EUROPEAN LANGUAGE GRID

D5.5

Data Management Plan (version 2)

Authors:	Mickaël Rigault (ELDA), Victoria Arranz (ELDA), Khalid Choukri (ELDA), Valérie Mapelli (ELDA), Pawel Kamocki (ELDA), Lucille Blanchard (ELDA)
Dissemination Level:	Public
Date:	31-12-2020

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, 42 months (GA amendment version: AMD-825627-7)
Deliverable number	D5.5
Deliverable title	Data Management Plan (version 2)
Type	Report
Number of pages	24
Status and version	Final – Version 2
Dissemination level	Public
Date of delivery	Contractual: 31-12-2020 – Actual: 26-12-2020
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	Mickaël Rigault (ELDA), Victoria Arranz (ELDA), Khalid Choukri (ELDA), Valérie Mapelli (ELDA), Pawel Kamocki (ELDA), Lucille Blanchard (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 © 2020 ELG Consortium

Table of Contents

List of Abbreviations

API	Application Programming Interface
CMDI	Component Metadata Infrastructure
CNIL	Commission nationale de l'informatique et des libertés (French National Commission on Informatics and Liberty)
DMP	Data Management Plan
DOI	Digital Object Identifier
DC	Data Controller
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
ELG	European Language Grid
ELDA	Evaluations and Language resource Distribution Angency
ELRA	European Language Resources Association
EU	European Union
FAIR	FAIR Principles (Findability, Accessibility, Interoperability, Reusability)
GDPR	General Data Protection Regulation
HLT	Human Language Technology
ICO	Information Commissioner's Office
IPR	Intellectual Property Rights
ISLRN	International Standard Language Resource Number
k8s	Kubernetes
LR(s)	Language Resource(s)
NCCs	National Competence Centres
NLP	Natural Language Processing
PID	Persistent Identifier
S3	Simple Storage Server

Abstract

This document introduces the updated version of the ELG Data Management Plan. It describes the recommendations and policies for the management of all the data types processed within the ELG platform, ranging from personal data to language data, technologies and services, with regard to technical, organisational and legal aspects. It provides the recommendations and policies that ensure compliance with both FAIR and GDPR principles and how these have been implemented in the European Language Grid (ELG). In this regard, the DMP is both a descriptive and a prescriptive document, describing the ELG platform content and guiding all users involved to manage such content.

1 Introduction

1.1 Objectives of the European Language Grid

The European Language Grid (ELG) project aims to address the fragmentation of the European Language Technology business and research landscape by establishing the ELG as the primary platform for Language Technology (LT) in Europe and to strengthen European LT business with regard to the competition from other continents. The ELG is a platform for commercial and non-commercial Language Technologies, both functional (running services and tools) and non-functional (data sets, resources, models). In order to achieve this, the consortium in charge of the ELG platform has enacted several priorities that include the processing of massive amounts of data from various partners and of different types. Such data intensive project requires clear data management policies, in particular considering the current GDPR constraints.

The Data Management Plan (DMP) that is detailed in this document is thus a concrete necessity for organisational, technical and legal management of all the data types that will be processed in the course of the project.

1.2 Objectives of the Data Management Plan for ELG

The objectives of the ELG DMP are:

- To document the variety of data types processed in the course of the project and report on how the data are going to be managed with regard to technical, organisational and legal aspects.
- To comply with today's best practices and, in particular, with H2020 requirements as well as GDPR obligations. This DMP defines useful practices to enhance compatibility with the FAIR principles[1], as endorsed and specified for H2020. These fall into four major categories:
  - Findability
  - Accessibility
  - Interoperability
  - Reusability

In addition, this DMP is also used as procedure description to prove compliance with the General Data Protection Regulation (GDPR[2]) in application of the accountability principle[3].

Moreover, the DMP also provides advice in terms of best practices for language resource creation.

Therefore, the objective of the DMP is twofold:

1. To document how the various types of data collected, received and/or processed during the course of the project are going to be managed in compliance with the external regulations on research activities and protection of individuals' rights.
2. To provide policies and best practices towards the appropriate production of language resources in compliance with both FAIR principles and the GDPR.

---

[1] https://www.go-fair.org/fair-principles/
[2] Regulation (EU) 2016/679
[3] Article 5-2 of the GDPR

2 Typology of data to be collected, supplied or processed within ELG

There are three major categories of data types that are being collected and/or processed during the course of the project:

1. Language data, technologies and services.
2. Data items containing personal data.
3. Feedback data.

The following section will describe each type briefly together with the major policy issues that will be further elaborated upon later on in this DMP.

2.1 Data items containing personal data

One of the main actions that is foreseen by the ELG is the development of a public grid connecting resources and tools and providing easy access to language tools and services to the public as well as to both research and industry stakeholders. In order to assist the development of this grid, the formation of a community that will use and interact with tools and services through this grid is foreseen as indispensable, too.

This community is being built through the collection of data related to users and stakeholders, including personal data. Their interactions with the platform are based on the provision of personal data to the platform for actions such as registration, access or purchase of language resources through the platform. As a result, all operations that require the processing of data items that are considered *personal data* (cf. Section 3.1) under the definition enacted by the GDPR[4] should be processed in compliance with the principles enforced by the laws and regulations.

