



**Innovative and Inclusive Democratic Spaces for Deliberation and Participation**

HE-101132431

**D6.6 Data Management Plan (initial)**

<b>Dissemination Level</b>	Public
<b>Contractual date of delivery</b>	30/06/2024
<b>Actual date of delivery</b>	
<b>Work package</b>	WP6
<b>Tasks</b>	T6.4
<b>Approval Status</b>	Draft/Final
<b>Version</b>	.03
<b>Lead Beneficiary</b>	UPF
<b>Contributing Beneficiaries</b>	ALL

**Summary**

This document describes the first version of the Data Management Plan (DMP) in accordance with the Grant Agreement. Providing a DMP in an early stage of the project is important to ensure that data is correctly communicated between the project partners. A DMP helps to identify potential problems in advance, avoid data loss, leakage of sensitive data and additional costs which may arise by the choice of inappropriate data format. It also ensures that the publication of data is set to be useful after the end of the project, being one of the main project goals.

**Plain Language Summary**

This document is the first edition of the Data Management Plan.  
 A data management plan is a document that explains how data will be treated during a research project and after it has ended.  
 It is important to have a plan for managing information early in the project.  
 This helps to avoid problems with sharing information between partners.  
 The plan helps us find problems early, prevent losing or sharing information, and save resources.  
 It also ensures that we can share the information after we finish the project.



**Funded by  
the European Union**

*This document is part of a project that has received funding from the European Union's Horizon Europe research and innovation program under the Grant Agreement No. 101132431. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. UOL is funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10103529)*

<b>Author List</b>		
<b>Organisation</b>	<b>Name</b>	<b>Contact Information</b>
UPF	Stefan Bott	stefan.bott@upf.edu
UPF	Sandra Szasz	sandra.szasz@upf.edu
UPF	Horacio Saggion	horacio.saggion@upf.edu
MAC	John O'Flaherty	j.oflaherty@mac.ie
CAPITO	Martin Gollegger	martin.gollegger@capito.ai
CAPITO	Verena Riegler	verena.riegler@capito.ai
NEXUS	Volkan Sayman	sayman@nexusinstitut.de
UOL	Serge Sharoff	s.sharoff@leeds.ac.uk
CIB	Lian Muñoz	lian.munoz@cibervoluntarios.org
PIM	Almudena Rascón Alcaina	almudenascon@plenamadrid.org
ANFFAS	Eleonora Severa	esevera@anffas.net
AAI	Claudia Mazzanti	claudia.mazzanti@actionaid.org
BOO	Eva Garcia Chueca	egarciach@bcn.cat
IMPD	Laura Trujillo	ltrujillo@bcn.cat

<b>Document History</b>			
<b>Version</b>	<b>Date</b>	<b>Action</b>	<b>Revised by</b>
0.1	05/05	First draft	UPF
0.1	23/05	Contributions uploaded	UPF, ALL
0.2	03/06	Submitted to reviewers	UPF
	10/06	internal review provided	UOL, Nouran Khallaf
	17/06	internal review provided	UPF, Josep Marti
0.3	19/06	Revisions applied and new contributions added	UPF, all
1.0	21/06	Final version prepared	UPF
	25/06	Submitted	UPF

## Table of Contents

<b>1. Overview</b>	<b>6</b>
1.1. Data Summary	6
1.2. Sensitive Data Storage	8
1.3. Findable, Accessible, Interoperable, Reusable (FAIR) data	9
<b>2. Data summary</b>	<b>10</b>
Purpose of the data generation or re-use and its relation to the objectives of the project	10
Data types and formats generated or re-used	12
Re-use and purpose of existing data	14
Origin or provenance of the data	15
Expected size of the data	17
Data usability/utility outside your project	18
<b>3. FAIR Data: Making data findable, including provisions of metadata</b>	<b>19</b>
Data identification (persistent identifier)	19
Rich metadata provision	20
Search keywords in the metadata	21
Metadata: harvested and indexed	22
Openly available data	23
Data availability, accessibility and repository	24
Methods or software tools to access the data	24
Software documentation	24
Relevant software inclusion	25
Deposit of data and associated metadata, documentation and code	25
Arrangements with the identified repository	25
Access specifications and restrictions on use	25
Data access committee	25
Described conditions for access	26
Person identification accessing the data	26
Data and metadata specification	26
Documentation provision for data validation analysis and data re-use	27
Data licence	27
Third parties data use	28
Data provenance documentation	28
Data quality assurance processes	29

---

<b>4. Other research outputs</b>	<b>30</b>
<b>5. Allocation of resources</b>	<b>31</b>
FAIR data making costs or other research outputs	31
Cost budget	31
Data management responsibility	31
Long term data preservation	32
<b>6. Data security</b>	<b>32</b>
Provisions	32
Safely stored data in trusted repositories for long term preservation and curation	33
<b>7. Ethics</b>	<b>34</b>
Ethics or legal issues that can have an impact on data sharing	34
<b>8. Other issues</b>	<b>34</b>
Use of other national/funder/sectorial/departmental procedures for data management	34

<b>Acronyms</b>	
ASCII	American Standard Code for Information Interchange
ATS	Automatic Text Simplification
BEA	Building Educational Applications
CVS	Comma-Separated Values
D	Deliverable
DMP	Data Management Plan
DOI	Digital Object Identifier
DS	Data Storage
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable data
GA	Grant Agreement
GDPR	General Data Protection Regulation
HE	Horizon Europe
iDEM	Innovative and Inclusive Democratic Spaces for Deliberation and Participation
M	Month
OP	Open Science
ORCID	Open Researcher and Contributor ID
OSC	Open Source Code
OSS	Open Source Software
SNIML	Simplified News in Many Languages
T	Task
TSV	Tabular Separated Variable
WP	Work Package

## 1. Overview

A data management plan (DMP) is a document which establishes the methodology for collecting, storing, handling and - if necessary - destroying the data produced or used in a research project. In the Horizon Europe programme it is an essential document which reflects upon the need for the data used in the project, how it contributes to the project objectives, and anticipates the risks associated with its handling. The development of a data management plan in an early stage of the project is important for a series of reasons: a DMP helps to identify potential problems in advance, avoid data loss or leakage of sensitive data and additional costs which may arise by the choice of inappropriate data formats. It can prevent data insufficiency caused by insufficient planning ahead of data collection. Last, but not least, the creation of a DMP is a good moment to reflect on what goals the project has, how these goals should be reflected in the data to be generated and how expensive data collections can be turned into resources which will serve the research community in the best possible way. The last point is central, considering that, in some projects, collecting data consumes considerable resources; it has the highest potential for being useful after the project ends and should therefore be treated with special care.

This document describes the initial iDEM Project Data Management Plan (DMP)<sup>1</sup> in line with the description drafted in the GA DMP in the impact section, compliant with Horizon Europe programme and it is updated throughout the lifetime of the project<sup>2</sup>.

It specifies the type of data to be collected, its need and utility, the procedures involved in its collection, its secure storage, the standards and guidelines followed for its annotation. Moreover, this plan details what and when data will be made publicly available following the principles of Open Science to ensure that experiments can be reproduced, always protecting sensitive data containing information and complying with the data protection legislation.

### 1.1. Data Summary

The iDEM project will produce and use a large array of data. The following is a list of the main data sources that will be used, both the data which is foreseen to be produced within the project and the already existing data which is owned by consortium members and can be used in the project.

As for the data to be generated or recompiled within the iDEM project we envision the following types of data.

