Prêt-à-LLOD

D1.7. Data Management Plan

Author(s): Víctor Rodríguez-Doncel, Mariano Rico

Date: 30 June 2019

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-ICT-29b

Grant Agreement No. 825182

Prêt-à-LLOD - Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors

D1.7. Data Management Plan

Deliverable Number: D1.7
Type: ORDP
Dissemination Level: PU
Delivery Date: 30 June 2019
Version: 1
Author(s): Víctor Rodríguez-Doncel, Mariano Rico.

Document History

Version Date	**Changes**	**Authors**
May 2019	initial version, following the template. A set of questions is made to the consortium	Víctor Rodríguez-Doncel, Mariano Rico
June 2019	version complete	Víctor Rodríguez-Doncel
June 2019	reviews	Thierry Declerck, Elena Montiel-Ponsoda

Table of Contents

List of Acronyms

API	Application Program Interface
DMP	Data Management Plan
DOI	Document Object Identifier
ELG	European Language Grid
FAIR	Findable, Accessible, Interoperable and Reusable
GDPR	General Data Protection Regulation
ODRL	Open Digital Rights Language
ORDP	Open Research Data Pilot
RDF	Resource Description Framework

D1.7. Data Management Plan

1. Introduction

1.1. Scope

This document contains the initial version of the Prêt-à-LLOD Data Management Plan (DMP). The DMP is a living document and will be regularly updated. Succesive stable versions of the DMP will be published in M24 and M36. This document is complemented by "D5.2 Policy-based language Data Management" (due in M24) and it is related to "D7.1 Ethics Requirements I" (delivered in M3).

The Data Management Plan adheres to and complies with the "H2020 Data Management Plan – General Definition" given by the European Commission (EC) online[1], where the DMP is described as follows:

> *"A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and reusable (FAIR), a DMP should include information on:*
>
> - *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)"*

Prêt-à-LLOD adopts policies compliant with the official FAIR guidelines [1] (findable, accessible, interoperable and re-usable), as mandated by the EC. Also, Prêt-à-LLOD participates in the Open Research Data Pilot (ORDP[2]) and is obliged to deposit the produced research data in a research data repository, as per Art. 29.3 of the Grant Agreement.

This Section 1 concludes with the presentation of preliminary concepts; Section 2 is the Data Management Plan itself and follows the template proposed by the EC[3].

[1] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[2] https://www.openaire.eu/what-is-the-open-research-data-pilot

[3] http://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx

1.2. Preliminary concepts

Zenodo

Zenodo[4] is a general-purpose open-access repository much used for publishing deliverables and data of H2020 projects. Zenodo exposes the data to OpenAIRE[5], a network of Open Access repositories to support the EC publication policies.

Resources in Zenodo (consequently also in OpenAire) are identified with a Document Object Identifier (DOI), they can be versioned, and there are good chances that they will enjoy long term preservation. Moreover, common search engines such as Google Scholar or Microsoft Research are aware of the assets hosted at Zenodo and they enjoy high visibility.

A Prêt-à-LLOD community has been created in the Zenodo portal.

`https://zenodo.org/communities/pret-a-llod/`

Both deliverables and research data will be published in this Zenodo community.

Prêt-à-LLOD Data Portal

The "**Prêt-à-LLOD Data Portal**" will be the data portal of the project and will host the description of relevant datasets (metadata). The data portal will also be used to host newly created language resources as long as their size is manageable. This data portal will be open to the general public and users will be able to search for datasets, visualize their description and eventually download the resource itself.

The Prêt-à-LLOD Data Portal will be built using the CKAN[6] technology (a standard software package for data portals) and Linghub, a Linked Data based portal already describing language resources [3]. Due to many internet users being already familiar with CKAN, their visual appearence will be respected, only being customized for the needs of this project and using the corporate image of Prêt-à-LLOD. A CKAN data access API will be exposed to offer infomation on the datasets metadata.

2. Data Management Plan

The sections of this document and the questions hereinafter are taken from the *Horizon 2020 FAIR Data Management Plan (DMP) template.* The use of the template is recommended by the EU commission.

