D1.1
Project DMP and ORDP

Authors: Jennifer Edmond, Erzsébet Tóth-Czifra, Maciej Eder, Caroline Odebrecht, Silvie Cinkova, Ingo Börner, Julie Birkholz, Sally Chambers, Christof Schöch, Matej Durco, Frank Fischer, Justin Tonra

Date: August 31, 2021

Project Acronym: CLS INFRA
Project Full Title: Computational Literary Studies Infrastructure
Grant Agreement No.: 101004984

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Deliverable/Document Information

Deliverable No.: D1.1
Deliverable Title: Project DMP and ORDP
Authors: Jennifer Edmond, Erzsébet Tóth-Czifra, Maciej Eder, Caroline Odebrecht, Silvie Cinkova, Ingo Börner, Julie Birkholz, Sally Chambers, Christof Schöch, Matej Durco, Frank Fischer, Justin Tonra
Dissemination Level: Public

Document History

Version/Date	Changes/Approval	Author/Approved by
V0.1 10/06/2021	Initial draft	Erzsébet Tóth-Czifra, Jennifer Edmond
V0.2 14/06/2021	Input from the whole consortium	All authors
V0.3 19/07/2021	Extended draft including software components	Erzsébet Tóth-Czifra
V0.4 19/08/2021	Extended draft reviewed by project partners	All authors
V1.0 31/08/2021	Final version of the M6 DMP	All authors

Project summary

The digital age offers challenges and opportunities for completing research on Europe's multilingual and interconnected literary heritage. Even though many resources are currently available in digital libraries, a lack of standardisation hinders their access and reuse. The EU-funded CLS INFRA project will help build the shared and sustainable infrastructure needed to undertake literary studies in the digital age. The project will align these diverse resources with each other, with the tools needed to interrogate them, and with a widened base of users. The resulting improvements will benefit researchers by bridging gaps between greater and lesser-resourced communities in computational literary studies and beyond, ultimately offering opportunities to create new research and insight into our shared and varied European cultural heritage.

Executive summary

This deliverable presents the first version of Data Management Plan for the Horizon 2020 project CLS INFRA. To maximise the potential of the present document as a living and functional project management tool, the team decided to accommodate to this present DMP all sorts of digital scholarly objects that qualify as data or software and that are deemed worthy to be captured and preserved for reuse. Following the Horizon 2020 Data Management Plan template and guidelines, the document first defines the scope and selection criteria for project outputs (1.1.) and provides a structured overview of data and software collected, created, and (re)used for and by the CLS INFRA project. The second, FAIR data chapter describes the provisions made or envisioned to accommodate the FAIR principles and thereby future usage of CLS INFRA data resources. The third chapter provides a similar FAIR assessment of software resources. The fourth chapter describes roles, responsibilities and practical commitments and safeguards that are in place for the successful implementation of the DMP. The fifth chapter covers data security aspects. In the final chapter, we accommodated discussions around ethical aspects such as safeguards for personal data handling and safeguards for adherence to intellectual property rights are discussed. While preparing the DMP, special attention has been given to domain and disciplinary specificities. The CLS INFRA Data Management Plan is a living document and shall be updated continuously throughout the project in line with the new information gathered via conducting the project activities.

1.Overview

1.1. Scope

Although the concept of a Data Management Plan comes with a primary focus on data and metadata by definition, the CLS INFRA team is well aware that it cannot be comprehensive without a proper documentation and management plan of the software environment in which data is shaped, run and contextualised. Therefore, to maximise the potential of the present document as a living and functional project management tool, **the team decided to accommodate to this present DMP all sorts of digital scholarly objects that qualify as data or software** and that are deemed worthy to be captured and preserved for reuse. **An inventory of both the digital scholarly object types that qualify as research data and research software can be found in the overview charts below.**

Scoping limitations of the current DMP starts with drawing a line between software vs. services. At M6, we find it too early to initiate discussions about the sustainability of services to be developed within the framework of the project, but it is expected to accommodate these crucial discussions, from the overall project's point of view, in a later version of the present DMP. Where relevant however, conditions and safeguards for hosting flagship interfaces, such as the DraCor platform or the project website, will be touched upon.

The overview of data types the project is operating along reflects the scoping decisions.

OVERVIEW OF WORK PACKAGES:

WPs coming with admin/support data (easy to deal with in the DMP)	1,2,4,9 (but: WP1 is in charge of RDM)
Research data heavy WPs	3,5,6,7
Not data heavy ones (but of course carrying data aspects too)	8
Research software heavy WPs	7,8

1.2. Overview of the data types within the project

Grey rows indicate proposed closed access

Data type		WP/Task	Size, volume, format	Location (and access)
Project administration	Project administration data, including personal data (textual, tabular)	WP1, WP2, WP9	.PDF,.docx, .xlsx, Google docs	Nextcloud (https://cloud.clsinfra.io), Mattermost (https://mattermost.clsinfra.io/), central GitLab instance (https://gitlab.clsinfra.io/)[1] Project Level Access Only
	Event management data (including personal data)* *Likely subject to GDPR regulations, see note under chapter 5 below.*	WP2, T4.2.	.PDF, .docx, .xlsx, Google docs	Nextcloud (https://cloud.clsinfra.io) Project Level Access Only
	Platform usage data (user statistics of the project website, blog(s), social media platforms, webinar platforms)	WP2, T4.2.	Platform dependent statistics	Embedded in the respective platforms (Twitter, Big Blue Button) Project website hosted by IJP PAN Project Level Access Only

[1] Commitments towards the open research culture and the open source movement have been the primary motivations for selecting these services.