---

<sup>1</sup> A Final DMP is due M18, D6.7 Data Management Plan (final) as stated in GA.

<sup>2</sup> This deliverable is the result of the work developed in T6.4.

- The iDEM Corpus: A corpus of around 300 documents selected from publicly available sources in the broad domain of democracy and civic participation to be manually annotated (including metadata) by experts in WP2 and used in WP3.
- A further selection of publicly available documents (e. g. From the Conference of the Future of Europe documents or similar resources) or databases (e.g. public lexica) of various sizes (from hundreds of sentences to thousands of Wikipedia pages) which are publicly available in order to train or test natural language processing systems which will be re-used in text simplification and complexity.
- Data from WP4: data (text, transcriptions) from interviews, focus groups, workshops, the use case pilots and questionnaires obtained from people with/without intellectual disability who will voluntarily participate in the annotation design process (WP2) and in use cases design and execution (WP4),
- Public/official documents (around 30 per use case) to carry out testing in WP4.
- The MLSP-2024 dataset: UPF TALN has participated in the first month of the project in the preparation of a 10 languages dataset with manually annotated data on lexical complexity and lexical simplifications of selected words. UPF was responsible for the Spanish and Catalan data and partially responsible for Italian and Japanese data. This data is essential for the training and testing of lexical simplification, especially for Catalan and Spanish, where no comparable datasets existed before. The whole dataset was published as test data for a shared task<sup>3</sup> at the BEA 2024 (Building Educational Application)<sup>4</sup> Conference.

The iDEM project will also benefit from previously existing data. In section 2 below we will address data re-use in more detail, but we would like to mention the data here which is intellectual property of some of the consortium members or has been created by them for settings which are closely related to the problems to be tackled in iDEM:

- The Capito Corpus is a native German-based corpus collecting real-world simplification jobs. These are all human-translated and human-verified which serves as a very strong foundation for further research. All texts are produced by Capito or one of its franchise partners. No web-crawling was used. The corpus is translated into Spanish using DeepL translation which is not(!) human-verified or quality-checked up to this point. This has to be considered strongly.
- UoL has developed a document-aligned representation of articles from Vikidia and Wikipedia for Catalan, English, French, Italian and Spanish. This will be used in order to train or test natural language processing systems which will be re-used in text simplification and complexity.

---

<sup>3</sup> <https://sites.google.com/view/mlsp-sharedtask-2024>

[https://github.com/MLSP2024/MLSP\\_Data/](https://github.com/MLSP2024/MLSP_Data/)

<sup>4</sup> <https://sig-edu.org/bea/2024>

## 1.2. Sensitive Data Storage

Sensitive data needs special care when it is stored, so it cannot be accessed by anyone who is not explicitly intended to have access to it. First of all, this includes the data directly produced by participants in interviews, questionnaires, focus groups and use cases. But sensitive data refers not only to the collected responses from users but also information sheets and consent forms. Information sheets are brief documents which describe the experiment or procedure in which participants engage, explaining their rights, the purpose of the research, potential risks and benefits, and the way confidentiality will be maintained. The consent form also gives information about the procedure, risks and benefits and requires the participant to sign in order to give official consent and to participate in the above mentioned groups.

By sensitive data we also understand the points of view they express in an environment they trust and may not be intended to be known outside this environment. Even if we do not expect this to be the case in our use cases, some points of view expressed could be potentially harmful if third parties can attribute these views to the author. In case the content of a discussion will be used for public uses, we will always make sure that the author of a contribution cannot be directly or indirectly identified.

Only the necessary personal data from participants will be collected and securely stored in password protected and encrypted repositories compliant with (EU) 2016/679 General Data Protection Regulation. This personal data will only consist of contact information of the participants in order to organise the groups, and will be deleted after its use. No personal data collected in the EU will be shared with the partners in non-EU countries.

Only data produced in WP2 and necessary for training models - which raises no ethical concerns, as it is not of a personal type - will be made public and shared with the scientific community via open repositories.

Data preservation will also be limited: personal data is stored in pseudonymized and encrypted form for 18 months after the interview's date. For the purpose of pseudonymization, IDs are assigned to the interviewees and the key is stored separately. The key is stored in a physically secure location and is only accessible to authorised project staff. The key and all personal data will be deleted 18 months after the interview has been conducted. After deletion of the key and all personal data, interviews are stored in anonymized form until the end of the project period (12/2026). The destruction of data will be the responsibility of the assigned responsible for data security as specified below in section 6. Participants will be informed about the data processing and their rights (e.g. disposal of their records) in the informed consent forms, which will be reviewed by the ethics committee prior to starting interviews and focus groups.

Contact data collected for the initiation and scheduling of interviews (telephone number, e-mail, postal address) will be used by authorised partners exclusively for the stated purposes. Data will

be kept of those participants who unambiguously agree to participate and immediately delete the data of all those who clearly reject participation after the rejection. If a potential interviewee does not answer our request for an interview within one month, we will count this as a rejection and immediately delete all personal data. However, a frozen (blocked) copy of the contact data (emails) of individuals who declined participation will be retained solely to avoid sending duplicate invitations. The list of frozen contact data will be deleted as soon as the list of all participants of interviews is finally set.

### 1.3. Findable, Accessible, Interoperable, Reusable (FAIR) data

The last decades have been characterised by an increasing importance of data in science, business and the daily life of most people living in a modern society.<sup>5</sup> In order for data to be beneficial it not only has to be plentiful (“more data is better data”<sup>6</sup>), but it also has to comply with a series of requirements, such as openness and accessibility. Data is a valuable resource; as it is expensive to create and collect. Therefore good and widely accepted principles for data management are extremely important for all stakeholders which in some way handle or depend on data. One of the best attempts to give an answer to the needs of getting the best value for society out of existing and newly created data is the development of the principles of FAIR data, which state that data should be findable, accessible, interoperable and reusable. Based on much preliminary work, the FAIR principles were endorsed by the G20 in 2017<sup>7</sup>, which also shows their central importance. Today the research community can rely on services which make sure that the FAIR principles are respected. For the current purpose it is important to note that platforms like Zenodo (<https://zenodo.org/>) and GitHub (<https://github.com/>) are based on FAIR principles as their design foundation and do the best they can to support researchers, developers and data consumers in managing data in accordance with them. One of the main central purposes of these repositories by design is to make data accessible in a FAIR manner.

In order to impact the research community and provide value beyond the project itself, iDEM will embrace the Open Access mandate for its publications and participate in the Open Research Data pilot. It will also adopt an Open Source approach to software development publishing code in public repositories (e.g. GitHub, Zenodo) in terms to be specified in the consortium agreement. The open software will also contain processes to access and manipulate the text data. A GitHub repository for iDEM was already created.<sup>8</sup> A Zenodo repository is still to be created.

---

<sup>5</sup> This is sometimes referred to as the fourth industrial revolution. See, e.g.: <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>

<sup>6</sup> A famous phrase originally coined by Church and Mercer, which has later been questioned and amended over the time: Church, K., Mercer, R. (1993): Introduction to the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics 19:1, pp. 1-24.

<sup>7</sup> [https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT\\_16\\_2967](https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967)

<sup>8</sup> <https://github.com/iDEM-eu/github>

As the Horizon Europe programme requires that all publications arising from the work funded in the project must be published in Open Access (green and gold route), iDEM will provide its results as immediate open-access (open-access journal). Participants will also self-archive via institutional repositories (e.g., Horizon-Zen, a Zenodo initiative)<sup>9</sup> as recommended by OpenAire, but also the institutions' repositories, such as UPF e-Repository). Standards such as taxonomies and free keywords will be used to tag the bibliographic records to allow the information to be findable. Sharing the non-personal data (e.g. annotated text corpora) will accelerate the development of solutions for text to easy-to-read translation.

