EUROPEAN LANGUAGE GRID

D5.4

Data Management Plan (version 1.0)

Authors:	Pawel Kamocki (ELDA), Khalid Choukri (ELDA), Valérie Mapelli (ELDA), Lucille Blanchard (ELDA), Mickaël Rigault (ELDA)
Dissemination Level:	Public
Date:	30-06-2019

About this document

Project	ELG – European Language Grid
Grant agreement no.	825627 – Horizon 2020, ICT 2018-2020 – Innovation Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2019, 36 months
Deliverable number	D5.4
Deliverable title	Data Management Plan (version 1.0)
Type	Report
Number of pages	20
Status and version	Final – Version 1.0
Dissemination level	Public
Date of delivery	Contractual: 30-06-2019 – Actual: 30-06-2019
WP number and title	WP5: Grid Content – Language Resources, Datasets, and Models
Task number and title	Task 5.4: Legal support, DMP and GDPR
Authors	Pawel Kamocki (ELDA), Khalid Choukri (ELDA), Valérie Mapelli (ELDA), Lucille Blanchard (ELDA), Mickaël Rigault (ELDA)
Reviewers	Kalina Bontcheva (USFD), Penny Labropoulou (ILSP)
Consortium	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Institute for Language and Speech Processing (ILSP), Greece
	University of Sheffield (USFD), United Kingdom
	Charles University (CUNI), Czech Republic
	Evaluations and Language Resources Distribution Agency (ELDA), France
	Tilde SIA (TILDE), Latvia
	Sail Labs Technology GmbH (SAIL), Austria
	Expert System Iberia SL (EXPSYS), Spain
	University of Edinburgh (UEDIN), United Kingdom
EC project officers	Philippe Gelin, Alexandru Ceausu
For copies of reports and other ELG-related information, please contact:	DFKI GmbH European Language Grid (ELG) Alt-Moabit 91c D-10559 Berlin Germany Dr. Georg Rehm, DFKI GmbH georg.rehm@dfki.de Phone: +49 (0)30 23895-1833 Fax: +49 (0)30 23895-1810 http://european-language-grid.eu © 2019 ELG Consortium

Table of Contents

List of Abbreviations

DMP	Data Management Plan
ELG	European Language Grid
ELDA	European Language Distribution Agency
EU	European Union
GDPR	General Data Protection Regulation
IPR	Intellectual Property Rights
ISLRN	International Standard Language Resource Number
LR(s)	Language Resource(s)
NLP	Natural Language Processing

Abstract

This document contains guidelines and corresponding templates for a Data Management Plan (DMP) to be used in the European Language Grid project (ELG). These templates shall be used as a set of best practices to document and manage Language Resources and other data collected and/or processed in the ELG project. As such, each LR (even versions of each LR) shall be accompanied with a data management plan. It addresses issues such as accompanying documentation, metadata description, legal aspects, preservation, sustainability plan, etc. The present document is the first version of the ELG DMP. Two additional versions will be delivered later on in the project as D5.5 at M24 and D5.6 at M36 (see the Grant Agreement).

Note: The DMP is of crucial importance within a highly data-centric project such as ELG. In this first version we focus on identifying the main issues and questions that pertain to data management in ELG. We plan to submit an updated version of the present document, D5.4, by M12.

1 Introduction

1.1 What is a Data Management Plan?

The H2020 Participant Portal manual[1] provides the following definition:

Data Management Plans (DMPs) are a key element of good data management. 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 Re-usable (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).

The implementation of a DMP is an obligation under article 29.3 of the H2020 Grant Agreement:

"Regarding the digital research data generated in the action ('data'), the beneficiaries must (…) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:

1. the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible; (…)
2. other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan' (…)"[2].

[1] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm (access: 16.04.2019)
[2] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

Having a DMP in place has also been part of the Language Technology best practices in use inside the community for more than a decade which emphasises the value of such a plan.