	Video recordings of internal meetings* *Likely subject to GDPR regulations, see note under chapter 5 below.*	WP2	.MP4 (approx. 400 MB each)	Nextcloud, Big Blue Button (https://bbb.clsinfra.io/) hosted by IJP PAN Project Level Access Only
Training and dissemination	Video recordings of public meetings, training events)* *Likely subject to GDPR regulations, see note under chapter 5 below.*	T4.2	.MP4 (approx. 400 MB each)	YouTube, embedded in DARIAH Campus CC-BY 4.0
	Content delivered via the website (news items, event reports, TNA reports, blog, etc.)	WP2	.HTML plus related files: .css, .xslt, .js, .es as appropriate	Website (see above) Archived version to be made available at project end - (CC-BY 4.0)
	Training materials	WP4	.md, .mp4, .HTML .PDF plus related files: .css, .xslt, .js, .es as appropriate	DARIAH Campus Open Access (CC-BY 4.0)
	Deliverable reports (textual)	All WPs	.PDF	Upload to the Horizon 2020 interface plus Zenodo Open Access (CC-BY 4.0)

	Research papers resulting from the project (including survey, methodology and data papers)	WP3, WP5.1 and 5.2, WP6, WP7	.PDF, .TEI-XML, .EPUB .JSON	Hosted by the publisher plus the AAM[2] deposited at Zenodo Open Access (CC-BY 4.0)
Primary sources **(digitized literary corpora)**	Primary resources (digitized literary text corpora) - public domain: DraCor[3] ELTEC[4] Others as per Tasks 5.3/6.1 TBD	WP1, WP3, WP5, WP6, WP7	XML-based, JSON-based, cvs-based data and RDF realisations	Remain to be hosted at original hosting institutions (University of Potsdam, University of Trier) Open Access (Public Domain, CC-BY 4.0) - as much as the licenses of the primary resources allow.
	Primary resources (digitized literary text corpora) - under copyright - TBD	T7.4; T7.5.		Remain to be hosted at original hosting institutions (to name) In copyright, see access scenarios specified in GA p. 103/55.
Landscaping and survey data	Registry of data formats	T6.1, T6.2.	.CSV, other possible formats to be specified	Zenodo, see Deliverables under Dissemination Open Access (CC-BY 4.0)

[2] Author Accepted Manuscript.

[3] https://dracor.org/

[4] https://www.distant-reading.net/eltec.

	Registry of TEI[5] annotation frameworks	T 6.1.	.TEI-XML	Zenodo, see Deliverables under Dissemination Open Access (CC-BY 4.0)
	CLS resource catalogue: An inventory of existing European literary corpora; registry of TEI annotation frameworks	T5.1; T6.1.	.PDF other possible formats to be specified	Created on Nextcloud, published on Zenodo and elsewhere (refined/specified) as an OA-Article Open Access (CC-BY 4.0)
	Models and controlled vocabularies	T5.1, T5.2, T6.2.	UML Meta model for corpus metadata[6], RDF, SKOS	These resources will be stored together with tools for which it was built. For instance, a specialized languages model for UDPipe will be stored at the LINDAT-CLARIAH-CZ. Secondary locations: Zenodo, OEAW (Vocabs service (https://vocabs.dariah.eu/). All secondary storage locations will sync with the primary location. Open Access

[5] Text Encoding Initiative.

[6] http://www.db.informatik.uni-bremen.de/teaching/courses/ss2020_eis/week10/H-mm.pdf

	Tool kit report	T5.3	.PDF	Zenodo, (CC-BY 4.0)
	Surveys, interviews (raw data)* *Likely subject to GDPR regulations, under chapter 5 below.* *Likely subject to ethical review at WP Lead or coordinator institution*	T3.2.; T3.5.	.MP3 .MP4 files , transcriptions in .word .PDF	Nextcloud Project Level Access Only
	Survey, interview, anonymized (pseudonymized) derived data and metadata (including documentation, codebook etc.)* *Likely subject to GDPR regulations, under chapter 5 below.* *Likely subject to ethical review at WP Lead or coordinator institution*	T3.2.; T3.5.	.MaxQDA or equivalent .SPSS (.SAV) or data (.CSV) + setup (.TXT) Or In Vivo	Open Access (CC-BY 4.0)
Derived data/enrichments	CLS specific metadata and structural data	T6.2.,T5.1	.JSON-LD, Linked Data (RDF), OAI-PMH, YAML, UML	Programmable Corpora prototype platforms and Zenodo