The GDPR principles that will be detailed throughout the DMP for the different data types that are processed during the project are listed below. There are four major requirements that determine how each data type has to be managed in compliance with the GDPR:

1. **Lawfulness of processing**
   In order to comply with this principle, the following should be clearly indicated to the users:
   - the specific purposes of data processing,
   - the technical process behind that processing, and
   - the legal basis allowing the collection and processing of the personal information.
2. **Purpose limitation principle**
   The purpose limitation principle falls under a similar logic to that of the lawfulness principle as the purposes that are indicated to the users shall be the only ones for which the data can be processed and it prevents processing for any other purpose.
3. **Data minimization principle and storage limitation**
   These two principles have been bundled together as they both ensure that data processing operations are strictly limited to what is strictly necessary for the advancement of the project. The storage limitation principle, on a more specific issue, clearly limits the duration for which the data can be stored.
4. **Data accuracy, data integrity and data confidentiality**

[4] Article 4-1 of the GDPR

These principles ensure that all data processing operations of personal data protect the quality of data and of its processing, and ensure the accuracy of datasets as well as offer protection against security risks such as material destruction or unauthorised access.

ELG uses registration protocols to give users access to its assets (resources, tools, directories, etc.) and uses Cookies for those who browse through the platform. The platform also collects data on the contributors to the platform as this cannot be done anonymously (providers may supply copyrighted resources and hence have to be identifiable). So, our legal basis is hence clearly stated and is documented in the following:

- ELG's Privacy Policy[5]
- ELG's Terms of Use[6]
- The current DMP (see Section 3).

2.2 Language data, technologies and services

The project is producing a comprehensive catalogue of language data as well as language technologies and services. It is also making available a catalogue of projects, initiatives and stakeholders in relation to the Human Language Technology Sector.

To populate this catalogue, the ELG consortium collects and makes available data from several sources:

- Metadata records for resources and technologies harvested from authorised external repositories. The content files remain at the host repository and can be accessed only via a link included in the metadata record.
- External resources and technologies migrated and ingested into the platform (resources hosted in ELG).
- Language resources produced by the selected projects of the two ELG Open Calls.
- Metadata records for language resources and technologies registered in the platform by third parties, for which content files remain at the host repository and can be accessed only via a link included in the metadata record.
- Language resources and technologies registered in the platform by third parties and fully ingested (resources and technologies hosted in ELG).

The ELG DMP describes the procedures and practices to manage all types of data related to these assets.

In addition, ELG also produces technologies for the ELG platform development (cataloguing, data and metadata management, such as metadata editors, metadata converters, etc.). These will be shared with the HLT community as open-source code, but no DMP is required for them.

2.3 Feedback data

Finally, to fine-tune the operation of the platform there are also feedback data that may be collected through different means: from the user base via surveys, through the contact page or in community events. The purpose of these feedback data will be to create statistical data for the improvement of the aforementioned platform. These feedback data will be considered together with the personal data from now forth in this document (see Section 3.1).

[5] https://www.european-language-grid.eu/privacy-policy/
[6] https://live.european-language-grid.eu/page/terms-of-use

3 Data management policies for personal data

3.1 Typology of personal data collected across the projects

There are common features among the different items of personal data that may be processed during the course of the ELG. ELG provides thorough documentation about all the data types that fall within the definition of personal data as indicated in the GDPR:

Personal data provided by users through access and interaction with the project will undergo the following actions, as reported in the ELG Privacy Policy:

- Email addresses and other information processed through the contact forms: they are saved only for the purpose of responding to the message.
- Email addresses and other information collected through correspondence via email: the user's e-mail address and other personal contact data are used for the correspondence with the user. Due to legal obligation every e-mail correspondence is archived. Users have the right to access and modify their personal data (see further down).
- Email addresses processed through subscription forms (e.g., Newsletter): they are used for the purpose of receiving the subscribed service. These and other associated personal contact data are stored in our contact database, as for the previous point, but not shared with anyone outside the ELG Consortium. Only Consortium members and the ELG boards have access to such information for reasons of project running.
- Registration information for attendance to community events: they are used for the purpose they were collected (e.g., registering to attend an event) and are not shared with people outside the ELG Consortium. Only Consortium members and the ELG boards have access to such information for reasons of project running.
- Survey data and feedback data: they are not anonymous as the ELG needs to have people's names to know who wishes to contribute to or has questions regarding ELG. They are not shared with people outside the ELG Consortium. Only Consortium members and the ELG boards have access to such information for reasons of project running.
- Information on organisations and projects provided by the NCCs for the country profile: this is public information that can be published through the ELG Catalogue without any specific restriction.
- Email addresses for contact people provided by the NCCs: this is restricted information that is used by ELG to inform contact people about the metadata for the organisations (existence and request for enrichment). These email addresses are not made public.
- User/Stakeholder account information stored in the platform (e.g., logging information): this information is used for:
    - the purpose of using the platform (acquiring or providing some functional or non-functional services) and is stored in our contact database, as for the points above, and also
    - for the purpose of statistics and sending out notification messages whenever required (e.g., to inform providers about the publication of the items they have submitted).
- Billing information:
    - Special certification will be allowed to handle payment data. Using the services of a certified external provider to handle such critical data will be considered.
    - This information should be saved only during the duration of the financial transaction.

    - Since procedures for the processing of billing information are still under discussion the storage of payment details (acquisition of a data set or usage of some service) is still being considered and will be defined in the final version of the DMP due at the end of the project.
- Access data: on every access to the ELG website some usage, transmission and connection data will be collected, temporarily stored in a log file and regularly deleted after 90 days. These are listed in detail in the ELG Privacy Policy and in Section 3.2.2 further down.
- User statistics processed through automated means can be summarised as follows (with full details in the ELG Privacy Policy and in Section 3.2.2):
    - ELG's website uses Google Analytics.
    - Google will use this information on behalf of the operator of the ELG website for the purpose of evaluating the user's use of the website.
    - The information generated by the cookies about the user's use of the ELG website is transmitted to and stored by Google on servers in the United States. However, IP -anonymisation is activated on the ELG website.

