OPEN FRAMEWORK OF E-SERVICES FOR MULTILINGUAL AND SEMANTIC
ENRICHMENT OF DIGITAL CONTENT

DATA MANAGEMENT PLAN

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D7.4
DELIVERABLE TITLE	Data Management Plan
RESPONSIBLE AUTHOR	DFKI

GRANT AGREEMENT N.	644771
PROJECT REF. NO	H2020-644771
PROJECT ACRONYM	FREME
PROJECT FULL NAME	Open Framework of E-Services for Multilingual and Semantic Enrichment of Digital Content
STARTING DATE (DUR.)	06/02/2015 (24 months)
ENDING DATE	06/10/2017
PROJECT WEBSITE	www.freme-project.eu
COORDINATOR	Felix Sasaki
ADDRESS	DFKI Alt-Moabit 91c 10559 Berlin
REPLY TO	Felix.sasaki@dfki.de
PHONE	+49-30 23895 1807
FAX	+49-30 23895 1810
EU PROJECT OFFICER	Alexandra Wesolowska

WORKPACKAGE N. | TITLE	WP7 | Project Management
WORKPACKAGE LEADER	DFKI
DELIVERABLE N. | TITLE	D7.4 | Data Management Plan
RESPONSIBLE AUTHOR	Nieves Sande (DFKI)
REPLY TO	nieves.sande@dfki.de
DOCUMENT URL	http://www.freme-project.eu/resources/deliverables/
DATE OF DELIVERY (CONTRACTUAL)	31.07.2015
DATE OF DELIVERY (SUBMITTED)	30.07.2015
VERSION | STATUS	V5 Final
NATURE	R (Report)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Nieves Sande (DFKI)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
01	Document creation	19/06/2015	DFKI
02	First Draft	24/06/2015	DFKI
02	Updated Draft DFKI internally	09/07/2015	DFKI
03	Internal Reviewed	17/07/2015	AK
04	Internal Correction	20/07/2015	DFKI
05	Final Version	24/07/2015	DFKI

PARTICIPANTS	CONTACT
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI, Germany)	Felix Sasaki Email: felix.sasaki@dfki.de
Institut für Angewandte Informatik Ev (InfAI, Germany)	Sebastian Hellmann Email: hellmann@informatik.uni-leipzig.de
Tilde SIA (Tilde, Latvia)	Tatiana Gornostay Email: tatiana.gornostay@tilde.lv
iMINDS VZW (IMINDS, Belgium)	Erik Mannens Email: erik.mannens@ugent.be
VistaTEC EV (VTEC, Ireland)	Phil Ritchie Email: phil.ritchie@vistatec.com
Agro-Know IKE (AK, Greece)	Giannis Stoitsis Email: stoitsis@agroknow.gr
Wripl Technologies Limited (Wripl, Ireland)	Kevin Koidl Email: kevin@wripl.com
Istituto Superiore Mario Boella (ISMB, Italy)	Michele Osella Email: osella@ismb.

ACRONYMS LIST

DMP	Data Management Plan
H2020	Horizon 2020
EC	European Commission
API	Application Programming Interface
GUI	Graphical User Interface
BC	Business Case
ICT	Information and Communication Technologies
LLD	Linguistic Linked Data
ISLRN	International Standard Language Resource Number
PID	Persistent Identifier

EXECUTIVE SUMMARY

This deliverable provides the FREME data management plan version 1. The Deliverable outlines how the research data collected or generated will be handled during and after the FREME action, describes which standards and methodology for data collection and generation will be followed, and whether and how data will be shared. This document follows the template provided by the European Commission in the Participant Portal[1].

---

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

TABLE OF CONTENTS

1 BUILDING A DMP IN THE CONTEXT OF H2020

1.1 PURPOSE OF THE FREME DATA MANAGEMENT PLAN (DMP)

FREME is a Horizon 2020 project participating in the Open Research Data Pilot. This pilot is part of the Open Access to Scientific Publications and Research Data programme in H2020[2]. The goal of the program is to foster access to data generated in H2020 projects.

Open Access refers to a practice of giving online access to all scholarly disciplines information that is free of charge to the end-user. In this way data becomes re-usable, and the benefit of public investment in the research will be improved.

The EC provided a document with guidelines[3] for projects participants in the pilot. The guidelines address aspects like research data quality, sharing and security. According to the guidelines, projects participating will need to develop a DMP.