	TEI annotations	T5.3, T5.3. T8.3.	.XML, .JSON, other possible formats to be specified	Stored or linked together with the original texts (primary resources). The primary data hub will be Zenodo, additional deposits to LINDAT-CLARIAH-CZ and TextGrid are also likely, tbd.
	Annotations and annotation frameworks other than TEI; their conversion to W3C annotation standards; extended transformation matrix	T6.3.		Zenodo
	Data from the TEI -->LOD[7] conversion toolbox	T6.3.	LOD (and other data formats)	
	Training corpora for NLP pipelines	WP8	.TXT, python	LINDAT/CLARIAH-CZ, and/or training datasets can be made available on ugent.be domain; Zenodo

[7] Linked Open Data.

	Workflows (e.g. workflows for building new NLP tools for lesser resourced languages.)	T8.2.		GitLab Publication SSH Open Marketplace[8]
	Generic polarity lexicons in at least four languages.	T8.6.	.TXT, python	Datasets can be made available on ugent.be domain as well as on Zenodo

[8] The SSH open Marketplace is a discovery and virtual research environment, bringing together data, tools, publications and workflows relevant to Social Sciences and Humanities research. https://marketplace.sshopencloud.eu/

1.3. Overview of the software types within the project

Software type or underlying code behind the following tools/toolchains	**WP/Task**	**Location (and access)** **(Which GitHub/GitLab repo)**
R libraries for the analysis of the multi-lingual corpora stored in DraCor	T7.2.	Central GitLab instance(https://gitlab.clsinfra.io/) and Zenodo if not specified otherwise. The already existing DraCor GitHub repositories https://github.com/dracor-org/ could be mirrored to CLS INFRA's central GitHub as well as deposited on Zenodo. The r-library will also be released on CRAN to allow for easy installation
Python libraries	T7.2.	Central GitLab instance (https://gitlab.clsinfra.io/) and Zenodo if not specified otherwise The already existing DraCor GitHub repositories https://github.com/dracor-org/ could be mirrored to CLS INFRA's central GitHub as well as deposited on Zenodo. https://github.com/dr

		acor-org/pydracor is the repository of the Python-Package PyDraCor; for easy installation and usage in Python, it will also be accessible via PyPi.
TXM tools[9]	T7.4.	Central GitLab instance (https://gitlab.clsinfra.io/) and Zenodo if not specified otherwise.
UDPipe pipeline (NLP tools developed by WP 8; LeTs toolkit, LINDAT/CLARIAH-CZ toolchain, tokenization, segmentation, morphology, POS tagging, lemmatization and parsing tools)	WP8	LINDAT/CLARIAH-CZ, central GitLab instance (https://gitlab.clsinfra.io/) and Zenodo if not specified otherwise.
Transformation toolbox	T6.2.	Central GitLab instance (https://gitlab.clsinfra.io/) if it is meant to be available to users right from the start.
Jupyter notebooks	T4.2. (mentioned in a training context, to add other tasks too)	Central GitLab instance (https://gitlab.clsinfra.io/) and Zenodo if not specified otherwise.

[9] TXM is an open source research software that offers a wide range of textometric queries. For more information, visit: http://textometrie.ens-lyon.fr/?lang=en

2. FAIR data

This section mainly describes the provisions made or envisioned to accommodate the FAIR principles and thereby future usage of CLS INFRA data resources.

2.1. FAIR by design

Exploring and implementing FAIR data curation practices within Computational Literary Studies (CLS) and enabling the realisation of a FAIR data environment are among the key aims of the project (see detailed in p. 69 of the GA). **WPs 5, 6 and 7 will facilitate** the alignment and transformation of literary corpora to become findable, accessible, interoperable and reusable (**FAIR)** within the currently highly scattered field of CLS that operates along legal, technological and conceptual challenges when it comes to connecting and reusing digitized literary corpora and research data derived from them. In addition to the project outputs and the implementation of the resulting solutions across the field of CLS, **this work will also directly feed into the future iterations of this DMP.**

2.2. Findable

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

What naming conventions do you follow?

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

Do you provide clear version numbers?

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Reflecting and sensibly serving domain specific needs where the continuum of data curation is essentially shared across cultural heritage institutions, research teams and computer scientists, the CLS INFRA project makes serious commitments to making literary corpora and their enrichment available in a democratic, decentralized FAIR manner, without creating the need for

a large central organisation or central data hub. **Instead, it leaves data hosting and ownership in the hands of its owners, thereby avoiding the challenges both of technically hosting large volumes of data, but also of negotiating agreements for data transfer, migrating data to a rigid common format etc.** Instead, the Programmable Corpora approach, the data landscape and the use cases provided by CLS INFRA empowers data owners to align their data with the others in the ecosystem, thereby making their own native data FAIR and active within the ecosystem, which also facilitates its use with analysis tools similarly optimised as well as its reuse by others active in the system. To embrace the plurality of ownership but still reduce fragmentation of resources, digital resources from different national, institutional contexts will be connected by the data landscape review and will be made discoverable through an API layer. This will provide tailor-made access of the different user groups to the literary data in the ecosystem via programmable corpora prototypes (e.g. DraCor), richly documented, Open Access use cases as well as via a Zenodo collection of resources.