Users will have the right to access their personal data and, if so wished, rectify, complete, delete, restrict or object to processing. In order to do so, they are invited to contact ELG's DPO (see Section 3.2.2).

Personal Data included in Language Resources and Language Technologies:

- Personal information related to providers of language data, technologies and services: this information is stored partly in a) the person's user account and b) in the metadata record of the resource, tool or service (s)he is providing through the ELG platform. The use made of this type of personal data is similar to that for the "User/Stakeholder account information stored in the platform" mentioned earlier in this section.
- Language data or technology metadata: metadata are used for the purpose of catalogue content description, storage and retrieval. Metadata records are made public once published by the metadata editors. These metadata may contain information on the data and technology providers and creators. Such information is public information available to catalogue users through the (functional or non-functional) service description page/entry. Such information will be accessible for as long as the service is available in the catalogue.

Personal information processed through the ELG Open Calls:

All personal information from evaluators and applicants participating in the ELG Open Calls for projects:

a. is stored on the ELG Open Call Platform, which is only available to the call organising institution (Charles University – CUNI);
b. some limited information needed for evaluation purposes (name, country, gender, topics of expertise, as well as CV and pdf of submitted project proposals) is also available to the ELG Pilot Board;
c. the data are processed on encrypted storage.
d. No personal information from either project evaluators or applicants is made public by ELG to the outside.
e. These data will be kept for five years after payment of the balance (in accordance with ELG's GA Article 18).

The purpose for keeping these data is specified in the registration forms both evaluators and applicants fill-in and comply with:

- Project evaluators:

    a. Quoted from the registration form for evaluators where evaluators confirm their consent:
    "*I hereby grant a permission to Charles University, residing in Ovocný trh 560/5, 116 36 Prague 1, company: VAT Number: CZ00216208 (hereinafter "CU"), acting as a controller of personal data of all faculties and other parts of CU, to process my following personal data: name, surname, title/s, nationality, e-mail, phone and professional experience for the purpose of selecting members of expert panel evaluating pilot projects within the EU project European Language Grid No. 825627.*"

- Project applicants:

    b. Quoted from the submission form for project proposals where applicants confirm their consent:
    "We hereby grant a permission to Charles University, residing in Ovocný trh 560/5, 116 36 Prague 1, company: VAT Number: CZ00216208 (hereinafter "CUNI"), acting as a controller of personal data of all faculties and other parts of CUNI, to process the personal data filled in this submission form for the purpose of selecting pilot projects and providing the financial support for the selected projects within the EU project European Language Grid No. 825627."

3.2 Compliance with GDPR

Pursuant art. 5 (1) e) of the GDPR, personal data is only 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 is 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). All these aspects of personal data are further analysed in the sections below.

3.2.1 Data Privacy Impact Assessment

The first step to document the compliance process that is in place by the project consortium is to measure how the project and the data processing that are foreseen may impact privacy. This is mandatory under the GDPR in the event that these processes would cause high risks on the rights and liberties of the persons. Thus:

1. This impact analysis is a declarative procedure that should be conducted by projects potentially collecting massive amounts of personal data.
2. This is a required step in the event of an audit, but does not need to be linked to the sharing or distribution of a resource.

There are several methods to perform this analysis. Below we present two which are easy to follow and recommended. Either of them can be adopted by the Consortium to ensure that any needed corrective measures are taken immediately:

- The *French National Commission on Informatics and Liberty* (CNIL) provides **software**[7] that helps conduct this assessment. Identification of similar tools at the EU and the Member State levels will be carried out if and as needed.
- The UK's Information Commissioner's Office (ICO) provides a **form**[8] to carry out a Data protection impact assessment.

[7] https://www.cnil.fr/fr/outil-pia-telechargez-et-installez-le-logiciel-de-la-cnil
[8] https://gdpr.eu/wp-content/uploads/2019/03/dpia-template-v1.pdf

Should ELG incur into either (1) or (2), the project's coordinating team will liaise with ELG's legal team to carry out the assessment.

3.2.2 Implementation of a robust Privacy Policy

The main policies that document the compliance with the GDPR are the ELG Privacy Policy[9] and the Terms of Use policy[10] that focus on data that is being processed with direct input from the users or participants in the project. As such, this policy document enables the consortium to provide the necessary information to the user about the ways such data is being processed and used within the project.