## 2. Data summary

### **Purpose of the data generation or re-use and its relation to the objectives of the project**

The partners within the iDEM consortium have different goals and needs for the collection and usage of the data specified in section 1.1. UPF TALN and UOL are classic research institutions interested in creating models from data, use these models for experiments and report on the experiments in scientific publications. For the ONGs in the consortium (PIM, AAIT, ANFFAS, IMPD and BOO) the main purpose of data collection and use is to study and understand the factors that hinder accessibility. CAPITO and NEXUS will need data for the improvement of their services and products. MAC mainly needs data for technical development of the iDEM application and the iDEM API. The details of the data purpose of each project partner or group of partners is as follows:

#### *UPF TALN*

UPF TALN will use data for carrying out experiments on Automatic Text Simplification and develop Text Simplification software. Experimental outcomes will be used for dissemination and peer-reviewed scientific publications. The developed software will be used to meet the project requirements, specifically in tasks T2.3, T3.1. and T3.2. Both experiments and software development will produce either document simplifications, sentence simplifications, lexical simplifications (of individual words) and ratings on the complexity/difficulty of these units.

#### *CAPITO*

The generated data will depend on the results of WP1 which examines available historical evidence and various forms of deliberative democratic civic participation in Europe: from spontaneous forms of engagement to organised participatory and traditional engagement. These insights direct the selection of texts which are to be simplified. The selection of these texts is of the highest importance since they influence the results drastically. Selection criteria could be for example topic, domain, scope (international/ national politics), different styles (direct

---

<sup>9</sup> iDEM has already made contact (2/2024) with the Horizon-Zen project to become an early adopter.

speech etc), length and complexity. The input text serves as data in this context. It is yet to be defined and discussed if an existing linguistic corpus from Capito would be amplified to serve the project objectives.

#### *NEXUS*

The purposes of data collection by NEXUS are the production of knowledge and practical insights concerning the accessibility barriers to democratic participation for people with cognitive disabilities and their marginalisation in democratic spaces, the assessment of the engagements of hard to reach groups and vulnerable populations in participatory and deliberative processes and ways to enhance the production of inclusiveness by facilitators.

The data will be produced and interpreted deploying qualitative methods of social scientific research and analysis, namely focus groups, semi-structured interviews, a structured analysis of scientific and grey literature and qualitative content analysis. In a next step, the knowledge and practical insights are supposed to be fed into a solution ideation and prototyping process. Finally, the prototypes are tested in a series of use case pilots, which will also produce data, because the use cases will be documented.

The purposes of re-use of data will be validating or expanding existing data for the benefit of scientific communities, re-using existing data within the scope of the development of technologies for vulnerable populations and hard-to-reach groups and re-using the data for designing participative spaces for the inclusion of vulnerable populations and hard-to-reach groups.

#### *MAC*

MAC will need data for technical development. Users input data will be stored on their own devices, as they use the iDEM service. MAC will use email addresses of iDEM Approved Users only to test and get feedback on the iDEM services. Only the resulting anonymised data will be used for deliverables and reports on the iDEM Google Drive for internal use. The Open Source code resulting from the data exploitation and the resulting technical development will be made available on Github.

#### *CIB, PIM, ANFFAS, AAIT, IMPD and BOO*

During the project, iDEM ONGs and users' organisations will work with available pre-existing data (e.g. guidelines, brochures, past experiences) and new data (e.g. gathered in focus groups and use cases). The purpose of the data collection is to obtain information which could contribute to understanding text accessibility barriers so that actions can be proposed to improve accessibility and to increase the degree of inclusion/participation of people with disabilities and hard-to-reach groups in democratic deliberative processes.

Pre-existing data will be used relating to the organisations, papers and other sources that work on the reality of people with intellectual disabilities and hard-to-reach groups in different contexts and that have expert staff to attend or know about their needs.

New data will be generated through discussion groups, with people with intellectual or cognitive disabilities and hard-to-reach-groups, interviews with different experts and the activities that will be prepared for the piloting in three different countries. The data collected in focus groups and interviews will provide information about experiences, barriers and solutions for deliberative processes, and feedback for process improvement. This data will be used to strengthen/validate the iDEM theoretical framework (WP1) and to develop/test a technology that responds to end user's needs (WP2-4). Some types of data could contribute to the project dissemination (WP5) and the final solution altogether.

#### *UOL*

UOL will develop classification tools aimed at assessing the difficulty level of various texts. To train these classifiers effectively, we will utilise existing data. Our approach includes leveraging a document-aligned representation of articles sourced from Vikidia and Wikipedia, covering languages such as Catalan, English, Italian, and Spanish. This representation ensures that articles in different languages are aligned for better comparison and training accuracy.

We will first develop a readability classifier to evaluate entire texts and then create a classifier for sentence-level assessment. We will experiment with various models, including both monolingual and multilingual classifiers, and keep track of accuracy and performance to find the best classifier model. The next step will be to develop a text simplification model, further enhancing our tools' usability and effectiveness for educational and linguistic applications.

### **Data types and formats generated or re-used**

Within the consortium different partners will have different types of data and different data formats. Despite this variety of needs, all data will be in standard formats, which include UTF-8 or ASCII encoded text files, CSV and TSV tables, XML, Word Documents, Google Docs, Excel Tables and Google sheets. Nexus will have the need to use data in MAXQDA files, which is a proprietary format that requires a licence. Any dataset for either internal use or dissemination will be accompanied by a readme file in plain text or markdown format. This is especially important for data published in our GitHub and Zenodo repositories (see below).

#### *UPF*

Most of the data that UPF-TALN will use will be UTF-8 encoded and organised in tables formatted as Tabular Separated Variable (TSV). For Lexical Simplification the standard of representation is tabular, with fields for the context sentence, the target word and the possible substitute word. A similar format is used for Complex Word Identification and Lexical Complexity Prediction. For document and sentence simplification, a tabular format which places original text and simplified text next to each other will be used. These formats may be converted to or be converted from Excel (or similar), XML and JSON formats, but since TSV tables are usually a sufficient format and the easiest one, we will use TSV whenever possible and if not requested or provided otherwise.

---

*CAPITO*

Overall, the domain of this project uses text-based data formats. Linguistic data will be represented in an XML-based format. The criteria catalogue and the translated and simplified texts (part of WP2) will be in docx and pdf format.

*NEXUS*

All data derived from the implementation of focus groups and semi-structured interviews will be stored in Word files (.docx) and the data produced by a structured literature analysis will be stored in Excel files (.csv). Qualitative content analysis will be conducted using the software MAXQDA and the results of the analysis will be stored as MAXQDA files (.mx22 or .mc22).x

*MAC*

Users' own device-held data will be mainly stored as plain text in ASCII. Email addresses will be mainly held in CSV files. Internal reports and deliverables will be mainly Google Docs, Sheets, PDFs and JPEGs. The source code will be stored on Github in ASCII.

*CIB, PIM, ANFFAS, AAIT, IMPD and BOO*

During the project, we will work with both already existing data and data which is newly created in the use case pilots.

Firstly, with the bibliographic search and statistical databases, information will be obtained on the targeted population, its characteristics and in which entities in the city they participate.

