Funded by the Horizon 2020 Framework Programme of the European Union

D9.7 Data Management Plan

PROJECT DOCUMENTATION SHEET	
Project Acronym	Easy Reading
Project Full Title	Easy Reading
Grant Agreement	780529
Call Identifier	H2020-ICT-2016-2017
Topic	ICT-23-2017
Funding Scheme	RIA (Research and Innovation action)
Project Duration	30 months (January 2018 – June 2020)
Project Officer	Michael Busch, European Commission Directorate-General for Communications Networks, Content and Technology, Unit &-3, L-2557 Luxembourg, +352.4301.38082
Coordinator	Universität Linz (JKU), Austria
Consortium partners	Kompetenznetzwerk (KI-I), Austria Technische Universität Dortmund (TUDO), Germany In der Gemeinde Leben gGmbH (IGL), Germany FUNKA Nu AB (FUNKA), Sweden Texthelp Ltd (TEXTHELP), United Kingdom VÄSTRA GÖTALANDS LÄNS LANDSTING (DART), Sweden GEIE ERCIM (ERCIM), France AHTENA I.C.T. Ltd (ATH), Israel
Website	www.easyreading.eu

DELIVERABLE DOCUMENTATION SHEET	
Number	Deliverable D9.7
Title	Data Management Plan
Related WP	WP 9
Related Task	9.4
Lead Beneficiary	JKU
Author(s)	Reinhard Koutny
Contributor(s)	Peter Heumader
Reviewers	DART
Nature	ORDP
Dissemination level	Public
Due Date	30.6.2018
Submission date	30.6.2018
Status	Final

QUALITY CONTROL ASSESSMENT SHEET			
Issue	Date	Comment	Author
-	20.6.2018	Approved	Sandra Derbring

ACKNOWLEDGEMENT

This document is a deliverable of the Easy Reading project, which has received funding from the European Union's Horizon 2020 Programme for Information and Communication Technologies under Grant Agreement (GA) Nb #780529.

Executive Summary

The present document is deliverable “D9.7 Data Management Plan” of the Easy Reading project which is funded by the European Union’s Horizon 2020 Programme und Grant Agreement #780529.

The purpose of this document is to provide the plan for managing the data generated and collected during the project. The Data Management Plan (DMP) describes the data management life cycle for all data sets to be collected, processed and/or generated by a research project. It covers:

- The handling of research data during and after the project
- What data will be collected, processed or generated
- What methodology and standards will be applied
- Whether data will be shared/made open and how
- How data will be curated and preserved

The DMP is currently in an initial state, as the project has just started. Following the EU’s guidelines regarding the DMP, this document may be updated - if appropriate - during the project lifetime (in the form of deliverables).

The DMP currently identifies the following data as research data generated during the project:

- Structure of the user profile
- Usage statistics
- Anonymized user profile data
- User evaluations
- Administrative metadata

Most data sets will be provided openly on public web servers, via a REST-API[1] for real time data or other means of provision. As the user profiles may contain sensitive data even if they are anonymized, it is, at this stage, currently unclear what will be openly available and what not.

[1] https://en.wikipedia.org/wiki/Representational_state_transfer

Table of Content

Introduction

This document is the Data Management Plan (DMP). The consortium is required to create the DMP because the Easy Reading project participates in the Open Research Data pilot. The DMP describes the data management life cycle for all data sets to be collected, processed and/or generated by a research project.

Scope

The present document is the Deliverable 9.7 "D9.7 – Data Management Plan" (henceforth referred to as D9.7) of the Easy Reading project. The main objective of D9.7 is to provide the plan for managing the data generated and collected during the project.

According to the EU's guidelines regarding the DMP, the document may be updated - if appropriate - during the project lifetime (in the form of deliverables).

Audience

The intended audience for this document is the Easy Reading consortium and the European Commission.

Data Summary

The Easy Reading framework will improve the cognitive accessibility of original digital documents by providing real time personalisation through annotation (using e.g. symbol, pictures, video), adaptation (using e.g. layout, structure) and translation (using e.g Easy-to-Read, Plain Language, symbol writing systems). The framework provides these (semi-)automated services using HCI techniques (e.g. pop-ups/Text-To-Speech (TTS)/captions through mouse-over or eye-tracking) allowing the user to remain and work within the original digital document. This fosters independent access and keeps the user in the inclusive discourse about the original content. Services adapt to each user through a personal profile (sensor based tracking and reasoning of e.g. the level of performance, understanding, preferences, mood, attention, context and the individual learning curve).