The ELG Privacy Policy provides the following elements to ensure compliance with GDPR rules[11]:

- Indication of Data controller's identity and contact details:
  - Phone: +49 (0)30/23895-1833
  - E-mail: georg.rehm@dfki.de
- Indication of Data protection officer's identity and contact details:
  - Phone: +49 (0)631 / 205 75-0
  - E-mail: datenschutz@dfki.de
- Indication of the purpose of data processing:
  - Provision of the information offering in the course of the public communication of the DFKI (on behalf of ELG).
  - Establishment of contact and correspondence with visitors and users.
- Indication of legal basis under which the data is going to be processed:
  - *Anonymous and protected usage*: Visit and usage of the ELG website are anonymous. At our website personal data are only collected to the technically necessary extent. The processed data will not be transmitted to any third parties or otherwise disclosed, except on the basis of concrete lawful obligations. Within our information offering we do not embed information or service offerings of third-party providers.
  - While using our website the data transmission in the internet is being protected by a generally accepted secure encryption procedure and hence cannot easily be eavesdropped or tampered.
  - *Access data*: On every access to our website some usage, transmission and connection data will be collected, temporarily stored in a log file and regularly deleted after 90 days.
  - On every access/retrieval the following data are stored:
    - IP address
    - transmitted user agent information (in particular type/version of web browser, operating system etc.)
    - transmitted referrer information (URL of the referring page)
    - date and time of the access/retrieval
    - transmitted access method/function
    - transmitted input values (search terms etc.)
    - retrieved page respective file
    - transmitted amount of data

[9] https://www.european-language-grid.eu/privacy-policy/
[10] https://live.european-language-grid.eu/page/terms-of-use
[11] This is an excerpt. The full details are in the Privacy Policy document: https://www.european-language-grid.eu/privacy-policy/

- status of processing the access/retrieval

The processing of the access data is lawful because it is necessary for the purposes of the legitimate interests pursued by the ELG Consortium, represented by DFKI. The legitimate interests pursued by DFKI are the adaptation and optimisation of the information offering and the investigation, detection and prosecution of illegal activities in connection with the usage of the ELG website.

The stored data records can be statistically evaluated in order to adapt and optimize the website to the needs of visitors. No technique that offers the possibility to retrace the access characteristics of users (tracking) is applied. The creation of user profiles and automated decision-making based on it is precluded.

The stored data records are not attributable to specific persons. They are not being combined with other data sources. However, the stored data can be analysed and combined with other data sources, if we become aware of concrete indications of any illegal usage.

- What personal data we collect and why we collect it:
  - Contact Forms:
    - Personal information sent through the contact form, in particular e-mail addresses, will be saved for only the purpose of responding to the message.
  - Cookies:
    - We use so-called cookies on the ELG website. Cookies are small files that are being stored by the visitor's web browser. The cookies used on the ELG website do not harm the user's computer and do not contain any malicious software. They offer a user-friendly and effective usage of the website. We do not use cookies for marketing purposes.
    - We transmit so-called session cookies to the visitor's web browser. They are valid only for the duration of his/her visit to the ELG website and they do not have any meaning outside of the web site. The session cookies are needed in order to identify the user's session with a unique number during his/her visit and to transmit our contents in the preferred language. At the end of the visit the session cookies are automatically deleted upon termination of the web browser.
    - We also transmit permanent cookies to the visitor's web browser with a validity period of at most 365 days. We are exclusively using these cookies in order to respect the user's settings for the type of presentation (normal, inverted) and for the font size. Furthermore, it will be recorded whether the visitor has taken notice of the information about the usage of cookies in his/her web browser.
    - The visitor can adjust his/her web browser such that (s)he will be informed on setting cookies and allow cookies on an individual basis resp. exclude the acceptance of cookies for specific cases or generally. The user can also adjust the automatic deletion of cookies upon termination of the web browser. Upon deactivation of cookies the functionality of the ELG web site can be limited. In any case, the ELG's information offering is available to its full extent.
  - Embedded content from other websites:

- Articles on this site may include embedded content (e.g., videos, images, articles, etc.). Embedded content from other websites behaves in the exact same way as if the visitor has visited the other websites.
- These websites may collect data about the user, use cookies, embed additional third-party tracking, and monitor his/her interaction with that embedded content, including tracking his/her interaction with the embedded content if (s)he has an account and is logged in to that website.

- Analytics[12]:
  - ELG's website uses Google Analytics, a web analytics service provided by Google, Inc. ("Google"). Google Analytics uses "cookies" to help the website analyse how users use the site. The information generated by the cookie about the use of the website will be transmitted to and stored by Google on servers in the United States.
  - IP anonymisation is activated on this website. ELG's user's IP address will be truncated within the area of Member States of the European Union or other parties to the Agreement on the European Economic Area. Only in exceptional cases the whole IP address will be first transferred to a Google server in the USA and truncated there.
  - Google will use this information on behalf of the operator of this website for the purpose of evaluating the user's use of the website, compiling reports on website activity for website operators and providing them other services relating to website activity and internet usage.
  - The IP -address, that the user's Browser conveys within the scope of Google Analytics, will not be associated with any other data held by Google. The user may refuse the use of cookies by selecting the appropriate settings on his/her browser, however it should be noted that if this is done the user may not be able to use the full functionality of this website. The visitor can also opt-out from being tracked by Google Analytics with effect for the future by downloading and installing Google Analytics Opt-out Browser Addon for the user's current web browser: http://tools.google.com/dlpage/gaoptout?hl=en.
  - As an alternative to the browser Addon or within browsers on mobile devices, the user can click on the link Disable Google Analytics in order to opt-out from being tracked by Google Analytics within this website in the future (the opt-out applies only for the browser in which the user sets it and within this domain). An opt-out cookie will be stored on the user's device, which means that (s)he will have to click the link again, if (s)he deletes his/her cookies.
- Correspondence: The user has the option to contact the ELG team by e-mail[13]. The latter will use the former's e-mail address and other personal contact data for the correspondence with him/her. Due to legal obligation every e-mail correspondence will be archived. Subject to our legitimate interests the user's e-mail address and other personal contact data can be stored in our contact database. In this case the user will receive corresponding information on the processing of his/her contact data.
- Information regarding the users' rights regarding their data:

[12] This has been introduced in Section 3.1.
[13] https://www.european-language-grid.eu/contact/

- Besides the information in this data protection policy the user has the right to access his/her personal data. To ensure fair data processing, (s)he has the following rights:
  - The right to rectification and completion of his/her personal data
  - The right to erasure of his/her personal data
  - The right to restriction of the processing of his/her personal data
  - The right to object to the processing of his/her personal data on grounds related to his/her particular situation
  - To exercise these rights, the user is invited to contact ELG's data protection officer:
    - Phone: +49 (0)631 / 205 75-0
    - E-mail: datenschutz@dfki.de
  - Furthermore, the user has the right to lodge a complaint with a supervisory authority if (s)he considers that the processing of his/her personal data infringes statutory data protection regulations.

