools (such as Apache Flume and Apache Kafka) such that it can exploit the dynamic scale-out infrastructure of WP1 and achieve efficient data ingestion for large volume and rate of user events.

The Consortium will make sure, that all OpenCourseWare content stored in SlideWiki, be it slides, presentations, questionnaires, diagrams, images, user data etc., is regularly backed-up and archived.

6 FAIR Data Management Principles

The SlideWiki consortium monitors the application of the FAIR Data management principles, also listed here below.

6.1 Making Data Findable, Including Provisions for Metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
What naming conventions do you follow?
Will search keywords be provided that optimize possibilities for re-use?
Do you provide clear version numbers?
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

6.2 Making Data Openly Accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
How will the data be made accessible (e.g. by deposition in a repository)?
What methods or software tools are needed to access the data?
Is documentation about the software needed to access the data included?
Is it possible to include the relevant software (e.g. in open source code)?
Where will the data and associated metadata, documentation and code be deposited?
Preference should be given to certified repositories which support open access where possible.
Have you explored appropriate arrangements with the identified repository?
If there are restrictions on use, how will access be provided?
Is there a need for a data access committee?
Are there well described conditions for access (i.e. a machine readable license)?
How will the identity of the person accessing the data be ascertained?

6.3 Making Data Interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

6.4 Increase Data Re-Use (Through Clarifying Licences)

How will the data be licensed to permit the widest re-use possible?
When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
How long is it intended that the data remains re-usable?
Are data quality assurance processes described?

6.5 Allocation of Resources

What are the costs for making data FAIR in your project?
How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
Who will be responsible for data management in your project?
Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

6.6 Data Security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
Is the data safely stored in certified repositories for long term preservation and curation?

6.7 Ethical Aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

6.8 Other Issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

7 Conclusion

This deliverable outlines the guidelines and strategies for data management within the context of the SlideWiki and will be fine-tuned and extended throughout the course of the project. Following the guidelines on FAIR Data Management in H2020[12], Data Management in H2020[13], we described the purpose and scope of datasets of SlideWiki, and specified the datasets management for the SlideWiki project. Five kinds of stakeholders related to the SlideWiki data management plan are identified and described: original data producer, data wrangler, data analyser, system administrator/developer, and data end-user; generic data flow chain of SlideWiki is listed and explained: data discover, data ingest, data persist, data analyse, and data expose. Following the best practices of Linked Data Publishing, we specified the 13 steps of best practices for the SlideWiki dataset management. Based on the above, we present DMP guidelines for SlideWiki, and DMP templates for data management process during the lifetime of the SlideWiki project.

[12] European Commission, [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[13] European Commission, [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

8 References

[1] European Commission, [Online]. Available:

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission, [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] Linked Data Stack, [Online]. Available: http://stack.linkeddata.org