2.2.1. Data sources the project builds on

Primary resources reused in the project: establishing reuse rights to these resources as well as technical interoperability issues in making them readily available for computational analysis within the project are discussed under '2.5. Reusable'

2.2.2. Data repositories, hosting services

Zenodo is identified as the primary repository hub for the project in the form of a dedicated collection to CLS INFRA. Additional, rich documentation will be provided by adding README files to the deposits. In addition, where relevant, data will also be deposited by subject-specific repositories such as TextGrid[10], ARCHE[11] or the LINDAT/CLARIAH-CZ repository[12]. The granularity in the PID policy of the project (making decisions on the units of deposited records that belong under the same PID) will be decided at a later phase of the project.

Alongside with sharing data and software, from a project sustainability point of view it is essentially important to guarantee secure hosting of the platforms and interfaces associated with the project. IJP PAN will be responsible for hosting the project website and the University of Potsdam will be responsible for hosting the DraCor platform (https://dracor.org/) at least for the duration of the CLS INFRA project, with sustainability considerations in mind regarding the digital project's afterlife. In this respect, the code is already available under an open licence and hosted on Github (https://github.com/dracor-org, especially the core API: https://github.com/dracor-org/dracor-api). In T7.3 Technical stability of APIs for versioning and reproducibility, we plan to develop means of

[10] https://textgridrep.org/
[11] https://arche.acdh.oeaw.ac.at/browser/about-service
[12] https://lindat.mff.cuni.cz/repository/xmlui/

quickly setting up the platform (API + necessary services, e.g. metrics service for network metrics) as docker containers (bundled with a docker-compose), which will allow users to set up and run their "own" (DraCor-)Platform. Such sustainability aspects will be discussed at a later iteration of the present DMP.

2.2.3. Custom metadata

Metadata will be collected from the outputs of WP5, 6, 7 and 8. A CLS-specific metadata scheme or schemes will be identified **based on the CLS resource catalogue** to be developed by T5.1. and T6.1. At a later phase of the project, metadata will be included in a designated platform from which it will be made available under an interoperable format, following international standards (such as JSON-LD, Linked Data, OAI-PMH) via APIs. The metadata are formally modelled with the UML standard and be realized with RDF. Both outputs will be Open Access published (CC-BY4.0) and archived.

2.2.4. File naming conventions

At this point in the project lifecycle, there is **no strict naming policy in place**. Work packages and within work packages, task leaders will decide on the optimum level of granularity in this respect and will make sure that creators, curators harmonise their files along a coherent, shared policy. Relevant guidelines will be shared and coordinated across WPs via Nextcloud. Further, file naming conventions will be added to READMEs to make sure future users will understand the data/file naming.

2.2.5. Versioning policy

Currently, resources are controlled by Git-versioning. Setting up and implementing new concepts for the referencing of evolving data sets as well **as a standardised and reliable versioning routine will be delivered by T7.3** (D7.3 *Report on versioning requirements of APIs and corpora within CLS* [M36]).

2.2.6. Training materials will be findable on DARIAH Campus

Training materials resulting from WP 4 **will be hosted on the DARIAH Campus** training discovery portal.[13] This however does not yet include solutions for long-term archiving and sustainable referencing via PIDs. Therefore, **it is expected that training materials will first be deposited in one of the trusted repositories** identified above and from there they will be linked to the discovery portal.

2.2.7. EOSC integration

As part of sustainability discussions towards the end of the project lifecycle, a detailed strategy of making CLS INFRA resources findable through the European Open Science Cloud (EOSC)

[13] https://campus.dariah.eu/

catalogue will be established. It is expected that both the EOSC onboarding workflows and the sustainability plan of CLS INFRA will go through considerable changes until then.

The primary means of EOSC integration will be realized through making a selected set of CLS INFRA outputs (workflows, data sets, publications, API libraries) available in **the SSH Open Marketplace.**

2.3. Accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

How will the data be made accessible (e.g. by deposition in a repository)? What methods or software tools are needed to access the data?

Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Have you explored appropriate arrangements with the identified repository?

If there are restrictions on use, how will access be provided?

Is there a need for a data access committee?

Are there well described conditions for access (i.e. a machine readable license)? How will the identity of the person accessing the data be ascertained?