3.2.3 Limitation of access

As a general policy, limitation of access is a valid security measure that is implemented by ELG. In all operations related to personal data processing, there is a severe restriction of access and a strict selection of persons that are authorised to have access to personal data. These are members of the DFKI team.

There are also other technical measures put in place regarding this access forbidding the persons that may have access to personal data to exchange personal data outside the authorised personnel.

3.2.4 Anonymisation of personal data in language resources, technologies and services

Language resources, either harvested from external sources or produced within the ELG Open Calls, may contain personal data (e.g., Named Entities). Therefore, there is a specific procedure that fosters protection and distribution of those resources, if there are reasons to share these with outsiders. This is explained below.

There are three major concerns that need to be taken into consideration for the management of language data containing personal information:

1. The first concern regards resources harvested by the ELG Consortium from external sources. The consortium makes sure that these datasets have been processed according to GDPR rules and that all information required by the GDPR is distributed along with the resources (such as consent forms if necessary, or other compliance documents). In the event that this is not possible, the ELG consortium may run anonymisation or pseudonymisation operations on the datasets that are concerned in order to ensure compliance with the GDPR, especially for redistribution.
2. The second concern is about the distribution of resources to players located in countries that may not protect personal data at the same level as European Union countries do (the GDPR forbids transfer of personal data collected in compliance with GDPR outside a list of countries[14]). These transfers should be strictly limited by technical measures preventing unauthorised users from accessing the resources or legal measures requiring that users from third party countries perform satisfying data protection

[14] These are countries where appropriate safeguards are in place and EU considers that data protection regulations are appropriate and equivalent to the EU GDPR (https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en).

operations as required by GDPR. This is being analysed by the Consortium at the time of writing of this document.

3. Regarding language data or tools provided by external providers, the ELG Consortium can advise on pseudonymisation techniques and tools, but the GDPR compliance of the resource remains the sole responsibility of the provider.

3.3 Hosting and storage of personal data items

The issues of storage and hosting are clearly linked with security issues but here we focus on the structure that is being deployed for the proper functioning of the Grid.

A few key features are that:

- The storage facilities are protected by robust encryption (see Sections 3.1 and 3.2.2) to prevent external breaches.
- The hardware infrastructure is accessible to authorised persons only (Sections 4.4.2 and 4.4.4).
- There are regular backups of the whole infrastructure and the backups are stored according to rules established by the consortium:
  - Daily backups with 7 days retention are carried out for WordPress pages and the stagging website.
  - The stagging site mirrors the whole content of the european-language-grid.eu instance.
  - Content is manually transferred from stagging to live and vice-versa so that they are completely separate from each other.

4 Data management policies for language data, technologies and services

4.1 Typology of language data processed by the ELG

As already mentioned, the ELG platform collects and produces a wide range of language resources, technologies and services that are made available to the community at large. These can be categorised as follows:

- Language data, technologies and services provided by infrastructures operated by the Consortium partners (e.g., LINDAT/CLARIAH-CZ[15]);
- Language data, technologies and services identified from external providers (e.g., Zenodo[16]);
- Language data, technologies and services produced through the ELG Open calls;
- Language data, technologies and services produced by third parties and integrated into the platform;
- Language data produced as output from the ELG tools and services (data processed within ELG);
- Language metadata produced along with language data;
- Technologies produced for platform development.

The privacy issues related to the storage, access and distribution of language data were described in Section 3 together with those regarding personal data. The following sections focus on the specific data management issues related to the treatment of the language data by their creators and in agreement with ELG.

[15] https://lindat.cz
[16] https://zenodo.org

4.2 Compatibility with the ELG metadata scheme

The ELG metadata scheme[17] plays a key role in the infrastructure as it helps ensure compliance with FAIR principles such as Findability (see Section 4.4.1). The language data, technologies and services that are being produced during the Open Calls, or those provided either by Consortium members or by third parties under bilateral agreements, will comply with the ELG metadata by design or through a conversion procedure.

The major issue behind this regards data, technologies and services harvested from external producers and providers. To address this issue:

- The ELG consortium has defined a set of minimal metadata elements[18] that are mandatory in order to be compatible with the ELG metadata scheme and thus ensure proper compatibility and consistency with the ELG catalogue.
- When looking at repositories to be harvested, ELG partners have developed and will develop appropriate converters[19] to ingest these resources with the right metadata descriptions. This has been the case towards the ingestion of several repositories such as ELRA[20] and LINDAT-CLARIAH-CZ reported in Appendix A.

4.3 Metadata licensing and sharing

Metadata provide essential information to describe resources, technologies, and services and constitute a special data type. The Consortium has agreed to release the public part of the metadata descriptors under a very permissive license (Creative Commons). The metadata can be harvested by any interested party.