Once the people who will participate in the pilot projects, use cases and focus groups have been identified, new data relating to the specific study topic "how to make participatory processes more accessible" will be generated.

The data to be collected, derived from forms, and compiled, for example in the Focus Groups, will be in the following formats: text (pdf, docs), numbers (xlsx), images and graphics (pdf, jpeg/jpg, png, psd, pptx), voice (maybe recorded, mp3), Tables (xlsx) and video (mp4, only for dissemination purposes).

Observational data, testimonies, opinions and personal data will be stored in text format, using Microsoft Office (xlsx, docx), pdf, ASCII/UTF8 text and paper.

*UOL*

Data used by UOL will be mainly represented as plain text in ASCII or UTF-8. Tabular information will be represented in CSV/TSV format.

## Re-use and purpose of existing data

Within the consortium, different partners have different possibilities to access and re-use existing data. The common problem is that no data or little data is available for our specific project goals. In what follows we give a detailed overview of the project partner's potential pre-existing data which may serve our purposes.

### UPF

For the text simplification and complexity assessment tasks which will be carried out by UPF TALN, there are several available datasets which were developed as gold standards for similar tasks. The main problem is that few resources exist for the three main languages targeted by iDEM: Italian, Catalan and Spanish. And of those resources that exist, none covers the textual domain of deliberation and democracy. Still, for the first experiments and as additional training data UPF will use already existing datasets.

There are some datasets available for Spanish, Catalan and Italian. The SIMPLEX corpus<sup>10</sup> was produced within a research project with the same name; it contains 200 Spanish news texts with manual simplifications produced on the Easy-to-Read methodology. The Newsela dataset<sup>11</sup> contains English and Spanish news documents with simplifications on different difficulty levels. For Spanish it has 1221 documents. SNIML (Simplified News in Many Languages)<sup>12</sup> is composed of documents in 5 languages, including Italian. MultiChroCane is a dataset of around 5k cross-lingual sentence pairs from the medical domain and 100k noisy cross-lingual pairs. For lexical simplification (which targets singled-out words in text contexts), there are several datasets, some of which also include ratings on Complex Word Identification. The ALEXSIS<sup>13</sup> Spanish Dataset for Lexical Simplification contains 381 instances. Each instance is composed by a sentence, a target complex word, and 25 candidate substitutions. The MLSP2024 dataset contains target words for lexical simplifications, together with possible substitution words which are easier. It has data for 10 languages, including Spanish, Catalan, English and Italian. For each language, there are 600 target words, taken from 200 context sentences. The EASIER corpus<sup>14</sup> contains 8155 Complex words in context from 260 Spanish documents with 7892 suggested synonyms. The CWI2018<sup>15</sup> the dataset contains 17000 Spanish target words in context for the classification of complex vs. non-complex lexical items.

### CAPITO

It is yet to be decided whether and to which extent the existing CAPITO linguistic corpus is to be reused in the iDEM project. The original texts are pre-existing data, while translations and

---

<sup>10</sup> Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., & Bourg, L. (2011). Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47), 341-342.

<sup>11</sup> <https://newsela.com>

<sup>12</sup> <https://pub.cl.uzh.ch/wiki/public/sniml/start>

<sup>13</sup> <https://zenodo.org/records/5837149>

<sup>14</sup> [https://github.com/LURMORENO/EASIER\\_CORPUS](https://github.com/LURMORENO/EASIER_CORPUS)

<sup>15</sup> <https://sites.google.com/view/cwisharedtask2018/datasets>

adaptations are new data. The entire corpus has to be translated into Catalan which poses certain challenges of maintaining the semantic information, sentiments and context. Additionally, it has to be examined if the domains and topics present in the existing linguistic corpus are of relevance to this project. With the strong focus on political participation and democracy, it is highly beneficial to have a linguistic corpus that includes explicitly these topics and to resemble the goals and focus of the project well. The Capito corpus is currently not publicly available. The corpus can be accessed via a server owned by Capito. It is under investigation to create a modified version of the corpus (tailored to the iDEM requirements) that is hosted on a repository, however not publicly available.

The iDEM corpus will consist of already existing documents, but the compilation and selection of the corpus itself will constitute new data.

#### *NEXUS*

The analysis of specific aspects of scientific and grey literature will proceed by analysing published results of research and practical experiments, not data in a preprocessed form. In this sense, we will not re-use any data.

#### *CIB, PIM, ANFFAS, AAIT, IMPD and BOO*

We will use the data for research purposes within the scope of the project and with the consent of all partners involved. The reuse of other data than those deriving from bibliographical references or investigation is not foreseen. The consortium will use these data for analysing and validating concepts, producing deliverables on a theoretical framework. For the reference repository, the consortium decided to use Zotero.

#### *MAC*

We will mainly use our own generated data as described in the section above. The iDEM service's Open Source Software (OSS) will only use external tools that are themselves free and OSS.

#### *UOL*

We have developed a document-aligned representation of articles from Vikidia and Wikipedia for Catalan, English, Italian and Spanish, which we will use as existing data.

### **Origin or provenance of the data**

#### *UPF*

The existing data that UPF TALN will use stem mostly from the scientific community. The datasets CWI2018, and MLSP2024 were collected and published for shared tasks at the BEA (Building Educational Applications) workshops of 2018 and 2023. UPF was actively involved in

data collection and annotation for MLSP2024. The Simplext corpus is not publicly available under an open-source licence. However, since UPF was one of the project partners of the Simplext project, we have access to this data. Please see the last section for a more detailed list of existing resources.

#### *CAPITO*

The generated data (by translation and simplification) will be based on a text selection made within the consortium, which will form the iDEM corpus. This selection process is crucial since the domain and content of the texts play a major role. Since the sample size is rather small in the realm of linguistic model training - 100 texts per language (Catalan, Spanish, Italian) - it is even of greater importance which texts are selected. Therefore, this is the major data generation aspect of WP2. With 300 total data sets, the scope is rather small compared to other linguistic research.

In terms of reusing data, a final decision has to be made in which form the existing Capito corpus can provide value to the project and its objectives. Most likely, an adaptation and amplification regarding the topic of political participation, democracy and overall politics is sensible.

#### *NEXUS*

The data generated by focus groups and semi-structured interviews will be generated in real-world settings. The analysis of the literature will produce data which stems from published results of research and practical experiments.

#### *CIB, PIM, ANFFAS, AAIT, IMPD and BOO*

Some of the data which might be used is the project partner's pre-existing data if it is available and apt: adaptations to easy to read, original texts and websites. Also guidelines about people with intellectual disabilities participation in democratic spaces. Further data will come from scientific literature and bibliographic references.

The main data will be generated during the course of the project and activities (interviews with experts, focus groups) during the life of iDEM project. The data from this primary research will be pseudo-anonymized or anonymized before being shared with the other partners if needed for creating the potential solution and results.

#### *MAC*

We will not use any pre-existing data.

#### *UOL*

We will use Vikidia and Wikipedia as freely available resources.

---

## Expected size of the data

The size of the data used by different partners varies a lot, depending on the tasks they are working on. The following is a list of data sizes, as estimated by the different organisations. The list does not include those organisations which do not intend to make data available to the public or to other project partners.

### *UPF*

The SIMPLEX corpus contains 200 Spanish news documents. The relevant Spanish part of the Nesela datasets contains 1121 documents. SNIML contains 1355 documents in 5 languages. The Spanish and Catalan parts of the MLSP dataset comprise some 600 target words per language, taken from 200 context sentences per language. Above we list more details on a larger selection of potentially useful resources which includes size information.

