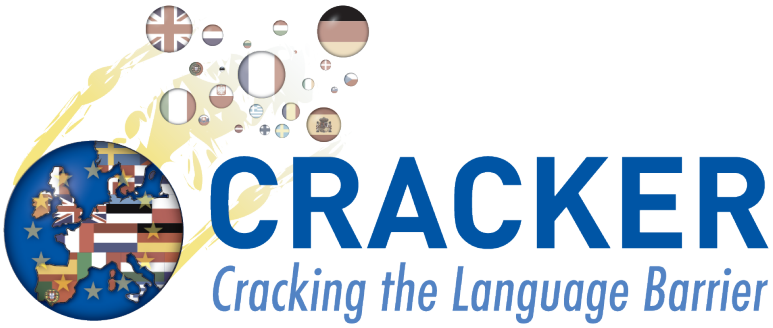
This document is part of the Coordination and Support Action CRACKER. This project has received funding from the European Union's Horizon 2020 program for ICT through grant agreement no.: 645357.



Deliverable 3.8

Data Management Plan
(Update)

Authors: Maria Koutsombogera (Athena RC)
Stelios Piperidis (Athena RC)

Dissemination Level: Public

Date: 30 June 2016

Grant agreement no.	645357
Project acronym	CRACKER
Project full title	Cracking the Language Barrier
Type of action	Coordination and Support Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	1 January 2015, 36 months
Dissemination level	Public
Contractual date of delivery	30/06/2016
Actual date of delivery	30/06/2016
Deliverable number	D3.8
Deliverable title	Data Management Plan (Update)
Type	Report
Status and version	Final
Number of pages	22
WP leader	ATHENA RC
Task leader	ATHENA RC
Author(s)	Maria Koutsombogera (ATH), Stelios Piperidis (ATH)
Contributor(s)	UEDIN, USDF, DFKI, ELDA
Internal reviewer(s)	Felix Sasaki
EC project officer	Pierre-Paul Sondag
The partners in CRACKER are:	• Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany • Charles University in Prague (CUNI), Czech Republic • Evaluations and Language Resources Distribution Agency (ELDA), France • Fondazione Bruno Kessler (FBK), Italy • Athena Research and Innovation Center in Information, Communication and Knowledge Technologies (ATHENA RC), Greece • University of Edinburgh (UEDIN), UK • University of Sheffield (USFD), UK

For copies of reports, updates on project activities, and other CRACKER-related information, contact:

DFKI GmbH
CRACKER
Dr. Georg Rehm
Alt-Moabit 91c
D-10559 Berlin, Germany

georg.rehm@dfki.de
Phone: +49 (0)30 23895-1833
Fax: +49 (0)30 23895-1810

Copies of reports and other material can also be accessed via http://cracker-project.eu.

Contents

History

Version	Date	Status	Notes
0	14/06/2016	Internal	Working version
0.1	23/06/2016	Internal	Updated dataset descriptions
1	29/06/2016	Public	Finalized after internal review

1. Executive Summary

This document describes the Data Management Plan (DMP) adopted within CRACKER and provides information on CRACKER's data management policy and key information on all datasets that have been and will be produced within CRACKER, as well as resources developed by the "Cracking the language barrier" federation of projects (also known as the "ICT-17 group of projects") and other projects who wish to follow a common line of action, as provisioned in the CRACKER Description of Action.

This second version includes the principles according to which the plan is structured, the standard practices for data management that are being implemented, and the description of the actual datasets produced within CRACKER. The final update of the CRACKER DMP document will be provided in M36 (December 2017).

The document is structured as follows:

- Background and rationale of a DMP within H2020 (section 2)
- Implementation of the CRACKER DMP (section 3)
- Collaboration of CRACKER with other projects and initiatives (section 4)
- Recommendations for a harmonized approach and structure for a Data Management Plan to be optionally adopted by the "Cracking the language barrier" federation of projects (section 5).

2. Background

The use of a Data Management Plan (DMP) is required for projects participating in the Open Research Data Pilot, which aims to improve and maximise access to and re-use of research data generated by projects. The elaboration of DMPs in Horizon 2020 projects is specified in a set of guidelines applied to any project that collects or produces data. These guidelines explain how projects participating in the Pilot should provide their DMP, i.e. to detail the types of data that will be generated or gathered during the project, and after it is completed, the metadata and standards which will be used, the ways how these data will be exploited and shared for verification or reuse and how they will be preserved.