4.4 Making ELG data compatible as FAIR data

Due to the specific nature and wide variety of the ELG content, the application of FAIR principles has been duly studied to put in place the best practices to ensure compatibility between those principles and the GDPR.

4.4.1 Findability

Findability is usually measured by 4 criteria as defined by the following FAIR principles:

1. F1. (Meta)data are assigned a globally unique and persistent identifier. This is a key element of description and identification. ELG uses PiDs as defined by the major data centres. The ISLRN[21] is the PiD used for metadata description. When other PiDs are used in the descriptions of harvested language data, these are also stored as reference (e.g., DOI in the Zenodo[22] datasets).
2. F2. Data are described with rich metadata: the ELG metadata scheme for LR, LT, and stakeholders (for the directories) is complete and metadata have been normalised to cover all the targeted data.
3. F3. Metadata clearly and explicitly include the identifier of the data they describe: this refers back to the PiD established in F1. The dataset's PiD (ISLRN) is explicitly included in its metadata description.

---

[17] https://european-language-grid.readthedocs.io/en/release1.1.0/all/A1_Metadata/Metadata.html
[18] https://european-language-grid.readthedocs.io/en/release1.1.0/all/A1_Metadata/Metadata.html#minimal-version
[19] For full details on the converters developed for the ELG release 1, please refer to ELG Deliverable D5.1: Victoria Arranz, Kahlid Choukri, Valérie Mapelli, Cristian Martinez (ELDA), Penny Labropoulou, Miltos Deliglianis, Stelios Piperidis (ILSP): Identification and collection of existing datasets, models, identified gaps and plans (version 1).
[20] http://catalogue.elra.info/en-us/
[21] http://www.islrn.org
[22] https://zenodo.org

4. F4. (Meta)data are registered or indexed in a searchable resource: This index is part of the ELG cataloguing process and complies with the principles of privacy protection:
    - Access to some parts is strictly limited to certain persons within the consortium (see Section 3.2.3) , this is specifically the case for providers' email addresses, which are not viewable by outsiders through the ELG Catalogue.
    - Further, when metadata are exported, emails are not be exposed unless the ELG has people's explicit consent and indication to do so.
    - However, first names and family names from authors/creators are exposed as best practice for bibliographic referencing and cataloguing following research practices and principles.

4.4.2 Accessibility

A specific access process has been implemented to ensure that only authorised users have access to the various types of ELG data:

a) User data derived from all personal information provided by users when they use the websites and services provided by ELG.
b) Personal data comprised in Language resources, tools, and services, and
c) Language data, technology and services themselves.

The ELG Terms of Use[23] (also described in Section 3.2.2) provide users with clear directives on how to access and use the above data in accordance with rules and regulations.

4.4.3 Interoperability

To be interoperable, the ELG platform has developed a metadata schema that is built on existing ones (META-SHARE[24], CMDI[25], etc.). ELG also advocates 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 is the use of containerisation techniques to make the data interoperable.

Interoperability is ensured in two ways:

- For personal data that must be interoperable: a simple exchange format (CSV or XML) will be used.
- For other types of data: largely adopted formats are used for speech/Audio, text, multimedia resources, as described in the ELG Platform report26.

4.4.4 Reusability

To allow reusability the ELG platform enforces the use of its metadata elements. Through this DMP associated to each Language Resource type, it also ensures that the purpose for which the data have been produced is documented with indication on the possibilities to reuse the data for other purposes (other technology developments). The ELG platform also ensures that all IPR issues are cleared.

---

[23] https://live.european-language-grid.eu/page/terms-of-use
[24] http://www.meta-share.org/knowledgebase/overviewOfTheMetadataModel
[25] https://www.clarin.eu/content/component-metadata
[26] Penny Labropoulou, Dimitris Galanis, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Stelios Piperidis (ILSP), Florian Kintzel, Nils Feldhus, Georg Rehm (DFKI), Ian Roberts, Kalina Bontcheva (USFD), Andis Lagzdiņš, Jūlija Meļņika (TILDE) : ELG Deliverable D2.4: ELG Platform (first release).

In addition, ELG ensures that the Language Resources are released with a clear and understandable data usage license. In particular, the ELG consortium advocates for very permissive licenses whenever this is possible. Special attention is paid to other legal issues that may hinder the reusability of the resources (e.g., GDPR compliance) and legal support is provided as part of the ELG helpdesk whenever required.

Reuse of personal data collected by the consortium through the ELG website, open calls for projects and other community events is strictly limited to the authorised persons within the ELG Consortium. These authorised people will have the right to perform strictly listed operations (website maintenance, traffic statistics, contacts with stakeholders registered in the platform, etc.). It is not foreseen to make available these data to third parties. Nevertheless, reusability is ensured within the consortium for the duration of the project and as long as necessary for the operations of the ELG. Sections 3.1 and 3.2 describe the authorisation protocols and personnel, reuse conditions and restrictions for such personal data.

Regarding language data containing personal data, these are distributed according to two procedures:

- For resources harvested from external sources or hosted in the platform:
  - The ELG platform simply replicates the license information used by the producer of the data and applies anonymisation only if necessary to ensure distribution of resources harvested by the ELG Consortium itself.
  - For those provided by external providers, it is the sole responsibility of the provider to ensure GDPR compliance (cf. Section 3.2.4 for full details).
- For resources produced during the project (directly or through the Open Calls):
  - These are compliant with GDPR restrictions and ensure that no sensitive data is contained (cf. Section 3.2). Should any anonymisation procedure be required, this is the responsibility of the data producer in order to allow sharing (cf. Section 3.2.4).
  - Regarding licensing of those resources, the Consortium is currently working on a licensing schema that complies with practices promoted by H2020 and Open Data policies. This will be available shortly, to be reported in the final version of the DMP.