The DMP describes the types of data that will be generated or gathered during the project, the standards that will be used, the ways how the data will be exploited and shared for verification or reuse, and how the data will be preserved.

This document has been produced following these guidelines and aims to provide a consolidated plan for FREME partners in the data management plan policy that the project will follow. The document is the first version of the DMP, delivered in M6 of the project. The DMP will be updated during the lifecycle of the project. This will be documented in deliverables D7.5 and D7.6 in M12 and M24 respectively.

1.2 BACKGROUND OF THE FREME DMP

The FREME DMP will be written in reference to the Article 29.3 in the Model Grant Agreement called “Open access to research data” (research data management). Project participants must deposit their data in a research data repository and take measures to make the data available to third parties. The third parties should be able to access, mine, exploit, reproduce and disseminate the data. This should also help to validate the results presented in scientific publications. In addition Article 29.3 suggests that participants will have to provide information, via the repository, about tools and instruments needed for the validation of project outcomes.

The DMP will be important for tracking all data produced during the FREME project. Article 29 states that project beneficiaries do not have to ensure access to parts of research data if such access would be lead to a risk for the project’s goals. In such cases, the DMP must contain the reasons for not providing access.

According to the forehand mentioned DMP Guidelines it is planned that research data management projects funded under H2020 will receive support through the Research Infrastructures Work Programme 2014-15 (call 3 e-Infrastructures). Full support services are expected to be available only to research projects funded under H2020, with preference to those participating in the Open Research Data Pilot.

---

[2] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[3] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

2 FREME DATA MANAGEMENT PLAN

2.1 OBJECTIVES OF THE FREME PROJECT

FREME general objective is to build an open innovative commercial grade framework of e-services for multilingual and semantic enrichment of digital content. Under digital content we understand any type of content that exists in a digital form, and available in various formats. FREME will improve existing processes of digital content management by utilising vast amounts of multilingual structured and unstructured datasets and reusing them in its enrichment services.

Under enrichment we understand annotation of content with additional information. We focus on semantic and multilingual enrichment. One aim of FREME is to transform unstructured content into its structured representation.

In terms of data and tooling, FREME will produce the following:

- Six e-Services realised as Web services for semantic and multilingual enrichment of digital content;
- Access to the e-Services via APIs and GUIs;
- Access to existing data sets for enrichment;
- Conversion of selected data sets into a standardised, linked data representation to make them suitable for enrichment;
- Facilities for FREME users to their own convert data sets into linked data for usage in enrichment scenarios.

The design of the FREME e-Services, the APIs and GUIs, and the selection of data sets is driven by the FREME business case partners, working on four business scenarios:

- BC 1: Authoring and publishing multilingually and semantically enriched eBooks
- BC 2: Integrating semantic enrichment into multilingual content in localisation
- BC 3: Enhancing cross-language sharing and access to open data
- BC 4: Empowering personalised content recommendation

One crucial aspect of FREME project will be to provide new business opportunities for these scenarios. Hence, the requirements on data management depend on the context of each business case and must not hinder the business opportunities.

2.2 DATA MANAGEMENT RELATED TO LANGUAGE AND DATA TECHNOLOGIES

FREME is a project building bridges between two communities: Language technologies and data technologies.

In terms of EC funding, the current focus of language technology is in ICT 17[4]. The ICT 17 mostly relevant project for data management is CRACKER[5]. CRACKER adopts and promotes methodologies developed within the META-NET initiative. For example, for its data management plan, CRACKER follows META-SHARE[6]. META-SHARE is a network of repositories of language resources, including data and tooling that provides a metadata scheme for documenting language resources.

With its “Cracking the Language Barrier” initiative CRACKER is promoting a collaboration that includes, among others, projects funded through ICT 17 and ICT 15. FREME signed the corresponding Memorandum of Understanding and is participating in this collaboration. As part of the effort FREME will

---

[4] See http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/88-ict-17-2014.html

[5] See http://www.meta-net.eu/projects/cracker

[6] See http://www.meta-share.eu/

include in its metadata description the META-SHARE template provided by CRACKER in its initial data management plan.

In terms of data technologies, in the last two years, the data community has gathered in the LIDER project[7] a group of stakeholders around Linguistic Linked Data (LLD). LLD means the representation of language resources using linked data principles[8]. The LIDER project will end in October 2015. One outcome of LIDER is guidelines on working with linguistic linked data[9]. FREME will adopt the general guideline of how to include data in the linguistic linked data cloud and the specific guidelines of using the DataID metadata format.