In principle, projects participating in the Pilot are required to deposit the research data described above, preferably into a research data repository. Projects must then take measures, to the extent possible, to enable for third parties to access, mine, exploit, reproduce and disseminate, free of charge, this research data.

The guidance for DMPs calls for clarifications and analysis regarding the main elements of the data management policy within a project. The respective template identifies in brief the following five coarse categories[1]:

1. **Data set reference and name**: an identifier for the data set; use of a standard identification mechanism to make the data and the associated software easily discoverable, readily located and identifiable.
2. **Data set description**: details describing the produced and/or collected data and associated software and accounting for their usability, documentation, reuse, assessment and integration (i.e., origin, nature, volume, usefulness, documentation/publications, similar data, etc.).
3. **Standards and metadata**: related standards employed or metadata prepared, including information about interoperability that allows for data exchange and compliance with related software or applications.
4. **Data sharing**: procedures and mechanisms enabling data access and sharing, including details about the type or repositories, modalities in which data are accessible, scope and licensing framework.
5. **Archiving and preservation (including storage and backup)**: procedures for long-term preservation of the data including details about storage, backup, potential associated costs, related metadata and documentation, etc.

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

3. The CRACKER DMP

3.1 Introduction and Scope

For its own datasets, CRACKER follows META-SHARE's (http://www.meta-share.eu/) best practices for data documentation, verification and distribution, as well as for curation and preservation, ensuring the availability of the data throughout and beyond the runtime of CRACKER and enabling access, exploitation and dissemination, thereby also complying with the standards of the Open Research Data Pilot[2].

META-SHARE is a pan-European infrastructure bringing online together providers and consumers of language data, tools and services It is organized as a network of repositories that store language resources (data, tools and processing services) documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access. It serves as a component of a language resource marketplace for researchers, developers, professionals and industrial players, catering for the full development cycle of language resources and technology, from research through to innovative products and services [Piperidis, 2012].

Language resources in META-SHARE span the whole spectrum from monolingual and multilingual data sets, both structured (e.g., lexica, terminological databases, thesauri) and unstructured (e.g., raw text corpora), as well as language processing tools (e.g., part-of-speech taggers, chunkers, dependency parsers, named entity recognisers, parallel text aligners, etc.). Resources are described according to the META-SHARE metadata schema [Gavrilidou et al. 2012], catering in particular for the needs of the HLT community, while the META-SHARE model licensing scheme has a firm orientation towards the creation of an openness culture respecting, however, legacy and less open, or permissive, licensing options.

META-SHARE has been in operation since 2012, and it is currently in its 3.0.3 version, released in May 2016. It currently features 29 repositories set up and maintained by 37 organisations in 25 countries of the EU. The observed usage as well as the number of nodes, resources, users, queries, views and downloads are all encouraging and considered as supportive of the choices made so far [Piperidis et al., 2014]. Resource sharing in CRACKER will build upon and extend the existing META-SHARE resource infrastructure, its specific MT-dedicated repository (http://qt21.metashare.ilsp.gr) as well as editing and annotation tools in support of translation evaluation and translation quality scoring (e.g., http://www.translate5.net/).

This infrastructure, together with its bridges, provides support mechanisms for the identification, acquisition, documentation and sharing of MT-related data sets and language processing tools.

3.2 Dataset Reference and Name

CRACKER opts for a standard identification mechanism to be employed for each data set, in addition to the identifier used internally by META-SHARE itself.

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

Reference to the a dataset ID can be optionally made with the use of an ISLRN (International Standard Language Resource Number), the most recent universal identification schema for LRs which provides LRs with unique identifiers using a standardized nomenclature, ensuring that LRs are identified, and consequently recognized with proper references (cf. figures 1 and 2).