CLS INFRA makes a **strong commitment to improve access to digitized literary corpora** and their enrichments and to connect the pieces of this currently scattered data landscape via an API layer detailed under '2.1. FAIR by design'. Data outputs of the CLS INFRA project will be made available in ways outlined in '2.2. Findable'. The project aims to release all created resources following internal Open Access and open data policies and implement a licensing policy where CC-BY 4.0 or its equivalent in the Apache framework, Apache 2 for software outputs (as discussed below at the 3. FAIR software' chapter) are the default options.

2.3.1. Open Access, open data policy

Papers and reports resulting from the project **will be published Open Access and a copy of them will be deposited in the Zenodo collection of the project.** In the case of resulting data sets and enriched corpora, these will also be made available Open Access unless legal restrictions of third-party materials make it impossible. **Solutions to clear reuse rights and facilitate access to copyright materials will be delivered by T5.2. And T7.5.** (see detailed under '2.5. Reusable'). In addition, results of 'T5.3 *Policies for sharing data as an institution or an individual*' will directly feed into a more detailed data access policy shared within the project.

2.3.2. Well-documented access conditions to CLS data via APIs

A set of **APIs** developed in WP7 (especially within T7.3.) **will guarantee stable, clear and federated access protocols** to CLS data taking into consideration the diverse needs of different target groups and application scenarios.

2.3.3. Making accessible a subset of CLS data via SPARQL endpoints

API development (T7.1.) will also entail providing access to metadata and texts in different formats and making them available **through a SPARQL endpoint for Linked Open Data (LOD) queries.**

2.4. Interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

CLS INFRA's serious commitment to enable FAIR in the curation and enrichment of literary data also manifested in **standardization efforts** so that research data and analysis tools are accessible and connectible to a rich pool of reusers arriving from different languages and disciplinary communities and with varying skill sets.

WP7 will mitigate both the difficulties of negotiating access to and hosting for vast amounts of research data. In addition, WPs 5 and 6 will be working to build the networks, relationships and social instruments (formal agreements, best practice documentation, etc.) needed to underpin this approach. **A final decision on the most applicable format will be made at month 24 of the project (Milestone 3).**

2.4.1. Technical and social interoperability

Standardization efforts will be centered around TEI annotation standards, serving as a dominant but internally diverse community standard of literary text annotation. As a next step, TEI annotation frameworks (together with their alternatives) will be transformed to W3C annotation standards in extended transformation matrix in order to connect them with the Linked Open Data Cloud and open them up with analysis via semantic web technologies.

Broken down to specific tasks:

D6.1 will provide an inventory of existing data sources and formats (TEI frameworks and annotation frameworks other than TEI) supported/required by tools and services and will also provide a transformation matrix capturing available and needed transformation paths (M14, in alignment with D5.1)

D6.2 will provide a transformation toolbox & ingest and processing workflow (including the development of CLS-specific metadata standards) (M48)

D6.3 will provide an extended transformation matrix / alternative formats (M28)

Further, training efforts of WP4 will facilitate the widespread, cross-community adoption of these standards.

2.4.2. Legal interoperability

Ethical and legal issues associated with CLS data are discussed in Chapter 5.

2.5. Reusable

How will the data be licensed to permit the widest re-use possible?

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

How long is it intended that the data remains re-usable? Are data quality assurance processes described?

Within the CLS INFRA project, reusability is a bidirectional concept.

On the one hand, the project heavily relies on reusing already digitized literary corpora (upstream reuse). These data are embedded in different regional, national, institutional and technical silos

and therefore making them readily available for reuse within the project is a cornerstone RDM challenge[14] of the project that needs to be discussed in the DMP.

On the other hand, a clear licensing policy and commitments to rich documentation is necessary to enable the widest possible reuse of project outputs (downstream reuse).

2.5.1. Upstream reuse - obtaining reuse rights

Relevant corpora that are subject to copyright restrictions have yet to be identified in WP5 (T5.1 *Data landscape overview*). Copyrighted sources will be held at the relevant GLAM institutions. One possible recommendation for this plan is to use the Cultural Heritage Data Reuse Charter (https://datacharter.hypotheses.org/charter-templates) for obtaining reuse rights and set up a partnership with the partner institutions. The project will research possibilities of generating derived text formats (cf. Schöch *et al.*[15]) that comply to the EU regulations on text and data mining cf. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, http://data.europa.eu/eli/dir/2019/790/oj). The project might store and provide derivate formats that are in line with EU copyright regulations only.

2.5.2. Downstream reuse

To facilitate downstream reuse, data will be stored or persistently linked together with tools for which it was built. For instance, a specialized languages model for UDPipe will be made available in the same location as all other models, that is, a LINDAT-CLARIAH-CZ and the place where the tool developers store the models currently in use when you call the service by its API. All secondary storage locations will be synced with the primary location, by automatic updates of metadata items only. Further, reuse is also supported through metadata and rich documentation (see the README files as well as the use cases and workflow publications as project outputs mentioned above).