The usage of DataID builds a bridge to the DBpedia community and the DBpedia association. DataID is also used in other H2020 projects, especially the ALIGNED project[10]. The tooling for creating metadata DataID metadata records will also be used in FREME.

In summary, the FREME approach to data management will be:

- Provide META-SHARE metadata following the dataset description provided by the CRACKER DMP; and
- Provide DATAID metadata following the metadata blueprint provided by the DBpedia community and the ALIGNED DMP.

---

[7] See http://lider-project.eu/

[8] See http://www.w3.org/DesignIssues/LinkedData.html

[9] See http://www.lider-project.eu/guidelines

[10] See http://aligned-project.eu/

3 METADATA FOR DATA MANAGEMENT

This section provides details on META-SHARE and DataID and their role for data management within FREME.

3.1.1 META-SHARE

META-SHARE is a sustainable network of repositories of language data, tools and related web services documented with high-quality meta-data, aggregated in central inventories allowing for uniform search and access to resources. Data and tools can be both open and with restricted access rights, free and for-a-fee. META-SHARE targets existing but also new and emerging language data, tools and systems required for building and evaluating new technologies, products and services.

This network of data, tools and web services repositories started with the integration of nodes and centres represented by the partners of the META-NET consortium, and its best practices in datasets description are followed by CRACKER DMP.

META-SHARE is gradually being extended to encompass additional nodes and centres, and to provide more functionality with the target of turning into a largely infrastructure as possible.

FREME will follow META-SHARE practices for data documentation, verification and distribution, as well as for curation and preservation, ensuring the availability of the data and enabling access, exploitation and dissemination.

3.1.2 DataID

DataID is a machine-readable metadata put forward within the community of linguistic linked data. DataID is used in the DBpedia community and in the ALIGNED DMP. The FREME consortium partner InfAI is also partner in ALIGNED project.

The effort of DataID team is a tool called DMP generator. The generator takes as input a DataID file and produces an HTML report that can be directly used as basis for DMP. Currently the generator is in early prototype stage. FREME will decide in an updated version of its DMP whether to rely on this tool for automatic DataID generation.

The DataID model establishes a system to describe metadata for datasets. This system improves datahub.io, a data management platform by the Open Knowledge Foundation adding richer semantics in several properties relevant to LOD datasets.

4 DATA DESCRIPTION

For FREME DMP, both data set description models - the one following the principles of CRACKER DMP, and the other one following the principles of ALIGNED DMP – will be used. In future steps of the project it will be decided which structure fits better to which of the FREME requirements.

There is an exhaustive list of FREME data sets and tools in FREME deliverable D1.1 “Initial requirement analysis and specification for the FREME framework and the e-Services”. This list is just a collection of datasets that will be probably taken in consideration for FREME.

In the following data set descriptions DBpedia serves as an example following both different data set models described above, since it is likely to be used in FREME.

4.1 DATA SET DESCRIPTION FOLLOWING META-SHARE PRINCIPLES

4.1.1 Data Set Reference and Name

FREME following the META-SHARE principles described in CRACKER DMP model will employ a standard identification mechanism for each data set. A PID or the ISLRN will be used to identify the dataset.

4.1.2 Data set description

The data set description is based on the DMP template circulated by CRACKER. This template will be circulated and further elaborated upon within the “Cracking the Language Barrier” initiative mentioned in section 2.2. It will include the following information items:

Resource Name	Complete title of the resource
Resource Type	Choose one of the following values: Lexical/conceptual resource, corpus, language description
Media Type	The physical medium of the content representation, e.g. video, image, text, numerical data, n-grams, etc.
Language (s)	The language(s) of the resource content
License	The licensing terms and conditions under which the tool/service can be used
Distribution Medium	The Medium i.e. the channel used for delivery or providing access to the resource, e.g. accessible through interface, downloadable, CD/DVD, etc.
Usage	Foreseen use of the resource for which it has been produced
Size	Size of the resource with regard to a specific size unit measurement in form of a number
Description	A brief description of the main features of the features