During the project execution, we do not expect that we can collect data which is larger than the MLSP datasets we already produced. We may collect Gold Standard data which is similar in size to the MLSP data if we can find time and resources for this data collection.

### *CAPITO*

We expect 600 documents (300 texts overall, but each with an original version and a simplified version). The length and extent of these documents is currently not yet defined. The final size also depends on the amount and granularity of computer-linguistic features that may or may not be needed. However, we expect it to be below 3GB.

### *NEXUS*

We expect to produce data with a size less than 1GB with individual files being between 1MB to 100 MB.

### *CIB, PIM, ANFFAS, AAIT, IMPD and BOO*

The data produced within the Focus Groups, the interviews and the use cases will be of limited size. We expect the text data to occupy not more than 1GB. The audio data will occupy between 10GB and 50GB and video data may occupy 100GB to 200GB.

### *MAC*

The data stored on user devices data will likely be a few MB at most. The email addresses will be less than 1MB. Internal reports and deliverables are likely to be <1GB, and the needed space on GitHub will be less than 5MB

### *UOL*

We will use 4,000-10,000 texts per language produced through scraping Vikidia and Wikipedia resources (Spanish, Catalan and Italian, as well as English). Total size 60MB.

## Data usability/utility outside your project

### *UPF*

The Spanish, Catalan and Italian parts of the MLSP2024 dataset, which were collected and annotated by UPF TALN are a contribution to the research community that works on lexical simplification. The iDEM consortium benefits especially from these datasets because it represents the first dataset on Lexical Complexity Prediction for the 3 languages. For Catalan it is the first dataset that was created for Lexical Simplification. Without this resource work on lexical simplification and lexical complexity assessment would not be possible within iDEM. Apart from the use within iDEM the MLSP2024 dataset is an important contribution to the research community dedicated to automatic text simplification. If UPF TALN has the possibility to collect further data it will also be made available as a contribution to the scientific community.

### *CAPITO*

The data generated in this project will be useful for other researchers and the scientific community that works in the area of easy language and readability assessment. This data (the iDEM corpus) is the translated and simplified text (tailored to the political domain). The iDEM corpus will be made publicly available to serve and support further research. With this data it may be possible to derive further observations and linguistic pattern recognition. Possible use-cases would be accessibility research or the development of readability metrics.

There is a pre-existing CAPITO corpus which is the IP of CAPITO. This corpus may be accessed by the Consortium, but not by the public. The linguistic corpus may be valuable data to support the process, but is independent of the project result.

### *NEXUS*

Beyond the consortium the data might be useful to a) researchers and practitioners interested in barriers for vulnerable groups in democratic spaces, the engagements of hard-to-reach-groups in participative processes and moderation techniques to enhance inclusivity from a scientific perspective, b) to research or applied research projects experimenting with technologies to enhance inclusivity in political participation and c) to research, applied research projects or commissioned participative processes aiming at including vulnerable populations in processes of political participation.

### *MAC*

The data held on users' devices data will be useful just for each individual user. Within the consortium, anonymised formative evaluation users' feedback and usage statistics will be available. The OSS will be available on Github to further developers and researchers.

---

*CIB, PIM, ANFFAS, AAIT, IMPD and BOO*

Raw and sensitive data will be shared only within the Consortium. Outside the project, it will only be accessible as results/conclusions of the project. The data used within the consortium will be helpful in creating scientific publications, verifying and validating the different tools and services that we would create and promoting the results of the research and final solution to reach the objectives of the project. In particular, investigation and testing will be carried out to verify and validate the theoretical framework and technology developed during the project.

The summarised data arising from the use cases will be useful for its participants and the institutions that helped collect it, in order to have a better understanding of the different processes.

Beyond the Consortium, there are different ways that the processed and summarised data (with all sensitive information removed) might be useful. The data could be used by general researchers interested in getting knowledge about the topic, in particular, academic researchers and students but also by companies that develop technologies similar to those developed in iDEM.

UOL

The data sets will be shared within the consortium as the working baseline to produce the scientific publications.

### **3. FAIR Data: Making data findable, including provisions of metadata**

#### **Data identification (persistent identifier)**

The consortium has agreed on the use of a Zenodo (<https://zenodo.org/>) repository which is to be created for the publication of any project-related data which is intended to be made publicly available. Zenodo automatically assigns DOIs (Digital Object Identifiers) to every upload. Unless otherwise stated below, the project partners will adhere to this agreement. By its very nature, all sensitive data is excluded from a FAIR treatment. Documents such as peer-reviewed articles will be identified through a DOI, either because the publisher already assigns one, or because it will be published on an open-access repository like Arxiv (<https://arxiv.org>). Identifiers for publications may also include ARK and URN identifiers. For the project partners which do not plan to publish datasets, including most ONGs (IMPD, BOO, AAIT, CIB, ANFFAS), this does not apply, because the need for persistent identifiers does not arise. For the Open Source Software (produced by MAC, UPF TALN and UOL) we will use repositories on GitHub. Since GitHub does not automatically assign DOIs, we will link any repository from Zenodo, which will ensure the assignment of a DOI.

## Rich metadata provision

Rich metadata in the context of linguistic data refers to additional information or descriptors that provide context and detail about the language-related content. This metadata enriches the linguistic data by offering insights into various aspects such as the origin of the text, its genre, authorship, publication date, language variety (e.g., dialect or register), and more.

For example, in the context of iDEM, rich metadata should be created in the process of selecting, translating and simplifying the chosen texts. That could include the country (where it was from), the time period (if for example an historical political reference is included), the author, the intended audience and a label of the topic (e.g. democracy, participation, dictatorship...), and sentiment (e.g. pro-democracy, anti-democracy, etc).

Only members of the iDEM consortium that plan to make data public will have to consider metadata. The partners that do not plan to make data public are not listed here.

### *UPF*

For the recompilation of the MLSP2024 dataset, in which UPF took part, very general demographic information was compiled: Age, years in education, weekly hours of reading, native language, and the number of foreign languages spoken. This information was pseudonymized and is not part of the dataset which was made publicly available. This metadata only exists in tabular form detached from the published main dataset. In the case that UPF collects similar data in future stages of the iDEM project, similar metadata will be created.

### *CAPITO*

Providing data sets with rich metadata would support the overall intent to make results publicly available to have the biggest positive social impact on this project possible. Overall, the results can be better understood and interpreted by other linguists. It allows for further nuanced analysis, even if the Capito corpus may not be publicly available. It could serve as a research basis for new text simplification models as well as societal awareness. The linguistic data is identified by a persistent identifier on the document level (called `global_id`). This `global_id` is linked to a hash to identify different simplification versions of the same original text. For example, 990-9f0cd46. 990 is the `global_id` and 9f0cd46 is the hash to ensure traceability. Each document has the original text and one simplification for each language level and a date or time stamp.

### *NEXUS*

We will follow the minimal standards of providing metadata required by the EU Open Research Repository.

These are:

- Visibility: Both public and restricted (with or without embargo and/or access request)
- Resource types: All resource types.
- Licences: Public and embargoed records MUST specify a licence.
- Funding information: Records MUST specify at least one grant from the European Commission.
- Creators: Creators SHOULD be identified with a persistent identifier (e.g. ORCID, GND, etc), and affiliations SHOULD be identified with a persistent identifier (e.g. ROR, ISNI, ...)
- Subjects: Records SHOULD specify one or more fields of science from the European Science Vocabulary.

