european lexicographic infrastructure

# D10.2

# DATA MANAGEMENT PLAN

Author(s): Simon Krek, Iztok Kosem

Date: 31 January 2020

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

# D10.2 Data Management Plan

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

| | |
|---|---|
| Deliverable Number: | 1.0 |
| Dissemination Level: | PUBLIC |
| Delivery Date: | 31 January 2020 |
| Version: | 1.0 |
| Author(s): | Simon Krek, Iztok Kosem |

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

| | |
|---|---|
| Project Acronym: | ELEXIS |
| Project Full Title: | European Lexicographic Infrastructure |
| Grant Agreement No.: | 731015 |

## Deliverable/Document Information

| | |
|---|---|
| Project Acronym: | ELEXIS |
| Project Full Title: | European Lexicographic Infrastructure |
| Grant Agreement No.: | 731015 |

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

# Table of Contents

1 Data Summary ... 5

2 FAIR data ... 5

2.1 Making data findable, including provisions for metadata ... 5

2.2 Making data openly accessible ... 6

2.3 Making data interoperable ... 6

2.4 Increase data re-use (through clarifying licences) ... 7

3 Allocation of resources ... 8

4 Data security ... 8

5 Ethical aspects ... 9

6 Other issues ... **Error! Bookmark not defined.**

7 Further support in developing your DMP ... 9

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

# 1 Data Summary

This document contains the **initial** Data Management Plan for ELEXIS, at the stage before extensive collection of ELEXIS data. This document answers questions on data collection, storage, security and preservation as well as any ethical or legal issues that need to be addressed within ELEXIS regarding data sharing. The Data Management Plant will be updated every time changes will be implemented with respect to the data management. Changes will be made available on-line, and provide in annex to periodic reports.

The general purpose of ELEXIS collection of **lexical data** is to create links between dictionaries, and make the links available in the Linguistic Linked Open Data cloud. This is related to the objective of the ELEXIS project to create a scalable, multilingual and multifunctional language resource by linking lexical content and interlinking it with other structured or unstructured data - corpora, multimodal resources, etc.

Original ELEXIS lexical data comes in different digtal formats, and si converted in the process of linking to XML TEI Lex0 and/or RDF Ontolex-Lemon data model, which are the target formats in ELEXIS. Size of ELEXIS data is not known yet, as the extensive collection process hasn’t begun at the point of submission of this version of DMP.

Existing lexicographic resources that are available in the ELEXIS consortium and in the observer community under open licenses will be harmonised and converted into new data sets with common standardised formats and data models. Data with restricted access will be linked, and links will be made available under open access creating a new resource called “dictionary matrix”.

In addition to lexicographers, taget users of the collected data and the new resource (dictionary matrix) include Semantic Web, artificial intelligence, NLP and digital humanities research communities. Data will be included in Linguistic Linked Open Data cloud: https://linguistic-lod.org/.

# 2 FAIR data

## 2.1 Making data findable, including provisions for metadata

The main repository for ELEXIS resources will be Lindat CLARIN.SI repository which uses handles as persistent identifiers. ELEXIS project and infrastructure will follow conventions endorsed by CLARIN.SI as part of CLARIN ERIC, in relation to the use of metadata, versioning, and other standards used by CLARIN services:

- Depositing services: storing ELEXIS lexicographic resources in a sustainable repository at a CLARIN centre
- Virtual Language Observatory: discovering ELEXIS language resources using a browser or a

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

map

- Easy access to protected resources: authentication with institutional username and password.

## 2.2 Making data openly accessible

The new resource called "dictionary matrix" – a collection of links between dictionaries as a result of the linking process – will be available under open license (Creative Commons). For other data, copyright owners will specify availability under main CLARIN license categories (PUB, ACA, RES). Following the work in WP6, recommendations about open access to data and software in ELEXIS have been produced and published as part of the D6.2 deliverable (Recommendations on legal and IPR issues for lexicography):

- It is recommended that open licenses are used whenever possible, using standard licensing schemas.
- Licensing and Intellectual Property rights issues connected to data or software should be carefully considered at the very beginning of a lexicographic project, i.e. at the planning or proposal writing stage. Consulting legal experts and possibly lexicographic community is recommended.
- If an open license cannot be used for the entire lexicographic dataset, using different licenses for different types/parts of lexicographic data should be considered.
- For software, one of the FLOSS licenses is recommended, preferably GNU Public License v3.0, GNU Library or Lesser General Public License v.3.0 and Apache License v2.0.
- To address potential safety concerns and IPR abuse, it is recommended to use established repositories that use authentication protocols.