During the project, data will be generated to improve the Easy Reading framework, to model the capabilities and preferences of the user and to evaluate the success of the project. The purpose of the data collection/generation can be subdivided into the following points:

- **Modelling the user:** The model of the user is required as a basis to provide services on top of this information which helps users with cognitive disabilities browsing the web.
- **Framework performance monitoring and improvement**: Collected data will be used to improve the tool. Evaluations of usage statistics for example will be used to determine which functions were more helpful than others, which configurations were used and on which content areas issues occurred for the user.
- **Matching services to user needs:** Based on user profile data and usage data, suggestions for services/functions can be made automatically.
- **Pre-configuration of functions to simplify web content:** Depending on user profile data and usage data, functions can be pre-configured to provide helpful support and a good user experience from the start when installing a new function.
- **Adjusting functions according to user profile:** Besides the initial configuration of the functions, adjustments and fine-tuning on-the-fly using usage statistics will be possible too.
- **Learning about the target group:** From a research perspective, major contributions in the field of web accessibility and people with cognitive disabilities can be made by collection and analysis of usage statistics.
- **Deducing rules for cognitively accessible web content:** Based on these findings, rules for cognitively accessible web content may be deduced and support web developers and content creators to make website easier to understand for everyone.

Types and Formats of Data

Currently the following data sets were identified. As mentioned before, these are subject to changes during the project lifetime.

- Structure of the user profile:
  - Cognitive capabilities of a user
  - Current mood of a user
  - Level of confusion
  - Usage statistics
  - Which target group uses which kinds of services

- Accumulated usage statistics
    - Time of the day
    - Mood
- Disabilities / CapabilitiesAnonymized user profile data
    - Function configuration preferences: Describes which configuration of a function the user uses
    - Time consumption per UI-element: Describes at which elements of the content the user spends most of the time
    - Level of confusion per element: Describes if there are elements of the content which especially confuse the user
    - Understandabilty of elements: Describes which parts of the content the user understands/are easy to interact with, and which are not
- User evaluations: Evaluations will be conducted during the course of the project to to ensure requirements of this manifold user group are considered adequately from the very beginning of the project. These evaluations involve user satisfaction tests to ensure that the user interface as well as the features of the tool meet the actual needs of people with cognitive disabilities.
- Administrative metadata: When and how data was created e.g.

The main data exchange format for data sets will be JSON, while the data itself is stored in a relational database. Also other kinds of structured data will probably be made openly available using a REST-service[2] using JSON[3] as data format. Manually created data, like evaluations or data used in publications, will be made available as PDFs.

Reuse of existing data

At this state of the project, no existing data is planned to be reused. This might be subject to change during the lifespan of the project.

Origins of Data

Most data will be generated and retrieved during the actual usage of the tool, either by user interaction or by tracking of the user's actions. In addition, relevant data will be created due to configuration processes either by the user or a caregiver. Evaluations will also provide further useful data.

- Data generated by user interaction: During user interaction, data can and will be collected to achieve the goals mentioned before. Amongst others, the following data may be useful:
    - The functions the user prefers to use
    - The way the user uses functions and how the tool and its features are configured
    - The kind of websites the user visits. This is very sensitive data of course, so some other data might be used instead. For example not the actual website, but measures like the complexity of the website layout or the complexity of the text.
    - The time users spend on website, or parts of a website. Again, this is of course very sensitive data and due to ethical considerations, other measures which reflect the most interesting outcomes will be used instead.

[2] https://en.wikipedia.org/wiki/Representational_state_transfer
[3] https://en.wikipedia.org/wiki/JSON

    - Level of confusion of the user
- Tracking of the user with sensors will be possible:
    - Mouse and keyboard tracking
    - Head and eye tracking
- Due to client-side configuration of functions, data will also be generated
- Data will also be generated by carers who can do an initial configuration or users of the target group who can use a wizard independently, which allows to manually add information about capabilities and preferences
- User evaluations through questionnaires, interviews and other means.

Expected Size of the Data

At this state of the project, the expected size of the data is unknown, as the user profile and its structure is not fully defined and implemented.

Data Utility

The data will be useful to the project (consortium), to other research projects in a similar field which are concerned with people with cognitive disabilities consuming web content and for companies which want to create products or content of this kind with people with cognitive disabilities in mind.