#### *MAC*

For OSS we will adhere to GitHub requirements, otherwise we will just use normal directories.

#### *CIB*

Metadata will be provided to allow discovery. It will contain information such as title, authors, ORCID information, keywords etc. The exact structure or content of the metadata will depend on the nature of the data. Standards for metadata creation and management will always be followed.

For example, following the Grant Agreement, metadata of deposited publications must provide information about title, authors, date of publication, publication venue, Horizon Europe funding, grant project name (including acronym and number), licensing terms, persistent identifiers, authors involved in the action and, if possible, their organisations and grant. Also, metadata standards will depend on the repository for the corresponding deposited data. For example, preprints deposited in arXiv will follow the arXiv meta-data format (following The Open Archives Initiative Protocol for Metadata Harvesting).

### **Search keywords in the metadata**

Search keywords are an easy and efficient way to make data and publications more visible. All iDEM partners are committed to using them in the best possible way when either data, software or scientific articles are published.

#### *UPF*

All data published on Zenodo and GitHub will have search keywords, like "text simplification", "iDEM", "lexical simplification", "complex word identification", etc., if applicable.

---

*CAPITO*

It is intended to include labels in the metadata. Relevant search keywords will probably be the topic and sentiment.

*NEXUS*

We will provide keywords for each dataset following the standard vocabulary of social scientific research.

*MAC*

The iDEM OSS is public on Github at [iDEM-eu/.github](https://github.com/iDEM-eu/iDEM-eu.github.io). In-line with Github procedures and to facilitate searching on various keywords the link is expanded with the Github README as follows:

[GitHub - iDEM-eu/.github](https://github.com/iDEM-eu/iDEM-eu.github.io): To make Democratic Deliberation Spaces & participatory processes more accessible and inclusive, the iDEM project is co-creating services to:

- (1) detect difficult to understand texts,
- (2) automatically fix them,
- (3) provide AI tools to generate readable texts.

The project has received funding from the EU Horizon Europe Programme under GA 101132431

*IMPD*

If needed, we might include keywords from the glossary of the project. These may include: Accessibility, people with disabilities, plain language, easy reading, right to participation, cognitive accessibility, inclusion, intellectual disability, participatory democracy

**Metadata: harvested and indexed**

The iDEM consortium will use a Zenodo repository for data publication. Zenodo supports Metadata Harvesting through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), in case UPF will publish relevant metadata in addition to the core datasets.

**Openly available data**

The iDEM consortium is committed to making data openly available whenever it is possible. There are, however, two limitations to this.

- 1) The commercial partners, especially Capito, have data which is crucial to their core business and they therefore have a very legitimate interest in protecting it as Intellectual Property.

2) The ONGs working in the use case pilot studies work with persons who have a right to data protection. This is especially important because we will work with vulnerable populations.

Much of the produced data can therefore not be made openly available, or only so in aggregate, pseudonymised or otherwise processed form. Also, general conclusions and learnings from WP4 will be made public, in scientific publications and the iDEM webpage as well as on the iDEM repository on Zenodo repository. The raw data will, however, only be shared between the partners under the condition of not sharing it with third parties. The following list covers only partners which plan to make data public.

#### *UPF*

All data produced by UPF TALN as part of data collection and data annotation initiatives will be made openly available through Zenodo. In case UPF uses data from partners which are proprietary and cannot be made public, only those parts which do not contain the proprietary information or would make proprietary information recoverable indirectly will be made openly available. Further dissemination of results obtained from the data will be used in scientific peer-reviewed publications.

#### *NEXUS*

We plan to share data produced in T1.2, T4.1 and T4.2. as open-access data.

#### *MAC*

MAC plans to make only OSS available through GitHub.

#### *CIB*

The data produced will be openly available through different mechanisms: There will be open publications, to make sure that it will reach the maximum scope, and the final tool developed from these use cases will also be made available to everyone through an open-source licence.

At the same time, some conferences will be held and recorded, to be finally uploaded online through official channels.

Additionally, the different tools that we will create will be shared freely and openly on the website under the resource's name.

#### *UOL*

The dataset produced through scraping Vikidia and Wikipedia resources will be made openly available via Zenodo and GitHub.

---

## Data availability, accessibility and repository

The consortium has agreed on the use of a Zenodo repository for the publication of data which can be published as open data. The software will be published on GitHub and linked from Zenodo. As for the data which is reserved for the internal use of the consortium, especially the data from the use cases within the iDEM repository, this data will be only accessible for research purposes within the consortium in line with the consents signed, not allowing any third party to conduct research with this data without the user's consent. This last category of data mainly affects data from the use cases.

CIB's data from the use cases' results as well as the use cases' consent forms will be stored in CIB's institutional repository (SharePoint) and also deposited in the iDEM's official document repository to facilitate the internal sharing.

## Methods or software tools to access the data

The data produced in iDEM will be published in standard formats, mostly UTF-8 encoded text and TSV files. For them, no extra software will be needed, although the data will be readable by a wide range of computer programs. Also, customised software will be easily able to read the data. Some data provided by Nexus will be in .docx format, which can be read and processed with Microsoft software and also with a range of open software. Since we will use Zenodo as a data repository, this repository provides all basic services in a robust way, especially web pages and application programming interfaces (APIs).

## Software documentation

The data published by the consortium will be distributed in standard formats that comply with common usage and best practices in the research area of Automatic Text Simplification and Machine Learning in general. Short plain text UTF-8 files (readme files) will be included to explain the data, the size of the data and the structure of the data. Since we agreed to publish all open iDEM software on GitHub, we will follow the repository conventions there, which include the provision of ReadMe files in Markdown format. The MD files are displayed as descriptive text when a repository or a subsection of the repository is accessed on GitHub.

## Relevant software inclusion

All relevant software that is needed to make work published by the iDEM consortium, especially in the case of UPF TALN, UoC and Mac, will be made available through GitHub

---

(<https://github.com/iDEM-eu/github>). In case the software is not produced by project partners, open-access software is publically available from third parties.

### **Deposit of data and associated metadata, documentation and code**

In accordance with what we have specified above, we will publish data and software on Zenodo and GitHub.

### **Arrangements with the identified repository**

Both Zenodo and Github are popular and well-tested open-access repositories. A GitHub repository has been created (<https://github.com/iDEM-eu/github>). A Zenodo repository is still to be created. iDEM has already made contact (2/2024) with the Horizon-Zen project, a Zenodo initiative, to become an early adopter.

### **Access specifications and restrictions on use**

All data that will be published by iDEM will be published under a non-restrictive licence, but the acknowledgment of the source of the data will be required in any resulting publication that uses it. The data that cannot be published because of ID protection and data privacy reasons is exempt from this.

### **Data access committee**

Considering the scope of the project and the small scale of collected and processed linguistic data as well as research experiments, no data access committee is needed.

### **Described conditions for access (i.e. a machine-readable licence)**

Zenodo provides well-described conditions for access, see <http://about.zenodo.org/policies/>.

### **Person identification accessing the data**

iDEM does not plan to make the identification of the person accessing data a precondition for access to data, but we will publish data under non-restrictive licences which require us to cite our work in any resulting publication. In the case access control should become necessary, which we do not foresee, the identity of the person accessing the data will be ascertained through standard protocols and methods whenever applicable.