4.5 Security measures

4.5.1 Limitation of access

Regarding access to the ELG platform and contents, access limitation is implemented as follows:

1. Access to the platform infrastructure:
   a. This provides access to both Kubernetes development and production clusters.
   b. The ELG administration team has access to potentially all information stored in the platform.
   c. Access to this is granted via kubeconfig (from Kubernetes[27]) credentials.
   d. The tokens are limited in scope (not everyone needs to have access to the whole cluster). Full administration access is granted only to the DFKI team.
   e. Critical component here:
      i. Postgresql database
      ii. Access to it is secured via username/password login.

[27] https://kubernetes.io

iii. Technical root access to the platform infrastructure also provides access to the database.
iv. Only the development teams of the database (DFKI, ILSP) have access to that area.
v. Any information accessed this way is only intended for operational purposes, i.e., debugging.

2. Access to the postgresql database via the catalogue-UI:
    a. This is the usual way users are interacting with the platform: normal logging into the platform.
    b. Access to the information contained is restricted via OAuth tokens and the configured roles for each user.
3. Access to datasets:
    a. This is restricted based on licensing conditions.
    b. Users must be registered to the ELG platform in order to access datasets, tools or services.
    c. This is managed by the backend catalogue system and is crucial.
4. Access to metadata (viewing):
    a. This is allowed for all if metadata records are published.
    b. As regards personal data, the ELG metadata only displays first names and surnames (agreed upon contact information). No email addresses are displayed.
5. Access to metadata (editing):
    a. In full: this is restricted to the metadata curators.
    b. Under limited authorisation: if you are authorised, you can view/edit the records you are permitted to. So, service providers are able to edit their own records, but not those of others.

4.5.2 Infrastructural and technical measures

In addition to the limitation of access, the ELG platform development team has explored infrastructural and technical measures to enforce the security of the data and prevent external breaches. As the project considers these data sets as critical assets, technical measures have been put in operation to ensure both integrity and low risk of public exposure. These are the following:

- Kubernetes “secrets” mechanism requires credentials to access the repositories.
- Service API publishing: APIs exposed by LT services will not be published directly on the internet so as to ensure security.
- Keycloak[28] is the access management solution which is used for user management, authentication and authorisation: user roles are granted and access tokens provided.
- For each ELG service, the appropriate code is inserted to the respective k8s (Kubernetes) config files that specify the mapping between a publicly accessible URL/endpoint to the respective internal k8s backend service name (and port).
- Access control and security for the storage of files (data and tools) is done through an S3-compatible Object Storage solution. This is organised in “buckets” with different access policies.
- S3 credentials have been improved to increase security: multiple accesses are implemented for different components instead of a single access to the whole S3.
- User’s permissions are checked for the upload, download and storage of non-functional content in ELG. Only registered users are authorised to carry out these operations.

[28] https://www.keycloak.org

- For security and isolation reasons, LT service providers can ask for a namespace with restricted access for their services.
- Appropriate technical safeguards are being planned to ensure the rightful access to LRTs through a secure licensing and billing feature. This specific measure will be detailed in the final version of the DMP due at the end of the project.

These measures are handled by the Technical team in charge of the ELG operations that set an appropriate security assessment team. Such team will regularly review this policy and report to the ELG management any newly identified threats or vulnerabilities of the platform components with the appropriate measures to address them. These will be documented in the Platform development reports.

4.6 Specifications for the new language data, technologies and services

A set of specifications will be provided to the pilot projects that will be funded following the ELG Open calls to ensure that the new resources and services produced comply with the ELG requirements as elaborated in this DMP. These specifications will be also recommended to any other data producer targeting compliance with the ELG platform.

As part of its mission, ELG also promotes best practices that emerge from the community for the development of new resources and services to improve their interoperability. These best practices are summarised in the two sections below.

4.6.1 Best practices in language resource production

The ELG platform takes on board new LRs developed within the project or uploaded by potential providers planning to use the platform. For that purpose, the consortium advocates the use of ELG policies with regards to all aspects of the DMP. Each phase (specification, production, post-production) should 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, production, etc.
- the metadata should be fully compliant with ELG metadata
- all mandatory metadata fields should be filled
- the production standards and procedures should be listed
- the quality assessment
- the validation procedures
- the IPR issues (for data access and re-use etc.)
- the management of ethics, privacy, confidentiality (data anonymisation, 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

The **post-production** phase comprises:

- the allocation of a unique identifier (ISLRN)
- 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

Among all the best practices recommended, some of them are more critical than others. For instance, not many initiatives are able to afford the bug reporting in the post-production phase. However, these best practices are to be implemented as recommendations and taken into account whenever possible.

One important aspect to bear in mind is that data creators should also specify 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. 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 accordingly.

Last but not least, the purpose for which personal data is processed within the project needs to be clearly stated (as seen in Section 3). The data controller has to be clearly identified in the DMP (as seen in Section 3.2.2). 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.

4.6.2 Sustainability of language resources

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 customisation, 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 ensure that the LRs in question are properly taken care of.

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

Regarding LR **sharing**, this requires interoperability and access facilities which are managed by the use of ELG containers. Data formatting practices are also promoted to enhance data sharing.