1.2 What are the objectives of the ELG Project?

LRs are absolutely necessary for the realization of the ELG project. It aims to address the challenges of the fragmentation of the LR market by providing access to a catalogue of LRs for a wide range of modalities (speech, text, audio, video, etc.). In order to meet this objective, the need for a DMP that focuses on the particular nature and lifecycle of the LRs is mandatory as these data will make up the majority of data processed in the course of the ELG project.

This project also aims to consolidate the European Language Technology community, among others, by organising workshops and events. In order to account for the various kind of data that may be collected and processed in the course of those events guidelines for management of such data has also been addressed in this document. A sensible DMP shall be put in place for all kinds of data collected and processed during the project.

1.3 What Data are Processed[3] in the ELG Project?

Language Resources

ELRA (European Language Resources Association) defines Language Resources as follows:

"The term Language Resource refers to a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users.

Examples of Language Resources are written and spoken corpora, computational lexica, terminology databases, speech collection, etc. Basic software tools are also important for the acquisition, preparation, collection, management, customisation and use of these Language Resources and other resources"[4].

Stakeholders' personal data collected during the ELG project

Personal data, e.g., collected during event registration, or operation of the ELG user platform.

Newsletter/Blog Content Data and Project Report Data

Copyright-protected items produced by project participants for dissemination purposes.

Website Traffic Data

Data gathered via the project's website and APIs (e.g., IP addresses, log data, navigation time, browser etc.).

Survey Data

Data gathered via surveys organized within the project.

---

[3] "Process" here is meant in the sense of Article 4.2 of the GDPR: set of operations which is performed on (…) data (…) such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction

[4] http://www.elra.info/en/about/what-language-resource/

1.4 What are the FAIR Principles and How are They Addressed in the ELG Project?

According to the Horizon 2020 requirements, members of the ELG consortium as well as EU funded projects should make all created or collected research data compliant with the FAIR principles defined as findability, accessibility, interoperability and reusability.

In order to assess the levels of FAIRness 14 metrics have been defined and inserted in this management plan for all resources. FAIRness within ELG is defined as follows.

Findability

For data to be findable, the ELG consortium will adopt an internal and a global persistent identifier attribution policy for all LRs to be made available via its platform. Such identifiers will follow the community practices to ensure identification and findability outside the ELG platform (DoI, ISLRN, etc.). Furthermore, all Language Resources offered on the ELG platform are to be described with the ELG metadata scheme that would also help improve findability. The metadata records can be exported to compatible LR catalogues to ensure that it can be also used to find the data once indexed in resource catalogues.

Accessibility

To be accessible, the ELG platform is cataloguing Language Resources by storing and documenting them. This will allow browsing and searching resources via various metadata elements. Most metadata elements associated with the LRs will be retrievable from the platform. ELG will make its metadata schema and the platform management software open and freely available for re-use. Preservation of the metadata will be ensured via export of ELG metadata records to other community platforms for long-term accessibility.

Interoperability

To be interoperable, the ELG platform is developing a metadata schema that is built on existing ones (META-SHARE, CMDI etc.). ELG will also advocate for the use of best practices within each of the Language Technology areas it addresses (data format, knowledge representation, etc.). An important added value of ELG will be the use of containerization techniques to make the data interoperable.

Re-usability

To allow re-usability the ELG platform will enforce the use of its metadata elements. It will also, through this DMP associated to each Language Resource, ensure that the purpose for which the data has been produced is documented with indication on the possibilities to re-use the data for other purposes (other technology developments). The ELG platform will also ensure that all IPR issues are cleared. In addition, ELG will ensure that the Language Resources are released with a clear and understandable data usage license. In particular, the ELG consortium will advocate for very permissive licenses whenever this is possible. Attention will be also paid to other legal issues that may hinder the re-usability of the resources (e.g., GDPR compliance) and legal support will be provided as part of the ELG helpdesk to complete the DMP.