### **Data and metadata specification**

(vocabularies, standards or methodologies we will follow to make our data interoperable to allow data exchange and re-use within and across disciplines. Community-endorsed interoperability best practices)

#### *UPF*

All data produced by UPF TALN will be published in standard formats: UTF-8, TXT and TSV. No proprietary software will be needed to access, read and interpret the data.

#### *CAPITO*

It is planned to follow the common standards and guidelines of easy language and computational linguistics. These approaches are widely accepted and supported by the scientific community as well as users (in this context, readers).

On the data side, this means using standardised formats like XML and API protocols (most likely not needed) for data exchange. The metadata and documentation are another important standard to ensure the data is used and understood correctly by everyone.

On the linguistic side, the target audience and experts are engaged for feedback on the text simplifications. This ensures quality standards. This is done for example in a focus group or interviews.

#### *NEXUS*

Published data will be interoperable as qualitative data across disciplines insofar the data will be stored in standardised formats. We plan to apply the standards to our data which are described in: [FAIRsharing.org](https://fairsharing.org). Further, Zenodo allows for the linking or referencing of similar data, thereby enhancing interoperability by design.

#### *CIB, PIM, ANFFAS, AAIT, IMPD and BOO*

Standard vocabularies and formats will be used to make the data interoperable. If non-standard or unfamiliar vocabularies (such as specific context-dependent abbreviations or similar) are used, they will be explained. Community-endorsed interoperability best practices will be followed, whenever possible. This includes the use of common (open) formats and standards for data, controlled vocabularies, or avoiding the creation of data that needs proprietary software to be used.

In case it is unavoidable, the use of uncommon or project-specific ontologies or vocabularies, mappings or explanations in terms of standard ontologies or vocabularies will be provided.

#### *UOL*

We will use simple formats for keeping texts for fine-tuning language models in plain text in one document per line. We will re-use annotated data collected by the partners in their respective formats.

### **Documentation provision for data validation analysis and data re-use**

For all data, descriptions of the dataset will be included in the form of text in the repository where it is published and readme files. Whenever possible and needed, we will provide Jupyter notebooks and/or example scripts which exemplify the usage and interpretation of the data.

The publicly available part of WP2, to be produced by CAPITO, consists of the results (simplified text versions). Therefore, no codebooks, readme files or scripts are relevant.

NEXUS will provide additional data files which describe the methodology as .docx files and the structure of codebooks as .csv files. This enables a contextualization of published data.

Software published on GitHub will be documented for use.

IMPD and BOO plan to include the necessary descriptive information to give context to their results, so that they can be interpreted properly (context, objectives, participants' profile).

### **Data licence**

All data produced by iDEM is intended to be re-used by the research community. This will be reflected by the use of non-restrictive licences, which only request citation of the source and acknowledgment of authorship. As stated above, the data will be published in open-access repositories. Nexus plans to use CC-BY 4.0. for their data. MAC will use an Apache 2 licence and a Creative Commons CC-BY licence for public documentation and an Open Data Commons – ODC BY licence for public databases and APIs. UPF TALN will adhere to the mentioned licences. UOL will use a Attribution-NonCommercial licence (by-nc).

The case of data produced in the use cases is more complicated because much of the data will be sensitive, so only non-sensitive data can and will be published. CIB will make non-sensitive data available in the public domain through Zenodo, under CC BY 4.0 licence. The corresponding metadata will be available in the same way under CC0 licence. Articles, conference papers and related content will be available through publication in open-source journals or journals which allow publication of a version on an open access repository. IMPD considers that most data

won't be up for publication as it is sensitive data. If at some stage there is a consideration to publish data, only non-sensitive data will be published.

### **Third parties data use**

All data which is publishable will be published during the project or by the end of the project. As we have stated above there are several cases which make certain data publication impossible. The CAPITO corpus is protected IP and as such not open for public distribution. CAPITO will, however, give full data access to the consortium, meaning the project results (texts plus corresponding simplification) and also to the corpus. It is intended that only the translated texts as well as metadata of the simplification results are publicly accessible.

Much of the data produced in the use cases is sensitive data, which after anonymisation, if possible, will be published in an aggregated form. By sensitive data we understand any data which contains personal information on participants and the points of view they express in a trustworthy environment. Some points of view expressed could be potentially harmful if third parties can attribute these views to them, even if we do not expect this to be an issue in our use cases. If possible, other non-sensitive data or pseudonymized data will be published. Some data will be selectively published within the projects' publications.

#### *NEXUS*

We do not have any particular plans to share data after the end of the project.

#### *PIM*

Results of use cases will be presented to the institutions participating in use case 1: Las Rozas Innova, Las Rozas City Hall and Trébol Association of people with disabilities as a conclusion. They will not be granted full access to data.

### **Data provenance documentation**

In the case of the usage of third-party data, we will cite all sources following scientific standards. Data about institutions and activities will be published if the persons and entities involved do not object to it. By default we will pseudonymised all personal data related to physical persons in order to protect their data - in the case the data is not too sensitive to be withheld from publication entirely. This implies also that we will not be able to credit persons for their data unless they explicitly wish so.

The provenance of data will also be documented in accompanying metadata and README files which were mentioned above. The provision of metadata with respect to the provenance of data will be done in accordance with the GDPR.

The personal data produced by IMPD, CIB, PIM, ANFFAS, AAIT and BOO will be kept in their own systems and will only be shared in anonymized form.

#### *UOL*

As described above, the scrapping of data from Wikipedia and Vikida has been done by UoL. Both sources will be cited and acknowledged, thus the providence of texts from these sources will be identifiable.

### **Data quality assurance processes**

Nexus will implement quality control measures at the following stages: prior to data gathering, while gathering data, before closing the data gathering phase, before data storage and after data storage. The measures will include several steps to check whether the data has been correctly recorded, transcribed, stored, formatted and protected.

The partners working on text simplification (UPF TALN, CAPITO, MAC) will control quality by using standard procedures and be consistent in applying these. After data collection and annotation, we will implement a cycle that checks every part of the dataset before further analysis and publication. There will be a stage for the detection of noise in this data.

The project partners working on the use cases will handle raw data produced within these use cases and see no need to perform explicit internal quality assurance. There will, however, be an external quality assurance. Task 4.6 foresees an evaluation through subcontracted external experts. We will make sure the data security and privacy protocols are also respected by these external collaborators. UPF will serve as contractor, intermediary and responsible in the contact to external experts.

#### *CAPITO*

Since the simplification process itself is very complex, it can be difficult to ensure quality standards. At Capito, we have established our own quality standard to evaluate easy-to-read texts. The translation process and the approach can vary between each translator (or model), therefore it is crucial that there are clear rules to follow. This is important because it is the only way to ensure persistent quality.

This standard consists of three steps:

- 1) All information is translated according to the CAPITO criteria catalogue. The criteria catalogue defines what is understandable for which language level.

- 2) People from the target group check whether the information is comprehensible. If there are ambiguities, the information is revised again.
- 3) Texts that have been translated in accordance with the catalogue of criteria and found to be comprehensible by the test group fulfil the quality standards.

#### **4. Other research outputs**

(software, workflows, protocols, models, etc.)

Apart from data, iDEM will produce research software to carry out experiments which will then be reported in peer-reviewed scientific publications. These experiments will probably produce modifications of existing language models or even produce completely new models. The resulting models will be made public unless licence restrictions on the models that they depend on do not allow this. This type of software includes text simplification systems and difficulty classifiers for lexical items, sentences and texts produced by UPF TALN, UOL.

Software produced by Capito will not be made publically available but project partners will have access to them during the project duration either as software or as services provided through APIs. Software produced by MAC will be available throughout the iDEM GitHub account.