Table 1 Table for datasets description according to CRACKER DMP

4.1.3 Example of Data set description: DBpedia

Resource Name	DBpedia 2014 Dataset
Resource Type	Lexical/conceptual resource
Media Type	Linked Data
Language (s)	126 Languages
License	CC-BY-SA 3.0
Distribution Medium	http://dbpedia.org/services-resources/datasets/datasets2014#h434-1
Usage	DBpedia as an Open Dataset
Size	1.200.000.000 Triples
Description	DBpedia is a Crowd-Sourced Community Effort

Table 2 Table for DBpedia datasets description according to CRACKER DMP

4.2 DATA SET DESCRIPTION FOLLOWING DATAID PRINCIPLES

The following template of data set description follows the DataID schema described in the ALIGNED DMP:

Data set reference and Name	Name Metada URI Homepage Publisher Maintainer
Data Set description	Description Provenance Usefulness Similar Data Re-Use and integration
Standards and Metadata	Metadata description Vocabularies and Ontologies
Data Sharing	License URL Data Set Description Openness Software Necessary Repository
Archiving and preservation	Preservation Growth Archive Size

Table 3 Table for datasets description according to ALIGNED DMP

4.2.1 Example of Data Set description: DBpedia

Data Set Reference and Name	DBpedia 2014 dataset **Metadata URI**: http://downloads.dbpedia.org/2014/dataid.ttl#dataset **Homepage:** http://dbpedia.org/ **Publisher:** DBpedia Association **Mantainer:** DBpedia Association
Data set Description	**Description:** dbpedia is a crowd-sourced community effort to extract structured information from wikipedia and make this information available on the web. Dbpedia allows you to ask sophisticated queries against wikipedia, and to link the different data sets on the web to wikipedia data. This work will make it easier for the huge amount of information in Wikipedia to be used in some new interesting ways. Futhermore it might inspire new mechanism for navigating, linking and improving the encyclopedia itself. **Provenance:** Wikipedia (Wikimedia foundation) **Usefulness:** DBpedia is a useful resource for interlinking general datasets with encyclopedic knowledge. Users profiting from DBpedia are open data developers, SMEs and researchers in data science and NLP. **Similar data:** freebase or yago provide similar datasets **Re-use and integration:** http://datahub.io/dataset/dbpedia
Standards and metadata	Metadata description is done in linked data using dataid, a metadata description vocabulary based on dcat. DMP reports are automatically generated and maintained up to date using this metadata. **Vocabularies and ontologies:** http://downloads.dbpedia.org/2014/dbpedia_2014.owl
Data sharing	**License:** cc-by-sa 3.0 **Odrl license description:** http://purl.org/net/rdflicense/cc-by-sa3.0de **Openness:** dbpedia is an open datase **Software necessary:** DBpedia needs no additional software to be used. DBpedia provides complementary software for extraction, data management and enrichment under http://datahub.io/dataset/dbpedia **Repository:** http://datahub.io/dataset/dbpedia
Archiving and preservation	**Preservation:** preservation of the DBpedia es guaranteed by archival of old versions on the archive server, the intent of the DBpedia association to keep the project running, as well as the DBpedia language chapters and the dbpedia community **Growth:** DBpedia is an ongoing open-source project. Goal of the project is the extraction of the Wikipedia, as complete as possible. Currently 126 languages are being extracted. In the future DBpedia will try to increase its importance as the center

	of the lod cloud by adding further external datasets. **Archive:** http://downloads.dbpedia.org **Size:** 1.200.000.000 triples

Table 4 Table for DBpedia datasets description according to ALIGNED DMP

4.3 DATA CURRENTLY BEING PRODUCED IN FREME

This version of the FREME DMP does not include the actual metadata about the data being produced in FREME. Access to this metadata will be provided in an updated version of the DMP. In FREME data sets and the actual FREME architecture are being produced based on requirements from FREME business case partners. Details about these requirements including information about relevant datasets are available in FREME deliverable D1.1 “Initial requirement analysis and specification for the FREME framework and the e-Services”.

REFERENCES

Guidelines on Data Management in Horizon 2020 Version 16 (1.0) December 2013, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 Version 1.0, 11 December 2013, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

ALIGNED Data Management Plan, May 2015 – (access June 2015): http://aligned-project.eu/wordpress/wp-content/uploads/2015/03/D7.2-Data-Managment-Plan-v1.04.pdf

CRACKER Data Management Plan, June 2015 – (access June 2015): Website not public yet