2.5.2.1. Clear licensing policy

As was already mentioned above, the CLS INFRA project is committed **to make available datasets as project outputs under a CC-BY 4.0 license by default**. Where it is not possible, solutions to clear reuse rights and facilitate access to in copyright materials will be delivered by T5.1 and T7.5. In addition, **metadata will be made available under CC0 license.** This licensing policy will be implemented in a way that is easily readable both for humans and machines.

[14] https://hal.archives-ouvertes.fr/hal-02961317

[15] Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, Jörg Röpke: Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel 2020. text/html Format. DOI: 10.17175/2020_006

2.5.2.2. Training and support for adoption of computational methods and data reuse

To facilitate the reuse of project outputs (data sets, API libraries, analysis tools) for a wide range of scholarly communities coming with varying skill sets and ease the adoption of computational analysis methods, **WP 4 will produce a rich pool of training materials.**

2.5.2.3. Sustainability plan for the project outputs

Sustainability of the project outputs **will be ensured by DARIAH ERIC**, as full partner of the project, who has committed to use its extensive network and technical acumen not only to maintain the tools and services developed within the project. **More specifically, T1.6 *Strategic Roadmap for future research infrastructure and innovation*** (DARIAH, IJP PAN) is in charge of developing a clear notion of sustainability and preservation of a selection of the project outputs by M48.

3. FAIR software

The CLS INFRA project is **strongly committed towards the principles of open source software development** and, as much as dependencies allow, will release software outputs as open source, openly licensed (research) artifacts (or objects) with rich documentation. Further, to ensure sustainable operation and maintenance of the CLS INFRA components efforts will be dedicated to align to and comply with the following sustainable and FAIR software guidelines:

- Carsten Thiel, Michelle Weidling, Yoann Moranville, 2018. The EURISE Network Technical Reference. https://technical-reference.readthedocs.io/en/latest/

- Jiménez RC, Kuzak M, Alhamdoosh M *et al.* Four simple recommendations to encourage best practices in research software [version 1; referees:3]. F1000Research 2017, 6:876 (doi: 10.12688/f1000research.11407.1)

- Hong, N. P. C., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., & Honeyman, T. (2021). FAIR Principles for Research Software (FAIR4RS Principles). *Research Data Alliance*. DOI: 10.15497/RDA00065

- Lamprecht, Anna-Lena et al. 'Towards FAIR Principles for Research Software'. 1 Jan. 2020: 37 – 59. DOI: 10.3233/DS-190026

Their implementation, broken down to the FAIR principles, is detailed below. An overview of the key software outputs of the CLS INFRA project can be found under '1.3. Overview of the software types within the project'. Within the project, WP 7 and WP 8 will be primarily involved in sustainable software development and sharing.

3.1. Findable

3.1.1. Software publication in open repositories

As a primary findability criterion, **source codes will be hosted in public version controlled repositories** (GitLab: https://gitlab.clsinfra.io/) Further, following community standards, **long-term availability** of the software outputs will be ensured in 2 ways:

1) Following community standards, relevant GitHub and/or GitLab repositories will be connected to the project's **Zenodo collection** to deposit software releases from GitHub to Zenodo, together with the provision of appropriate metadata, providing contextual information about the software. Zenodo mints DOIs for each released version of the software, and also creates a concept DOI which refers to all versions of a given software. This way, a PID will be assigned to all versions and specific deployments of source codes.

2) **Software Heritage**[16] is another publicly funded, European open archive that harvests all public GitHub repositories to ensure long-term availability, traceability and citability of source codes of research software.

3.1.2. Sharing software outputs in discovery environments and registries

To further increase findability of software outputs of CLS INFRA, GitHub repositories (and their long-term deposits) will be linked 1) to scholarly publications resulting from the project 2) to the DraCor platform 3) to the SSH Open Marketplace as well as to other domain-specific code libraries and registries like the CLARIN Resource Families. Furthermore, as stated under '2.2.7. EOSC integration', possibilities will be explored to integrate the key outputs of the project to the EOSC service catalogue.

3.1.3. Rich metadata and a clear indication of dependencies

For documentation purposes, a **README file will be provided to each software component** including information about: the project, where to find a specific version of the software, how to cite it, who are the authors/contributors, what are the inputs and outputs, and what are dependencies plus information about provenance elements, with, again, special focus on dependencies (data, code, environment and workflows in which the software is operating). The documentation will provide information on how to install, run and use software, dependencies will be clearly stated.

In terms of documentation, the EURISE guidelines will be followed: https://technical-reference.readthedocs.io/en/latest/developer-guidelines/03-documentation.html

[16] https://www.softwareheritage.org/mission/approach/

3.1.4. Clear versioning policy

For clear versioning, the project is using Git, currently the most popular technology for software source code version control.

Besides, T7.3. Is tasked with developing a technical stability of APIs for versioning and reproducibility.

3.2. Accessible

Access protocols will be specified and **managed through the GitLab (and for legacy software, in some cases, GitHub) repositories** as well as by Zenodo and Software Heritage.