The H2020 Guidelines on Data Management[5] contain a template for a Data Management Plan that aims to achieve these goals (see Section 4). This H2020 template has been extended to cover all categories of data produced within the ELG project.

However, to take into account questions of paramount importance for the consortium these guidelines and principles will be adapted for the management of LRs acquired or produced during the project. These adaptations have been made possible thanks to decades of experience in the collection and processing of Language Resources in projects such as FlareNet, META-NET, CRACKER and others. While participating in these projects, ELDA has developed an expertise on the formulation of sensible Data Management Plans for Language Resources which ensure Findability, Accessibility, Interoperability and Re-Usability of LRs, but are structured differently than the H2020 template and cover other critical aspects of data processing and management. Indeed the DMP for Language Resources (apart from the introductory Description) is based on the consortium's experience gathered from different European projects catering to the processing of Language Resources, as mentioned above. As such it is tailored to fit the specific challenges posed by the three main stages of LR 'lifecycle': Acquisition, Storage and Sharing, which are of crucial importance for the realization of the ELG project.

2 ELG Guidelines on Data Management for Language Resources

This section contains recommendations on how consortium members, ELG pilot projects and LR providers shall document, manage or otherwise process Language Resources.

In the current version of the DMP, the focus has been to put in place requirements of certain information items necessary to document LRs (e.g., language, copyright etc.). Revised versions shall be distributed by taking into account the actual resources to be uploaded on the ELG platform by LR providers, using a bottom-up approach. As an example, for both newly created and pre-existing LRs, the nature of the data should be specifically described as well as their initial purpose, and reference to applied policies should be included.

2.1 Data Acquisition

Pre-existing Language Resources

Basic description elements for LRs have been consolidated in META-SHARE[6] to define the most accurate criteria. In the course of the ELG project the following basic description elements shall be attached to the LRs for them to be integrated in the platform (the full ELG metadata set will be delivered in a separate deliverable):

- data type,
- languages,
- modality,
- access and availability,
- copyright and user license type,
- conditions for re-use, re-purposing, re-packaging
- documentation.

[5] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[6] http://www.meta-share.org/knowledgebase/homePage

In order to facilitate the discovery of LRs needed for the purpose of ELG, each criterion should be as specific as possible. In particular, all the elements that can facilitate IPR clearance and license negotiations with the right holders should be mentioned.

It should also be specified whether the LR contains personal data (within GDPR[7]). The consortium assumes that datasets collected before the entry into force of the GDPR (25 May 2018) have been collected in compliance with the adequate European and national legislation at the time of collection, and should be documented as such. However it must be noted that all acts of data processing performed after 25 May 2018, even on data collected before the entry into application of the GDPR, shall be made in compliance with the legislation, and documented within the DMP.

The data controller (the entity that defines the means and purposes of personal data processing) should be clearly identified in the DMP. If other entities than the controller process personal data within the project (e.g., subcontractors), this section should specify that these other entities (called data processors) will process personal data only on documented instructions from the controller and will not engage another processor without prior written authorisation from the controller. The controller shall also clearly specify the purposes for which personal data are processed, since further processing in contradiction with the intended objectives are prohibited.

New Language Resources

The ELG platform will take on board new LRs developed within the project or uploaded by potential providers planning to use the platform. To advance this purpose the consortium will advocate the use of ELG policies with regards to all aspects of the DMP. Each phase (specification, production, post-production) shall comply with the ELG recommendations.

The **specification** phase is crucial and sets:

- the description of data
- the use of standards and best practices widely adopted within the community, for curation, annotation, processing etc.
- the metadata should be fully compliant with ELG metadata
- the quality assessment
- the validation procedures
- the IPR issues (for data access and re-use etc.)
- the management of ethics, privacy, confidentiality (data anonymization, non-disclosure agreements)
- the packaging
- the promotion in case of distribution and sharing

The **production** phase includes:

- the actual production of the data
- the quality controls on a regular basis
- the validation

[7]Art. 4 of the GDPR defines personal data as "any information related to an identified or identifiable natural person".