2.1 Data summary

1. Data summary
a) What is the purpose of the data collection / generation and its relation to the objectives of the project?

[4] http://zenodo.org

[5] Open Access Repositories in Europe, http://openaire.eu

[6] Comprehensive Knowledge Archive Network, http://ckan.org

The declared objectives of this project are:

- to support the exchange of multilingual cross-sectoral data
- to develop interoperable language technology services and language data
- to favour the sustainability of language technologies and language resources

Consequently, the collection and generation of data are core activities for this project, and its purpose can be summarized as ‘prepare linguistic data so that it can power multilingual applications in a digital single market’.

An initial list of 52 processing activities is documented in annex A of “D7.1. Ethic Requirements I” [2].

b) What types and formats of data will the project generate / collect?

The vast number of formats that will be handled by this project does not allow a preliminary enumeration. Data in different formats will be collected and eventually transformed. The preferred type for the generated data is the one which most favours interoperability — this will be RDF (Resource Description Framework[7]) in its different serializations.

The types of data in Table 1 have been identified, with respect to their meaning:

Name	**Description**
Catalogue metadata	Description of existing data resources
Open linguistic data	Existing open linguistic data already available prior to Prêt-à-LLOD.
New linguistic data	Transformation of existing resources or creation of new resources in the LLOD (Linguistic Linked Open Data cloud [8]) by the Prêt-à-LLOD project. These assets are considered results of this project.
Experiment-related data	Data produced in the course of reports generations, execution of experiments (e.g. experiments for automated linking), etc., often related to research publications.

Table 1. Types of data generated or collected by Prêt-à-LLOD according to their meaning

The following types of data have been identified, according to their openness.

Name	**Description**
Private to partners	Available to the partner who owns it
Available to partners	Not public, only available to the partners. No Non-Disclosure Agreements (NDAs) are not necessary and the Consortium Agreement suffices.
Published as Open Data	Both public and available with an open license.

Table 2. Types of data in Prêt-à-LLOD, according to their openness.

c) Will you re-use any existing data and how?

This project will extensively reuse linguistic resources, eventually republishing them possibly after a transformation.

d) What is the origin of the data?

[7] https://www.w3.org/RDF

Datasets available in the LLOD cloud and resources available in other data catalogues (OLAC[8], LRE Map[9], META-SHARE[10], Clarin[11], Retele[12]), and private data resources that will not be exposed.

e) What is the expected size of the data?

The size of the data is broken down per data type:
Catalogue metadata: ~1Gb
Open linguistic data: not to be stored by Prêt-à-LLOD
New linguistic data: ~100Gb
Experiment-related data: ~10Gb
These figures have been estimated considering the experience of some of the Prêt-à-LLOD partners in the past FP7-funded LIDER project[13].

f) To whom might the data be useful ('data utility')?

Two large communities are identified: (i) the community of researchers and practitioners of linguistics and social sciences and (ii) the community of computer scientists and developers with interests in natural language processing.

2.2. FAIR data

2. FAIR data

2.1 Making data findable, including provisions for metadata

a) Are the data produced and / or used in the project discoverable and identifiable?

Catalogue metadata will be available at the Prêt-à-LLOD Data Portal**.** Data will be discoverable because each dataset will have a description using the standard DCAT vocabulary[14] (see Figure 1) --in particular, DCAT-AP: the "DCAT application profile for European data portals", developed in the framework of the EU ISA Programme[15], which has become a de-facto standard.
New linguistic data produced by this project will also be offered through the Prêt-à-LLOD Data Portal. Identifiability will be supported because each data in all datasets will have a unique identifier (IRI[16]) accessible through the Web.
Experiment-related data will be published in Zenodo, in turn connected with OpenAIRE and every major indexer of scientific documents. Eventually, small pieces of data will also be available from source code repositories (e.g. a Gitlab instance hosted in the premises of the coordinating institution in Ireland).