Stable access to documentation and metadata will be provided by/through PIDs, discussed above, under ‘Findable’.

Access conditions will be specified by clearly stated and machine actionable licenses specified below under ‘Reusable’.

3.3. Interoperable

An important aim of CLS INFRA is to provide solutions for computational text analysis that **will not lock the user into a specific software product or data format** (p. 23 of the DoA part B of the Grant Agreement). The **documentation** accompanying the software outputs will provide details on how the software interoperates with other digital objects: input and output data types and formats, communication interfaces (through the APIs to be developed by the project), and/or deployment options.

The research software outputs of the project will be hosted in a distributed manner, but will also be published durably **as an interconnected package**, either through rich linking, data deposit, or a next-generation rich open monograph publication (p. 25 of the of the DoA part B of the Grant Agreement)

3.4. Reusable

3.4.1. Licensing

The legal framework associated with software outputs is a crucial element of reusability as it determines how software can be built, modified, used, accessed and distributed. For software outputs, the project will apply an **APACHE 2.0 or 3.0 license by default.** This license must also be compatible with the requirements of the licenses of the software's dependencies so that the software can be legally combined.

Software documentation and **metadata will be released under CC0 license by default.**

On both levels, special care will be taken towards the **machine readability** of license information.

3.4.2. Software citation

To ensure appropriate crediting and acknowledgement, CLS INFRA will follow **software citation good practices** (such as this one: https://www.software.ac.uk/how-cite-software) and will use **machine-readable citation formats**, such as a .CFF file (https://citation-file-format.github.io/).

3.4.3. Rich provenance information

In the context of 'Reusable', it is worth mentioning that similarly to data resources, reusability also includes building on already established software components and data (corpora) that are associated with them. Provenance information on the documentation of CLS INFRA software outputs will include **information on how the software has been compiled and which dependencies it incorporates, including machine-readable references to other software components.**

4. Allocation of resources

What are the costs for making data FAIR in your project?

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project?

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

4.1. DMP development timeline and responsibilities to update it

The CLS INFRA project aims to use this DMP as a project management tool that facilitates developing a common understanding and shared data management solutions across the project participants. As such, **all WPs of the project contribute to it via meetings with the WP leaders and written consultations carried out in an iterative fashion**. **T1.5. (DARIAH, IJP PAN) will be in charge of coordinating and leading regular updates of this DMP** which will evolve into an Open Data Research Plan by M48. Four versions of the DMP are planned **at M6, M18, and M30 and M48.**

4.2. Resources allocated in project budget

Costs for making the data associated with the CLS INFRA project FAIR are covered by the grant. **A total of 6 PMs is reserved** for dealing with all issues around data management. During the project lifetime, DARIAH-EU will guide the process that ultimately leads to the deposit of relevant components of the project in data repositories.

4.3. Initial sustainability discussions

See discussed above, under ‘2.5.2.’ Downstream reuse.

5. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Is the data safely stored in certified repositories for long term preservation and curation?

Task leaders are expected to have their back-up and storage policies in place **followed by their own, local policies and back-up protocols using institutional cloud storage solutions** (handled by their IT departments) and following the rules of GDPR. In case a partner institution does not have sufficient infrastructural components in place for secure storage, it is possible to coordinate with another project partner.

As discussed above under '2.2. Findable' and '3.1. Findable', **data and software outputs will be deposited and made openly available on the long term in trusted repositories**.

6. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

The Ethical aspects of the CLS INFRA project are described in detail in the Grant Agreement.

All project partners agree to apply the ethical standards and guidelines of Horizon 2020, as well as professional and international standards and relevant national, EU and international legislation. With respect to personal data, all partners will comply with the General Data Protection Regulation (GDPR).

6.1. Safeguards for personal data handling

In practice, four concrete cases have been identified which involve personal data protection, three of which are coming from the context of project administration while one qualifies as research involving human participants. These are:

- Personal data concerning the project participants (managed by WP1)
- Personal data concerning TNA beneficiaries (managed by WP9, WP1)
- Personal data collected from attendees of public events (online or face to face, managed by WP2 and WP4)
- Survey data collected and managed by T 3.2. and T 3.5.

In principle, the data collection is not aiming at collecting sensitive data, however personal data may be collected. This data collection process will strictly adhere to the GDPR regulation. All personal data will be anonymised. In the case it is unavoidable to collect identifiable data (e.g. contact details for participants taking part in multiple sessions), these will be removed at the earliest opportunity.

Each partner that is subject to the obligation to appoint a DPO according to the GDPR, has appointed a Data Protection Officer (DPO). The contact details of the DPO will be available to all data subjects involved in the research lead by that specific partner (qualifiable as data controller) through the informed consent and the information sheet. In case there is any doubt about the procedures to follow, an independent Ethics Advisor will be consulted prior to the pilot study activities.