The **post-production** phase comprises:

- the allocation of a unique identifier (ISLRN or another persistent Id)
- the attribution of the appropriate license (if such data can be licensed)
- the documentation
- the selection of the appropriate storage infrastructures (storage devices or facilities, portability of data must be anticipated to ensure the preservation)
- the bug reporting
- the maintenance
- the dissemination

It should also be specified whether the LR contains personal data (within the meaning of GDPR). We assume that datasets, collected before the entry into force of GDPR (25 May 2018), have been collected in compliance with the adequate European and national legislation at the time of collection. This has to be documented.

However it must be noted that data processing performed after 25 May 2018, even on data collected before the entry into application of the GDPR, shall be made in compliance with the legislation and documented within the DMP. The data controller (the entity that defines the means and purposes of personal data processing) should be clearly identified in the DMP. If other entities than the controller process personal data within the project (e.g., subcontractors), this section should specify that these other entities (called data processors) will process personal data only on documented instructions from the controller and will not engage another processor without prior written authorisation from the controller.

The purpose for which personal data will be processed within the project should be clearly stated. The data cannot be processed for other purposes. If personal data is included in the LR, the following principles should be observed:

- only data that is necessary to achieve the purpose of the project (stated in the previous section) can be processed (data minimisation);
- personal data should be accurate and, when necessary, kept up to date (data accuracy)
- personal data can only be processed if the data subject (the person that the data refer to) consented to the processing or, exceptionally, when another criterion of art. 6 of the General Data Protection Regulation is met (lawfulness of data processing)

Pursuant art. 35 of GDPR, when the processing of personal data within the project is likely to result in a high risk to the rights and freedoms of natural persons the data controller in cooperation with the data protection officer (if designated) shall carry out a risk assessment prior to the processing. The impact assessment should contain at least:

- a description of the envisaged processing operations and the purposes of the processing;
- an assessment of the necessity and proportionality of the processing operations in relation to the purposes;
- an assessment of the risks to the rights and freedoms of data subjects (the persons that the data relate to);
- the measures envisaged to address the risks, including safeguards (such as pseudonymisation, encryption, approval by an ethics committee), security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with the GDPR.

ELG requirements on all these elements are clearly stated in Section 3 and will be refined over time as the development of the ELG platform progresses.

2.2 Storage, Preservation and Access

The ability of a given resource to survive over time without any explicit and external financial support is the usual definition of sustainability in our field, even though it only describes the self-sustainable nature of a resource. Indeed, this definition does not encompass the aspects related to the availability and the use of the resource, the management of its rights (e.g., licensing), its potential customization, its "openness" or its improvements (updates, corrections, repackaging), etc.

In the context of research projects where data is produced (or existing data re-packaged), data preservation is essential in order to guarantee sharing and future use. The use of the ELG platform and its storage, preservation and access facilities will ensure that the LRs in question are properly taken care of. The DMP may provide for alternative possibilities for storage, preservation and access outside of the ELG environment.

Furthermore, if a dataset contains personal data, they should be kept up to date. If they were anonymized, a periodical evaluation of the results of anonymization should be carried out in order to ensure that the data are still to be regarded as anonymous, despite technological progress.

Pursuant art. 5 (1) e) of the GDPR, personal data will only be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed (storage limitation). Moreover, personal data will be stored in a manner that ensures appropriate security of the data, including protection against unauthorised or unlawful access and against accidental loss, destruction or damage, using appropriate technical or organisational measures (data integrity and confidentiality).

2.3 Sharing

LR sharing requires interoperability and access facilities which will be managed by the use of ELG containerisers, data formatting practices will also be promoted to enhance data sharing.

Today, a large set of licenses exists for data sharing. Licenses range from the LR-dedicated licenses by ELDA[8], and by META-SHARE[9] to the more permissive public license suites such as Creative Commons[10].

