Recognition and Enrichment of Archival Documents

D2.8. Data Management Plan

Report for Period 1

Günter Mühlberger (UIBK)

Distribution: Public

http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	**READ**
Project full title	**Recognition and Enrichment of Archival Documents**
Instrument	H2020-EINFRA-2015-1
Thematic Priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date / duration	01 January 2016 / 42 Months

Distribution	Public
Contractual date of delivery	31.12.2016
Actual date of delivery	29.12.2016
Date of last update	23.12.2016
Deliverable number	D2.8.
Deliverable title	Data Management Plan
Type	Report
Status & version	Final
Contributing WP(s)	WP5, WP6, WP7, WP8
Responsible beneficiary	UIBK
Other contributors	All beneficiaries
Internal reviewers	
Author(s)	Günter Mühlberger
EC project officer	Martin Majek
Keywords	Data Management Plan – Initial Version

Table of Contents

Executive Summary

This paper provides an initial version of the Data Management Plan in the READ project. It is based on the DMP Online questionnaire provided by the Digital Curation Centre (DDC) and funded by JISC: https://dmponline.dcc.ac.uk/. We have included the original questions in this paper (indicated in italic).

The management of research data in the READ project is strongly based on the following rules:

- Apply a homogenous format across the whole project for any kind of data
- Use a well-known external site for publishing research data (ZENODO)
- Encourage data providers to make their data available via a Creative Commons license
- Raise awareness among researchers, humanities scholars, but also archives/libraries for the importance of making research data available to the public

1. Data summary

Provide a summary of the data addressing the following issues:

* State the purpose of the data collection/generation * Explain the relation to the objectives of the project * Specify the types and formats of data generated/collected * Specify if existing data is being re-used (if any) * Specify the origin of the data * State the expected size of the data (if known) * Outline the data utility: to whom will it be useful

The main purpose of all data collected in the READ project is to support research in Pattern Recognition, Layout Analysis, Natural Language Processing and Digital Humanities. In order to be useful for research the collected data must be "reference" data.

Reference data in the context of the READ project consist typically of a page image from a historical document and of annotated data such as text or structural features from this page image.

An example: In order to be able to develop and test Handwritten Text Recognition algorithms we will need the following data: First a (digital) page image. Second the correct text on this page image, more specifically of a line. And thirdly an indication (=coordinates of line region), where the text can be found exactly on this page image. The format used in the project is able to carry this information. The same is true for most other research areas supported by the READ project, such as Layout Analysis, Image pre-processing or Document Understanding.

Reference data are of highest importance in the READ project since not only research, but also the application of tools developed in the project to large scale datasets is directly based on such reference data. The usage of a homogenous format for data production was therefore one of the most important requirements in the project. READ builds upon the PAGE format, which was introduced by the University of Salford in the FP7 Project IMPACT. It is well-known in the computer science community and is able to link page images and annotated data in a standardized way.

2. Fair data

2.1 Making data findable, including provisions for metadata

* Outline the discoverability of data (metadata provision) * Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers? * Outline naming conventions used * Outline the approach towards search keyword * Outline the approach for clear versioning * Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

Part of the research in the Document Analysis and Recognition community is carried out via scientific competitions organized within the framework of the main conferences in the field, such as ICDAR (International Conference on Document Analysis and Recognition) or ICFHR (International Conference on Frontiers in Handwriting Recognition). READ partners are playing an important role in this respect and have organized several competitions in recent years.

One of the objectives of READ is to support researchers in setting up such competitions. Therefore the ScriptNet platform was developed by the National Centre for Scientific Research – Demokritos in Athens to provide a service for organizing such competitions. The datasets used in such competitions will be made available as open as possible.

For this purpose we are using the ZENODO platform and have set up the corresponding ScriptNet community: https://zenodo.org/communities/scriptnet/. In comparison to current competitions this is a step towards making Research Data Management more popular in the Pattern Recognition and Document Analysis community.

The format of the data is simple: As indicated above all data are coming in the PAGE XML format, together with images and a short description explaining details of the reference data.

Since all data in the READ project are created in the Transkribus platform and with the Transkribus tools, the data format is uniform and can also be generated via the tool itself. In this way we hope to encourage as many researchers but also archives and libraries to provide research data.

2.2 Making data openly accessible

* Specify which data will be made openly available? If some data is kept closed provide rationale for doing so * Specify how the data will be made available * Specify 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)? * Specify where the data and associated metadata, documentation and code are deposited * Specify how access will be provided in case there are any restrictions

All data produced in the READ project are per se freely accessible (or will become available during the course of the project). We encourage data providers to use the Creative Commons schema (which is also part of the upload mechanism in ZENODO) to make their data available to the public. Nevertheless some data providers (archives, libraries) are not prepared to share their data in a completely open way. In contrast rather strict regulations are set up to restrict data usage even for research and development purposes. Therefore some dataset may be handed over just on request of specific users and after having signed a data agreement.

2.3 Making data interoperable

* Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. * Specify whether you will be using standard

vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Due to the fact that data in the READ project are handled in a highly standardized way data interoperability is fully supported. As indicated above the main standards in the field (XML, METS, PAGE) are covered and can be generated automatically with the tools used in the project.

2.4 Increase data re-use (through clarifying licenses)

* Specify how the data will be licenced to permit the widest reuse possible * Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed * Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why * Describe data quality assurance processes * Specify the length of time for which the data will remain re-usable

As indicated above we encourage use of Creative Commons and support other licenses only as exceptions to this general policy.

3. Allocation of resources

Explain the allocation of resources, addressing the following issues: * Estimate the costs for making your data FAIR. Describe how you intend to cover these costs * Clearly identify responsibilities for data management in your project * Describe costs and potential value of long term preservation

Data Management is covered explicitly by the H2020 e-Infrastructure grant. All beneficiaries are obliged to follow the outlined policy in the best way they can.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

We distinguish between working data and published data. Working data are all data in the Transkribus platform. This platform is operated by the University of Innsbruck and data backup and recovery is part of the general service and policy of the Central Computer Service in Innsbruck. This means that not only regular backups of all data and software are carried out, but that a distributed architecture exists which will secure data even in the case of flooding or fire. Security is also covered by the Central Computer Service comprising regular security updates, firewalls and permanent evaluation. Published data are still kept on the Transkribus site as well, but are also made available via ZENODO.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

There are no ethical issues connected with the management of research data in READ. Nevertheless the only aspect which might play a role in the future are documents from the 20th century coming with personal data. For this case the Transkribus site offers a solution so that specific aspects of such documents - which may be interesting research objects - can be classified (e.g. person names) in a way that research can be carried out but without conflicting with personal data protection laws.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Not applicable

7. References

PAGE (Page Analysis and Ground-Truth Elements) Format Framework

- S. Pletschacher, A. Antonacopoulos: Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260
- PAGE XML Schema:
- http://www.primaresearch.org/schema/PAGE/gts/pagecontent/2016-07-15/pagecontent.xsd

Transkribus

- Transkribus client for creating annotated data sets and exporting them as PAGE files
- http://transkribus.eu/

ScriptNet

- Competitions site for Document Analysis and Recognition
- https://scriptnet.iit.demokritos.gr/competitions/

ZENODO – ScriptNet Community

- Published datasets
- https://zenodo.org/communities/scriptnet/