Today, a large set of licenses exists for data sharing. Licenses range from the LR-dedicated licenses by ELDA[29], and by META-SHARE[30] to the more permissive public license suites such as Creative Commons[31]. The licensing scheme established by ELG is currently being defined by the Consortium and will be made public shortly. What can already be said is that priority is given to a sharing as open and unrestrictive as the resources allow.

In that regard, legal support can be provided whenever necessary for more complex situations. With the support of legal experts from the ELDA team, ELG can 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?
- 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 rights-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 offers assistance on legal issues related to the management of data mainly on IPR and personal data issues[32] and provides access to legal advisors. These advisors are also available to provide assistance with the DMP.

4.7 Policy issues regarding language data, technology and services hosted in the ELG

This section looks at “hosting” and its regulation by the eCommerce Directive. The implications behind hosting language data, technology and services are being analysed and this sets the policies being defined.

From a legal point of view, the storage of information provided by a third party may be considered as “hosting” if ELG does not add any value (on the content, e.g., editing/describing/selection of parts, etc.). This “hosting” is regulated by the eCommerce Directive[33]. Under these specific regulations, some information needs to be given to the visitors such as:

- Identity of the service provider;
- Geo-localisation of the service provider;
- Contact information of the service provider;
- Registration information of the service provider.

Some of this information needs to be supplied by providers as part of the GDPR and the “hosting” requirements and additional information is requested by ELG to ensure compliance with its practices.

The “hosting” regulation grants protection to ELG regarding liability under the eCommerce Directive as long as the ELG does not have actual knowledge of any illegal activity of the provider. Therefore, the ELG Terms of Use

[29] http://www.elra.info/Licensing.html
[30] http://www.meta-net.eu/meta-share/licenses
[31] http://creativecommons.org
[32] http://www.elda.fr/en/services-around-lrs/legal-support-helpdesk
[33] https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A32000L0031

clearly indicate that the data providers for language datasets, technologies, and services need to certify that these are compliant with all regulations, i.e., Intellectual Property and privacy ones.

A "notice and take down" procedure has been implemented in order to remove or disable the access upon notification by the rights-holder or upon obtaining actual knowledge of an illegal act regarding the resources.

5 Conclusions

ELG's data intensive nature requires the setting-up of a Data Management Plan that guides the user to process both large amounts of data and a wide variety of data types, personal data included. The Data Management Plan (DMP) that has been described in this document is a necessity for organisational, technical and legal management of all the data types that will be processed. The ELG platform manages a very rich variety of data and these data have to be addressed in compliance with both GDPR and FAIR principles, in terms of legal compliance, as well as in agreement with language data, technologies and service needs.

In this regard, the current DMP provides recommendations, best practices and policies on how to address every data exchange and processing case for both personal and non-personal data. A granular typology of the data we may encounter in the ELG platform is provided, analysing all principles, constraints and conditions that will ensure aspects like security and compliance with GDPR in areas such as data access or storage. The GDPR guarantees factors such as data privacy impact assessment, implementation of a robust Privacy Policy, while watching over limitation of access, security and anonymisation aspects, all of them key aspects of our data processing needs.

Then, we move on to the application of FAIR principles to make all ELG data compatible as FAIR data. This helps us establish policies to reach the required Findability, Accessibility, Interoperability and Reusability factors, which also contribute to the sustainability of the data.

Last but not least, the DMP analyses data management policies for non-personal data: language data, technologies and services, the core of the ELG platform. All different types of data are listed and we take the opportunity to go over the specifications for the production of language resources, as well as for compatibility with the ELG metadata scheme and a brief outline of the policies behind data hosting.

As a preview to what the ELG platform already offers in terms of datasets, Appendix A lists the repositories that have been and are being ingested into the ELG catalogue and lists the number of resources available per resource type.

A. Repositories ingested and datasets available through ELG

This appendix lists the repositories that ELG has been working on till present in terms of dataset ingestion into the ELG catalogue. These are presented in Table 1.

	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	**Total**
ELRA	635	545	–	**1180**
ELRC-SHARE	844	43	–	**887**
META-SHARE	52	12	7	**71**
LINDAT/CLARIAH-CZ	243	66	–	**309**
ELRA-SHARE-LRs (2014-2018)	46	25	–	**71**
Zenodo	36	37	–	**73**
Total	**1857**	**727**	**7**	**2591**

Table 1: Repositories ingested by M24

The repositories that are in blue are considered as already concluded given that their conversion or harvesting has been done for all targeted LRs. Despite of this, some regular automatic harvesting protocols are being put into place so as to recover and ingest new arrivals to those catalogues. The repositories in black are still in progress, together with Quantum Stat and ELRA-SHARE-LRs 2020 (these latter are still under analysis).

The procedures to ingest these repositories have been different, depending on their metadata and protocols. The following have been applied:

- Metadata conversion through building of converters and then uploading of metadata descriptions: ELRA, META-SHARE and ELRA-SHARE-LRs
- Mapping and harvesting through OAI-PMH protocol: ELRC-SHARE and LINDAT/CLARIAH-CZ
- Mapping and manual enriching of metadata: Zenodo and Quantum Stat

Access to the ingested resources also follows different approaches:

- Some resources are hosted within ELG: META-SHARE and Zenodo
- Others just provide metadata records referring back to the source repository: ELRA, ELRA-SHARE-LRs, ELRC-SHARE and LINDAT/CLARIAH-CZ