## 2.3 Making data interoperable

Both TEI Lex-0 and Ontolex-Lemon are envisaged as standards for best practices in lexicography, and are supported within ELEXIS. However, to ensure semantic interoperability between these diverse dictionary structures, ELEXIS will establish a common model. Such a model is necessary to a) streamline the integration of lexicographic data into the ELEXIS infrastructure, b) to allow reliable linking of the data in the dictionary matrix, and c) to form a basic template for the creation of new

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

lexicographic resources, such that they can automatically benefit from the tools and services provided by the ELEXIS infrastructure.

The aim ELEXIS is not to develop a fully-fledged data model.  Neither does the project aim to replace existing models (TEI Lex-0 and Ontolex-Lemon).  The main aim is to ensure semantic interoperability between lexicographic resources predominantly using their own custom format.  The model developed in OASIS by members of LEXIDMA technical committee (TC).

The LEXIDMA TC's purpose is to create an open standards based framework for internationally interoperable lexicographic work. The TC will develop a simple, modular, and easy to adopt data model that will be attractive for all lexicographic industry actors across companies and academia as well as geographic locations. Adoption of that model will facilitate exchange of lexicographic and linguistic corpus data globally and also enable effective exchange with adjacent industries such as language services, terminology management, or technical writing.

The TC will describe and define standard serialization independent interchange objects based predominantly on state of the art in the lexicographic industry. Defining specific serializations, transaction models, standard interfaces, and web services based on the defined objects and object models is also in scope as far as it facilitates the high level purpose set out here. It aims to develop this lexicographic infrastructure as part of a broader ecosystem of standards employed in Natural Language Processing (NLP), language services, and Semantic Web.

## 2.4 Increase data re-use (through clarifying licences)

ELEXIS „Survey on licensing" (Recommendations on legal and IPR issues for lexicography – D6.2) established that the main concerns lexicographic institutions have with regards to sharing their data are:

- Commercial use of their data, especially by competitors. The concerns are especially connected with producing low quality products for profit generation only.
- Unclear status of data because they were obtained from corpora with licensing restrictions.
- Lack of standardized documentation for sharing lexicographic data.
- Misuse by others, e.g. use beyond the purposes allowed by the license. Also, misuse may result in breach of contract with data provider, e.g. when making a corpus.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

- Fear of someone else being faster with data analyses or source preparation.

In general, findings show that there are numerours issues in lexicography regarding the possibility to make data open and re-usable.

The survey showed that out of 38 institutions, when asked about different types of lexicographic data, most institutions reported to be willing to share, under different licenses, lemma lists (28 institutions). Many institutions would also be willing to share examples (23), synonyms (22), sense structure (22), morphological information (22), definitions (21), collocations (20), fixed expressions (19), frequency information (19), and syntactic information (18). Fewer institutions reported willingness to share etymological information (14), pronunciation information (12), and frequently misspelled word forms of lemmas (9).

Therefore, two recommendations are particulary important in relation to the possibility to re-use data through clarifying licences: (1) using different licenses for different types/parts of lexicographic data, and (2) use of established repositories with reliable authentication protocols.

# 3 Allocation of resources

At this point it is not possible to assess exact allocation of resources for making data FAIR. Collection of lexicographic data is part of WP1 (Lexicographic data and workflow), and WP2 (Interoperability and Linked (Open) Data) is in large part dedicated to linking lexicographic data, which contribute significantly to making lexicographic data available under FAIR principles. The amount of person-months allocated to the two work packages amounts to one quarted of the whole ELEXIS budget.

ELEXIS project is relying on CLARIN and DARIAH as long-term infrastructures to take over the responsibility for maintaining access to ELEXIS data.

ELEXIS PI is the reponsible person for data management.

# 4 Data security

Only authenticated users can deposit items in CLARIN.SI repository. Users without home organisation can register at clarin.eu and authenticate using clarin.eu website account. Alternatively, CLARIN.SI Help Desk can create a local account for users without the possiblity to register at clarin.eu.

In general, ELEXIS data is stored according to the rules of CLARIN.SI repository.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

# 5 Ethical aspects

As ELEXIS data is lexicographic in nature, there are no ethical issues.

# 6 Further support in developing your DMP

This version of Data Management plan is the initial version. A more elaborate version will be available after the launch of LEX1 infrastructure, and after M36 when conversion, linking and other services will be available.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.