[8] Open Language Archives Community, http://www.language-archives.org/
[9] Language Resources and Evaluation Map, http://lremap.elra.info/
[10] META-SHARE, http://www.meta-share.org/
[11] Common Language Resources and Technology Infrastructure, https://www.clarin.eu/
[12] Catálogo de Recursos Lingüísticos en Linked Data, http://catalogo.retele.linkeddata.es/
[13] Grant agreement ID 610782, http://lider-project.eu
[14] Data CatalogueVocabulary, http://www.w3.org/TR/vocab-dcat
[15] https://ec.europa.eu/isa2
[16] Internationalized Resource Identifier, RDF 3987, https://tools.ietf.org/html/rfc3987

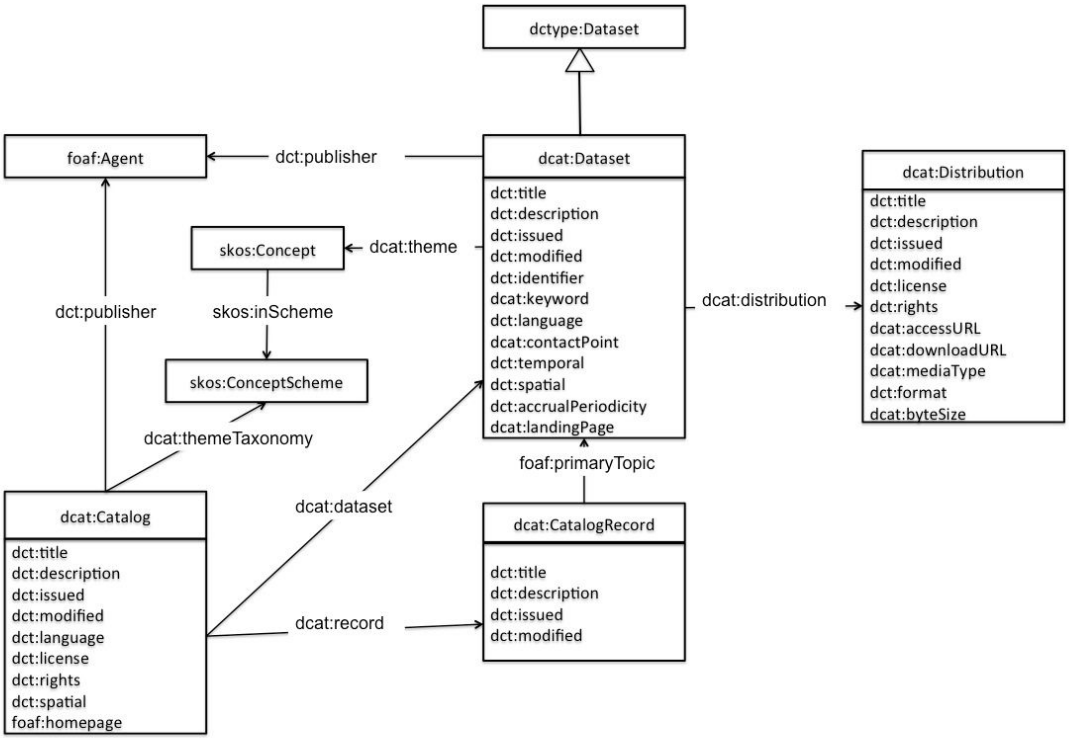

Figure 1. Metadata elements in the DCAT specification

b) What naming conventions do you follow?

We defined the following two policies:
(i) Identification of datasets. Datasets are identified by an slug (a user friendly and URL valid name of a resource).
(ii) URI minting policy, to be decided at a later stage of the project.

c) Will search keywords be provided that optimize possibilities for re-use?

The use of keywords is natural in the Prêt-à-LLOD Data Portal and in Zenodo. Zenodo’s commitment with FAIR policies is made explicit[17].

d) Do you provide clear version numbers?

The use of a semantic versioning is inherent to Zenodo. The Prêt-à-LLOD Data Portal will also provide versioning and provenance mechanisms.

e) What metadata will be created?