With the support of legal experts, ELDA has gained a long experience in managing IPR issues and supporting LR licensing. In order to find and define the best license for a specific use within ELG, ELDA will help rights-holders to answer critical questions such as:

- Are the data protected by an exclusive right (such as copyright or the sui generis database right)? If so, who holds it? Is a permission required to collect and re-use the data?
- Under which license should the data be shared (an LR-specific license? A Creative Commons license? A Free or Open Source Software License?)? Do we have all the necessary rights and permissions to share the data under this license?

[8] http://www.elra.info/Licensing.html
[9] http://www.meta-net.eu/meta-share/licenses
[10] http://creativecommons.org

- What kind of uses are allowed by the license under which the data are available? Should the data be used as is? Are we allowed to create derivatives? Can the data be used for commercial purposes?

The right holders will adopt a given license among the set of licenses that the ELG platform will support.

As part of the ELG legal helpdesk, ELDA will offer assistance on legal issues related to the management of data mainly on IPR and personal data issues[11] and provides access to legal advisors. These advisors are also available to provide assistance to complete DMPs.

3 Template for a Data Management Plan for ELG Language Resources

As indicated above, this DMP template for the management of LRs represents an extension to the H2020 template and the FAIR principles. However, the different sections of this template have been structured to comply with the FAIR principles.

Findability

For the LRs to be findable, ELG recommends that its members and participants comply with the use of identifiers, and use distribution channels recommended by ELG.

Accessibility

To comply with the FAIR principles, ELG recommends how data shall be packaged, stored, secured and shared.

Interoperability

All data sets shall comply to the ELG metadata scheme, whether from the start for resources produced thanks to ELG support or through conversion of existing resources.

Reusability

The reusability principle will be applied through recommendations made by ELG on issues such as licensing and IPR issues. the new LRs produced thanks to ELG support will have to comply with Open Data Principles and permissive licensing schemes.

3.1 General context

- What is the initial purpose of LR production?
- What are the institutions involved in the LR collection?
- What language data will be processed in the project and how?
- Will personal data be processed in the project?
    - What types of personal data will be processed?
    - What ground for lawfulness will the processing be based upon (e.g., consent)?
    - Who will be the data controller?
    - Will there be any data processors (persons or entities processing data on behalf of the controller)? Who will they be?

[11] http://www.elda.fr/en/services-around-lrs/legal-support-helpdesk

- Is the processing of the personal data necessary and proportional in relation to the project's purpose?
- What are the risks to the rights and freedoms of the persons that the data relate to?
- What measures are envisaged to mitigate the risks and to ensure protection of the personal data (e.g., pseudonymization, encryption)?

3.2 Data Acquisition

Production of Language Resources

- What standards/best practices will be implemented in the production process? How will they ensure interoperability of the data?
- What metadata will accompany the data?
- What are/were the estimated costs of data production? How will/were they be covered?
- What assessment and validation procedures will be followed (quick quality check, content validation, processing validation etc.)?
- How will the data be packaged?
- How (i.e., through which channels) will the data be distributed?
- What type of data will be collected (language, modality)?
- What is the origin of the data?
- Are the data protected by an exclusive right?
  - Who holds it?
  - Do you need additional permission for the intended use of the data? How will you obtain it?
- Under what license will the data be distributed?

ELG will promote open licenses such as the CC licenses for datasets produced with ELG funding. However, the consortium has identified existing resources that may be useful for the completion of the ELG project. Such resources shall be distributed according to their original licenses.

3.3 Storage, Preservation and Access

- Where will the data physically stored?

The ELG platform will be hosting some of the LRs, as such it should be indicated in the DMP. Remote access should be documented.

- What is the (expected) size of the data and its main characteristics?
- Are there storage backup solutions?
- What are the security management procedures with regards to storage?
- How will the sustainability of the resource be guaranteed?
  - Will it incur any additional costs? How will they be covered?
  - What are the foreseeable uses of the data?
  - How will accuracy of personal data be guaranteed?
  - How often will the results of anonymization be reviewed?
- How will data maintenance be guaranteed on short/long term?
- Will the validation reports be shared? If no, why?
- Will the data have a persistent identifier (ISLRN)?