6.1.1. Informed consent

Informed consent forms will be used 1) to gather data and/or consent from participants as the host of an academic event as part of the registration procedure 2) as part of invitations to contribute to the surveys conducted by T3.2. and T3.5. In both cases, consulting the DARIAH ELDAH Consent Form Wizard (https://consent.dariah.eu/) will ensure that consent is obtained for all possible reuse scenarios and that the information sheet is compliant with the requirements of GDPR. Consent forms will be attached as appendices to a later iteration of this DMP.

Informed consent for participation in the surveys is sought at the time of the invitation. Panel members are able to easily withdraw their consent and fully exercise their data access rights. Data protection is a primary consideration and is intended to provide efficient coordination of cross-national panel management between national and central teams.

As a rule, the CLS INFRA will only accept participants for studies who are fully able to give informed consent, or, in case of minors, consent will be gained from the parents, guardians, or legal representatives. All participants will be provided with an information sheet and, after being informed about the project, asked to sign a consent form. The consent form as well as the information sheet will clearly state the purpose of the project and the methodology used. It will

also describe how the data will be stored who has access to it. All participants are pointed to their right to withdraw and full contact details of a contact person will be provided.

The implementation of GDPR policies will be led by T5.3 Policies for sharing data as an institution or an individual (UBER, DARIAH).

6.2. Safeguards for adherence to intellectual property rights

Another crucial ethical and legal dimension of the CLS INFRA project concerns establishing reuse rights and enhancing access to literary corpora that are in copyright. There are three dedicated task forces within the project for the establishment of sharing rights and for the development of legal and technical protocols to access, analyze and enrich in copyright materials:

- **T5.3. Policies for sharing data as an institution or an individual (UBER, DARIAH)** (**D5.3** *Toolkit report for data sharing between researchers and institutions in the field of literary studies, including literature review, gap analysis, case studies and sharing tool templates* [M36])
- **T7.4 Case study on integrating existing apps into the ecosystem (ENSL, UP),** to develop concepts and technical solutions for CLS research in copyright material in order to explore the possibilities of reproducible research on 20th and 21st century literature (**D7.4** *Report on the implementation and prototyping of programmable corpora, addressing issues of genres, integration of existing applications (e.g. TXM platform), and potential for non-consumptive reuse scenarios and derived formats* [M48])
- **T7.5. Non-consumptive reuse scenarios (UP, UT),** to allow researchers unlimited access not to the full-text of the copyright corpora, but to datasets derived from these corpora and containing text statistics.

All the results of these task forces will be integrated with future versions of this DMP.

References

Carsten Thiel, Michelle Weidling, Yoann Moranville, 2018. The EURISE Network Technical Reference. https://technical-reference.readthedocs.io/en/latest/

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, Jörg Röpke: Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel 2020. text/html Format. DOI: 10.17175/2020_006

Hong, N. P. C., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., & Honeyman, T. (2021). FAIR Principles for Research Software (FAIR4RS Principles). *Research Data Alliance*. DOI: 10.15497/RDA00065

Jiménez RC, Kuzak M, Alhamdoosh M et al. Four simple recommendations to encourage best practices in research software [version 1; referees:3]. F1000Research 2017, 6:876 (doi: 10.12688/f1000research.11407.1)

Lamprecht, Anna-Lena et al. 'Towards FAIR Principles for Research Software'. 1 Jan. 2020 : 37 – 59. DOI: 10.3233/DS-190026

Tasovac, Toma, Sally Chambers, and Erzsébet Tóth-Czifra. 2020. Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper. https://hal.archives-ouvertes.fr/hal-02961317.

'Depositing Scientific Software into Software Heritage'. 2018. https://www.softwareheritage.org/2018/09/28/depositing-scientific-software-into-software-heritage/.

'Documentation — Technical Reference 0.3-Snapshot Documentation'. n.d. Accessed 30 August 2021. https://technical-reference.readthedocs.io/en/latest/developer-guidelines/03-documentation.html.

'DraCor – Drama Corpora Project'. n.d. Accessed 30 August 2021. https://dracor.org.

Druskat, Stephan. n.d. ‘Citation File Format (CFF)’. Citation File Format (CFF). Accessed 30 August 2021. https://citation-file-format.github.io/.

‘EUR-Lex - 32019L0790 - EN - EUR-Lex’. n.d. Accessed 30 August 2021. https://eur-lex.europa.eu/eli/dir/2019/790/oj.

‘European Literary Text Collection (ELTeC) – Distant Reading for European Literary History’. n.d. Accessed 30 August 2021. https://www.distant-reading.net/eltec/.

‘How to Cite and Describe Software’. n.d. Accessed 30 August 2021. https://www.software.ac.uk/how-cite-software.

‘Metamodeling with Metamodels Using UML/MOF Including OCL’. n.d. Accessed 25 August 2021. http://www.db.informatik.uni-bremen.de/teaching/courses/ss2020_eis/week10/H-mm.pdf.