FAIR Data

The research data generated by the Easy Reading project should be 'FAIR'; findable, accessible, interoperable and re-usable.

Findable Data

- **Discoverability of data:** Since the user and usage data is stored in a relational database, it can be accessed using SQL queries. The database structure/schema will be made available in some sort of wiki.
- **Identifiability of data:** Not specified at this stage of the project
- **Naming conventions:** Most of the data will be stored in a relational database. Therefore standard SQL database naming convention are being used:
    - Singular names for tables
    - Singular names for columns
    - Schema name for tables prefix (E.g.: SchemeName.TableName)
    - Pascal casing (a.k.a. upper camel case)
- **Search keywords:** Not specified at this stage of the project
- **Clear versioning:** Not specified at this stage of the project
- **Metadata creation standards used:** The use of specific standard for metadata creation is not yet decided upon.

Accessible Data

Data generated by the Easy Reading project contains sensitive data. For example the user profiles themselves, even if they are anonymized, are considered sensitive. Therefore the consortium is very cautious on making this data openly accessible. Currently following data might be made openly available:

- structure of the user profile

- user evaluations
- accumulated usage statistics (in contrast to individual user statistics)
- deduced data from evaluations as part of publications

At this stage of the project it is not fully decided if individual usage statistics, anonymized user profile data and administrative metadata will be made openly accessible due to the sensitive nature of this kind of data.

Openly available data will be made public by following means:

- Publications (evaluations, accumulated usage statistics)
- Direct download (user profile structure)

To access openly available data no special software or method will be required at the current stage. For publications and evaluations, a standard PDF viewer is sufficient. Usage statistics and the user profile structure can be displayed using a JSON scheme parser. This data and its associated metadata, documentation and code are deposited on the project website.

Data restrictions: There are no restrictions to openly available data. Other data is currently only available for consortium members and it needs to be decided which parts of this data will be made available and in which way.

Interoperable Data

Data is structured (relational database, JSON), but due to the highly sensitive nature of parts of this data, it is not openly available at this point in time. However, in the future interoperability of less sensitive parts of the data is easily possible (e.g. parts of usage statistics with REST service and data exchange format JSON). Data types are not fully decided at this stage of the project and will evolve over the course of the project, but standard types will of course be applied as often as possible.

Data Reusability

To ensure FAIR use of data and to ensure widest reuse possible, openly available data generated by the project is planned to be licensed under the Apache License 2.0. Third parties will be able to use this openly available data.However, it is at this stage of the project not possible to specify a date when the data is available for reuse. Neither is it fully decided which parts of the data will be openly available. For instance, parts of the usage statistics might be deemed as sensitive data in terms of privacy.

Allocation of Resources

Costs for making your data FAIR are covered by the project budget. The project leader (JKU) coordinates the data management of the project. The costs and potential value of long term preservation has not yet been determined due to the early stage of the project. This will be dealt with later on over the course of the project

Data Security

As sensitive data is stored and transferred, data security is of utmost importance for the Easy Reading project. At the moment, data used by the framework is stored on cloud server infrastructure from IBM (IBM Bluemix). Later on in the project this could be moved to Amazon AWS. Both platforms

provide sufficient measures and tools to provide data security. Data will be stored on Servers in the EU.

- IBM Bluemix Cloud Security: https://console.bluemix.net/docs/security/index.html#security
- Amazon AWS Cloud Security https://aws.amazon.com/security/

Regular backups using cron-jobs of the relational database will ensure easy data recovery. Data transfer will be exclusively done via HTTPS /WSS (Web Socket Security).

Ethical Aspects

At the current stage of the project all ethical aspects are covered in section 5 “Ethics and Security” of the Grant Agreement of the project.

Conclusions

The purpose of this document is to provide the plan for managing the data generated and collected during the project; The Data Management Plan. Specifically, the DMP describes the data management life cycle for all data sets to be collected, processed and/or generated by a research project. It covers:

- the handling of research data during and after the project
- what data will be collected, processed or generated.
- what methodology and standards will be applied.
- whether data will be shared/made open and how

Following the EU’s guidelines regarding the DMP, this document may be updated - if appropriate - during the project lifetime (in the form of deliverables).

Due to the sensitive nature of some data collected, it is unclear at this point of the project if all data will be made openly available. Data sets that will be openly provided to the public will be hosted on web servers or provided in real time via REST endpoints. Finally, data sets will be preserved after the end of the project on the pilot’s web sites, on web servers or other web-based solutions.