- How often will this Data Management Plan be reviewed?

3.4 Sharing and Re-Use

- When will the Language Resource be made available?
- Will there be any restrictions on access to and re-use the Language Resources (and if so, how are they justified) (e.g., only for research purposes in the field of Language technology)?
- Under which license will the data produced in the project be shared?
    - Do you have all the necessary rights and permissions to share the data under this license?
- Do you need additional permission(s) to lawfully share the Language Resource? How will you obtain it (them)?

4 Checklist for a Data Management Plan for Other Categories of ELG Data[12]

The template below constitutes a preliminary framework for the processing of data other than Language Resources, collected and processed exclusively in the course of the project. It includes essential elements in order for all stakeholders in the ELG to be compliant for a general Data Management policy but as such this template needs to be adapted to each type of data collected and processed and may be updated with the appropriate Terms of Service or Ethics analyses to be delivered during the project.

4.1 Data Summary

- What is the purpose of the data collection/generation and its relation to the objectives of the ELG?
- What types and formats of data will the ELG project generate/collect?
- Will you re-use any existing data and how?
- What is the origin of the data?
- What is the expected size of the data?
- To whom might it be useful ('data utility')?
- Will personal data be processed in the project?
- What types of personal data will be processed?
- Are the personal data necessary to achieve the project's purpose?
- What ground for lawfulness will the processing be based upon (e.g., consent)[13]?
- When and how will the data be anonymised?
- Who will be the data controller[14]?
- Will there be any data processors (persons or entities processing data on behalf of the controller)? Who will they be?
- Is the processing of the personal data necessary and proportional in relation to the project's purpose?
- What are the risks to the rights and freedoms of the persons that the data relate to?
- What measures are envisaged to mitigate the risks and to ensure protection of the personal data (e.g., pseudonymisation, encryption)?

[12] This is a modified version of the H2020 DMP template.
[13] Cf. Art. 6 of the GDPR.
[14] The GDPR defines "data controller" as the person who, alone or jointly with others, defines the means and purposes of data processing.

4.2 Compliance with FAIR Data Principles

Making Data Findable

Are the data produced and/or used in ELG discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g., persistent and unique identifiers such as Digital Object Identifiers)? For example aggregated data from workshop participation can be identified with a DOI no. 123456:

- What naming conventions do you follow?

ELG naming convention will be followed.

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

ELG will index data with appropriate keywords.

- Do you provide clear version numbers?
- What metadata will be created? In case metadata standards do not exist in the relevant discipline, please outline what type of metadata will be created and how.

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 (e.g., they contain personal data), 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?
- 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?

Making Data Interoperable

- Are the data produced in the project interoperable, i.e., allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e., adhering to standard 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?

Increase Data Re-Use

- 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?

4.3 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)?

4.4 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?

4.5 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?

A. Pre-existing Language Resources

		Language Group A			Language Group B			Language Group C			
	Open Access	Corpora	Lexicons	Models	Corpora	Lexicons	Models	Corpora	Lexicons	Models	Totals
META-SHARE	yes	617	447	16	55	54	0	84	51	1	1325
	no	582	550	1	44	65	0	198	94	0	1534
ELRC-SHARE	yes	317	114	0	3	1	0	0	0	0	435
	no	74	16	0	2	1	0	0	0	0	93
ELDA	no	563	1012	0	35	18	0	250	54	0	1932
ELG	mixed	74	108	43	0	0	12	4	1	21	263
Totals		**2227**	**2247**	**60**	**139**	**139**	**12**	**536**	**200**	**22**	**5582**

Figure 1: Table of pre-existing Language Resources identified by the ELG consortium for distribution on the ELG marketplace

B. DMP Template

Page 1

DMP v.x.x
LR n°xx
LR Name

Author:

Organisation:

Date:

Dissemination level:

Template of Data Management Plan for a Language Resource produced within ELG project

Page 2

Table of contents

Abstract (Purpose of LR)

C. DMP Content

What is the initial purpose of LR production?