The stored data are described by using the standard metadata schema Qualified Dublin Core and DCAT.
Zenodo's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichments[18].

2.2 Making data openly accessible

a) Which data produced and / or used in the project will be made openly available as the default?

[17] https://about.zenodo.org/principles/

[18] https://schema.datacite.org/

	By default, all metadata in Zenodo and Prêt-à-LLOD Data Portal are openly available as soon as the record is published.
b)	How will the data be made accessible (e.g. by deposition in a repository)?
	All data are stored in Zenodo and the Prêt-à-LLOD Data Portal. All metadata in Zenodo and the Prêt-à-LLOD Data Portal are publicly available in an Open Access modality. Eventually, language resources created by Prêt-à-LLOD will be introduced in the well-known language resources catalogues (OLAC, LRE Map, META-SHARE, Clarin, Retele).
c)	What methods or software tools are needed to access the data?
	The extensive use of open specifications and consolidated standards grants that there is no need for special software tools to access the data. Eventually, experiment-related data may require of additional software (e.g. GATE[19]).
d)	Is documentation about the software needed to access the data included?
	Not necessary for the time being.
e)	Is it possible to include the relevant software (e.g. in open source code)?
	Not necessary for the time being.
f)	Where will the data and associated metadata, documentation and code be deposited?
	The following data stores are foreseen: — The Prêt-à-LLOD Data Portal store defined in Section 1.1, hosted in Ireland, for catalogue data and some newly generated resources. — A Gitlab instance, hosted in Ireland, for small datasets. — Zenodo for research-related data. Data will also be mirrored, whenever possible, in projects with whom liaisons will be established. In particular, relevant data will be also passed to the ELG (European Language Grid[20]) project, “Towards the Primary Platform for Language Technologies in Europe”.
g)	Have you explored appropriate arrangements with the identified repository?
	The aforementioned repositories are either self-managed by Prêt-à-LLOD partners, or they are already deemed for these purposes. Formal arrangements with ELG are pending to be done.
h)	If there are restrictions on use, how will access be provided?
	No restrictions have been identified at this stage, but the commercial interest of the partners may lead to the creation of private data.

[19] GATE, General Architecture for Text Engineering, https://gate.ac.uk/

[20] European Language Grid, H2020 grant id 825627, https://www.european-language-grid.eu/

i) Is there a need for a data access committee?	
	No. Rules that concern governing of data access of the partner institutions will be followed and implemented, together with the FAIR principles followed by this Plan.
j) Are there well described conditions for access (i.e. a machine readable license)?	
	Licenses in Prêt-à-LLOD Data Portal are represented in a machine readable form, using the most common metadata descriptor (dct:license, see Figure 1) pointing to standard URL licenses'. Whenever linked data is published, standard practices will be followed to publish the rights information [5]. Moreover, in some cases, a fully machine readable representation of the licenses is given using the Open Digital Rights Management Language (ODRL)[21]. Licenses from the RDFLicense dataset are also used [4].
k) How will the identity of the person accessing the data be ascertained?	
	Not necessary for the time being.
2.3 Making data interoperable	
a) Are the data produced in the project interoperable?	
	Both Zenodo and the Prêt-á-LLOD Data Portal use standard interfaces, protocols and metadata, etc. Using standard metadata schemas in Zenodo, metadata can easily be converted into other metadata schemas.
b) What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	
	DCAT (described above) and the CKAN schema[22] based on it. Linghub currently makes use of the META-SHARE OWL ontology [6].
c) Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?	
	Yes, see above (2.3.b).
d) 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?	
	The use of RDF as the meta-format grants the easy definition of links between equivalent metadata elements.
2.4 Increase data re-use (through clarifying licences)	
a) How will the data be licensed to permit the widest re-use possible?	
	Open by default, using the CC-BY license (Creative Commons 4.0 Attribution International[23]) unless this hampers the business model of our partners.

[21] ODRL https://www.w3.org/TR/odrl-model/, https://www.w3.org/TR/odrl-vocab/.