AAIT is considering sharing models of methodology for focus groups, use cases and guidelines for interviews or for the evaluation process with internal staff.

iDEM services, especially the ones provided by MAC, will be provided on a FRAND (Fair, Reasonable, And Non-Discriminatory ) basis. As we have argued above, iDEM will all research outputs produced Findable and Accessible by using well known and heavily used open-access repositories (Zenodo, Github). These repositories already implement many measures that make the published data findable and identifiable. Zenodo automatically associates each upload with a DOI and supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

#### **5. Allocation of resources**

##### **FAIR data making costs or other research outputs**

The cost for making data FAIR will be minimal or null by using open repositories (ZENODO and GitHub) which are already designed to comply with FAIR principles. This is true for data management. However, publication in open-access peer-reviewed journals may have a cost, depending on the journal and type of publication.

UOL

The costs for depositing the dataset with the project, and subsequent resources required to make the dataset publicly available have been included within Work Package WP3 within the project.

### Cost budget

No additional costs for data preservation will be generated because we will use free open repositories and the size of the data is not very big. The costs of preserving the Nexus database will be covered by the project's budget or additional funding. Costs related to publication in peer-reviewed journals will be covered by each partner using their respective budgets in the iDEM project.

MAC

Various as appropriate in the D5.8/9 exploitation plans.

### Data management responsibility

Each partner has to respect the policies set out in this DMP. Datasets have to be created, managed and stored appropriately and in line with applicable legislation by each partner. There will be one assigned representative of each partner who will be responsible for the data management.

For UPF the responsible will be the UPF Data Protection Officer<sup>16</sup>. Responsibles for data management at Nexus are Dr. Thomas Blanchet and Volkan Sayman. Responsible for data management at PIM is Almudena Rascón Alcaina. In Plena Inclusión Madrid, The DPO (Data Protection officer) is Silvia Sánchez<sup>17</sup>.

CIB differentiates between two people responsible for data: The DPO (Data Protection officer) is Jorge Rastrilla Caballero<sup>18</sup> and the Data Administrator of the organisation, Alfonso Peón Nistal<sup>19</sup>.

In AAIT responsible for project data management is Claudia Mazzanti; ActionAid Italia has a Data Protection Manager, her name is Isabella Di Ruggiero<sup>20</sup>; AAIT has an international and national policy regarding the use and storage of data.

---

<sup>16</sup> <https://www.upf.edu/web/cirep/committee>

<https://www.upf.edu/en/web/universitat/-/data-protection-officer>

<sup>17</sup> [silviasanchez@plenamadrid.org](mailto:silviasanchez@plenamadrid.org)

<sup>18</sup> [rgpd@cibervoluntarios.org](mailto:rgpd@cibervoluntarios.org)

<sup>19</sup> [alfonso.peon@cibervoluntarios.org](mailto:alfonso.peon@cibervoluntarios.org)

<sup>20</sup> [isabella.diruggiero@actionaid.org](mailto:isabella.diruggiero@actionaid.org)

## Long term data preservation

The use of standard open-access repositories (GitHub and Zenodo) will be ensured under the reasonable assumption that these repositories will perdure themselves. Further possibilities for long-term preservation will be discussed and agreed upon within the whole consortium. More details will be documented in D5.8 iDEM Initial Sustainable Exploitation, Innovation and IPR Plans, M12 and D5.9 iDEM Final Sustainable Exploitation and IPR Plans, M36.

## 6. Data security

### Provisions

#### *UPF*

Data collected by UPF TALN will be stored on shared Google drives, in accordance with university regulations, and on UPF Secure storage. Both storage systems implement regular back-ups. The data collected by UPF will not be sensitive, which does not make special procedures necessary.

#### *NEXUS*

Data collected within the project will be digitalized and stored on a cloud service (nextcloud) operating according to the rules of the GDPR. Data recovery will be ensured by a third-party service provider managing the cloud service and complying with the rules of the GDPR. No sensitive data will be transferred.

#### *PIM*

The data is stored on a local server that is password protected. Daily backups are automated and encrypted, located in the office and on a rented server within a data centre. The IT administrator keeps the system and backup copies updated, to guarantee the security and integrity of the information, protecting it against threats of viruses, malware, ransomware, etc.

#### *MAC*

Users' own data will be stored on their own device. Project files and approved users' emails will be stored in the secure iDEM Google Drive.

#### *ANFFAS*

Data is stored on a shared drive that is password protected. Backups are made regularly. Maintenance, setup and backups of computers and shared drives is done by an information company. The system in use is periodically updated to guarantee the security of the information

and data stored, as well as protection from viruses and malware that circulate online and which can also arrive via email. .

#### *CIB*

For data deposited in public trusted repositories, security will be provided by the entity responsible for the management of the repository. For data deposited in local repositories of each partner, security provisions will be determined and provided by the corresponding institution. They usually include frequent backups, storage of copies on local drives, etc.

#### *AAIT*

AAIT has an IT office which, in agreement with the federation, is responsible for applying data protection solutions. The system in use is periodically updated to guarantee the security of the information and data stored, as well as protection from viruses and malware that circulate online and which can also arrive via email.

#### *BOO and IMPD*

The City Council of Barcelona has set up a reliable standard security systems and proceedings which covers BOO and IMPD.

#### *UOL*

UoL will not manage sensitive data and will rely on GitHub and Zenodo for all aspects of secure public data storage.

### **Safely stored data in trusted repositories for long term preservation and curation**

As for the data that will be destined for open access, we will trust Zenodo and GitHub repositories. They provide safe and durable storage, as well as comply with the FAIR principles.

Data which is restricted to internal use will be stored on internal servers and the storage system of each partner. This will be necessary for AAIT, CIB, ANFAAS, IMPD, PIM and BOO. The documents will be stored there as long as needed. Access will be restricted to persons who are working within the project.

## 7. Ethics

### **Ethics or legal issues that can have an impact on data sharing**

The project partners working on text simplification systems and the technical implementation (UPF TALN, CAPITO, MAC, UOL) of these see no ethical risks which are present in these technical tasks. Third party data have been derived from permissible licences. Attention has to be paid to translation/simplification bias which might stem from the use of Large Language Models. Nexus will share only data which will not contain any relations or references to real people or other sensitive data.

Regarding the personal data collected in use case pilots, the project partners which handle this data (PIM, ANFFAS, CIB, AAIT, IMPD, BOO), adhere to the principles of informed consent, ensuring that participants are adequately informed about the purpose of data collection, storage, and sharing. Our data-sharing practices are aligned with European data protection regulations, such as the General Data Protection Regulation (GDPR). We ensure that all data processing activities, including sharing, adhere to these legal frameworks.

The iDEM partners involved in data collection have sought (e.g. NEXUS) and will seek (e.g. PIM, CIB) ethical approval from the CiREP office at UPF regarding research ethical protocols which will be followed during focus groups and interviews data collection and data processing.

## 8. Other issues

### **Use of other national/funder/sectorial/departmental procedures for data management**

All iDEM project partners will use the protocols foreseen by the project.

We will ensure that all procedures comply with ISO 27001<sup>21</sup> and ISO 27002<sup>22</sup>, public administration guides, the Esquema Nacional de Seguridad (for Spain) and all agreements signed by the consortium.

---

<sup>21</sup> Information Technologies – Security Techniques – Information Security Management Systems

<sup>22</sup> Information security, cybersecurity and privacy protection — Information security controls