(e.g.) The initial purpose of the collection of this LR is to have parallel corpus of parliamentary proceedings in English and French

What are the institutions involved in the LR collection?

(e.g.) The Language Resource has been collected by the University of X and annotated by Y.

What language data data will be processed in the project and how?

Please accurately describe the types of data that are collected, processed in the project (speech, text,…).

Will personal data be processed in the project?

Yes/No, If yes answer the questions below

What types of personal data will be processed?

'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or 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.

What ground for lawfulness will the processing be based upon (e.g., consent)?

The grounds for lawfulness of processing of personal data are provided by Article 6 of the GDPR

- Consent
- Necessary for the performance of a contract
- Necessary to comply with a legal obligation
- Protection of subjects' vital interests
- Performance of a task carried out in the public interest or exercise of official authority
- Legitimate interests pursued by the controller or third party

Who will be the data controller?

'Controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law

Will there be any data processors (persons or entities processing data on behalf of the controller)? Who will they be?

'Processor' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;

Is the processing of the personal data necessary and proportional in relation to the project's purpose?

In order to comply with the GDPR, a data controller shall process personal data in a manner that is adequate, relevant and limited to the purpose of the process. For example personal data should not be processed if they are not strictly necessary for the good achievement of the purpose of a project.

What are the risks to the rights and freedoms of the persons that the data relate to?

For example collection of religious, political or health data may put the persons whose data are collectied in danger of being discriminated against.

What measures are envisaged to mitigate the risks and to ensure protection of the personal data (e.g., pseudonymization, encryption)?

Here detail the technical and/or organisational measures put in place to ensure protection of personal data (e.g., secure storage, access limited only to project participants.

Data Acquisition

Production of language ressources

What standards/best practices will be implemented in the production process? How will they ensure interoperability of the data?
What metadata will accompany the data?
Describe accurately which metadata will accompany the data (legal metadata, origin, language with ISO code, etc.) in compliance with ELG metadata scheme to be disclosed later in the course of the ELG project
What are/were the estimated costs of data production? How will/were they be covered?
What assessment and validation procedures will be followed (Quick Quality check, content validation, processing validation)?
How will the data be packaged?
How (i.e., through which channels) will the data be distributed?
Here providers should indicate the use of the ELG marketplace as a distribution channel, other distribution channels can be indicated, if applicable.
What type of data will be collected (language, modality)?
What is the origin of the data?
Are the data protected by an exclusive right?
Who holds it?
Do you need (additional permission) to make the intended use of the data? How will you obtain it?
Under what license will the data be distributed?
ELG promotes the use of open licenses such as CC licenses for datasets produced with ELG funding.
Storage, Preservation and Access
Where will the data physically stored?
What is the (expected) size of the data and its main charcteristics?
Are there storage backup solutions?
What are the security management procedures with regards to storage?
How will the sustainability of the resource be guaranteed?
Will it incur any additional costs? How will they be covered?
What are the foreseeable uses of the data?
How will accuracy of personal data be guaranteed?
How often will the results of anonymization be reviewed?
How will data maintenance be guaranteed on short/long term?
Will the validation reports be shared? If no, why?
Will the data have a persistent identifier (ISLRN)?
How often will this Data Management Plan be reviewed?
The DMP should be revised at the start of the project and whenever major changes occur.
Sharing and Re-Use
When will the Language Resource be made available?
Will there be any restrictions on access to and re-use the Language Resources (and if so, how are they justified) (e.g., only for research purposes in the field of Language technology)?
Under which license will the data produced in the project be shared?
Do you have all the necessary rights and permissions to share the data under this license?
Do you need additional permission(s) to lawfully share the Language Resource? How will you obtain it (them)?