[22] https://dcat-ap-donl.readthedocs.io/projects/ckanext-dcatdonl/en/latest/schema.html

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

b) When will the data be made available for re-use?
Data will be made available as soon as it is created and no data embargoes are foreseen.
c) Are the data produced and / or used in the project useable by third parties, in particular after the end of the project?
Making data ready to use is the motto of this project, and every possible measure will be taken to maximize its usability.
d) How long is it intended that the data remains re-usable?
Data in the Prêt-à-LLOD Data Portal may not be supported after the end of the project, but because it will be mirrored in the ELG, long time preservation will be possible. Research data will enjoy long term preservation as it will be uploaded to Zenodo.
e) Are data quality assurance processes described?
No. Future versions of this DMP may include a definition of such process.

2.3. Allocation of resources

3 Allocation of resources
a) What are the costs for making data FAIR in your project?
None that is not foreseen in the Grant Agreement: making data FAIR is an explicit objective of this project.
b) How will these be covered?
Not applicable.
c) Who will be responsible for data management in your project?
Víctor Rodríguez Doncel (UPM) will be the responsible for the management of open data in this project. The management of private data will be responsibility of the partners having produced it.
d) Are the resources for long term preservation discussed?
The cooperation agreements with the ELG project are headed towards long term preservation.

2.4. Data security

4 Data security
a) Is the data safely stored in certified repositories for long term preservation and curation?
Most data (catalogue data, newly created resources) will contain data to be published under an open license. This data does not need any security measure whatsoever. For the case partners generate privative data with personal information, security measures

	will have to be adopted to comply with the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679).
b)	What provisions are in place for data security?
	Not yet described.

2.5. Ethical aspects

5 Ethical aspects	
a) Are there any ethical or legal issues that can have an impact on data sharing?	
	Ethical aspects have been extensively documented in Prêt-à-LLOD deliverables “D7.1 Ethics Requirements I” and in “D7.4 Ethics Requirements 4”.
b) Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?	
	See Prêt-à-LLOD deliverable “D7.1. Ethics Requirements I”.

2.6. Other issues

6 Other issues	
a) Do you make use of other national / funder / sectorial / departmental procedures for data management?	
	— National University of Ireland Galway (NUIG) is subject to a *“Insight Open Source Release Process”* procedure. — Universidad Politécnica de Madrid (UPM) is subject to “*Normativa sobre protección de resultados de investigación de la Universidad Politécnica de Madrid*” and “*Reglamento del comité de ética de actividades i+d+i de la Universidad Politécnica de Madri*d” These procedures are compatible with the provisions made in this data management plan.

3. References

[1] H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 (2016), Version 3, Directorate-General for Research & Innovation, European Commission.

[2] Noussias, A. (2019) D7.1 Ethics Requirements I. Prêt-à-LLOD deliverable.

[3] McCrae, J. P., & Cimiano, P. (2015). Linghub: a Linked Data based portal supporting the discovery of language resources. SEMANTiCS (Posters & Demos), 1481, 88-91.

[4] Rodríguez Doncel, V., Gómez-Pérez, A., & Villata, S. (2014). A dataset of RDF licenses. in Proc. of the 27th Int. Conf. on Legal Knowledge and Information System (JURIX), R. Hoekstra (Ed.), ISBN 978-1-61499-467-1, pp. 187-189, IOS Press

[5] Radulovic, F., Poveda-Villalón, M., Vila-Suero, D., Rodríguez-Doncel, V., García-Castro, R., & Gómez-Pérez, A. (2015). Guidelines for Linked Data generation and publication: An example in building energy consumption. Automation in Construction, 57, 178-187.

[6] McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., & Cimiano, P. (2015, May). One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. In European Semantic Web Conference (pp. 271-282). Springer, Cham.

[7] McCrae, J. P. (2019) D7.4 Ethics Requirements 4. Prêt-à-LLOD deliverable

[8] McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G. & Osenova, P. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In Proc. of the Tenth Int. Conf. on Language Resources and Evaluation, 2435-2441.