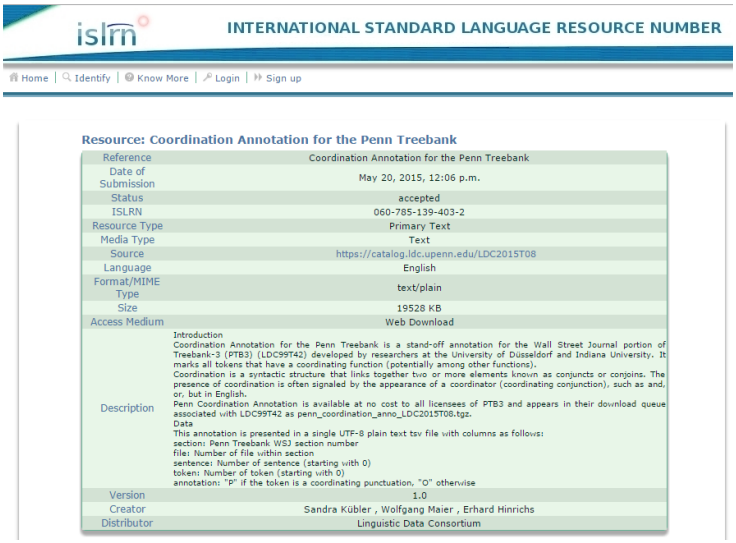


Figure 1. An example resource entry from the ISLRN website indicating the resource metadata, including the ISLRN, http://www.islrn.org/resources/060-785-139-403-2/.

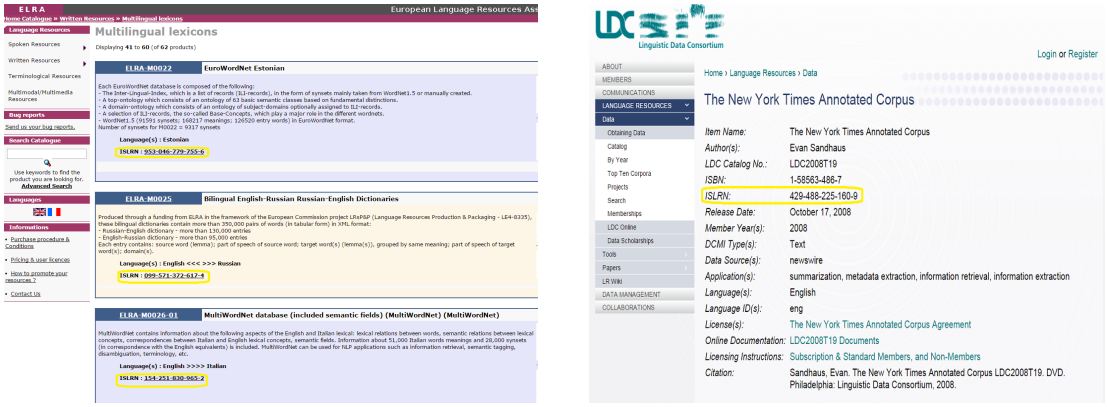


Figure 2. Examples of resources with the ISLRN indicated, from the ELRA (left) and the LDC (right) catalogues.

3.3 Dataset Description

In accordance with META-SHARE, CRACKER is addressing the following resource and media types:

- **corpora** (text, audio, video, multimodal/multimedia corpora, n-gram resources),
- **lexical/conceptual resources** (e.g., computational lexicons, ontologies, machine-readable dictionaries, terminological resources, thesauri, multimodal/ multimedia lexicons and dictionaries, etc.)
- **language descriptions** (e.g., computational grammars)
- **technologies** (tools/services) that can be used for the processing of data resources

Several datasets that have been and will be produced (test data, training data) by the WMT, IWSLT and QT Marathon events and, later on, extended with information on the results of their respective evaluation and benchmarking campaigns (documentation, performance of the systems etc.) will be documented and made available through META-SHARE.

A list of CRACKER resources with brief descriptive information is provided below. This list is only indicative of the resources to be included in CRACKER and more detailed information and descriptions will be provided in the course of the project.

3.3.1 R#1 WMT Test Sets

Resource Name	WMT Test Sets
Resource Type	Corpus
Media Type	Text
Language(s)	The core languages are German-English and Czech-English; other guest language pairs will be introduced in each year. For 2015 the guest language was Romanian. We also included Russian, Turkish and Finnish, with funding from other sources.
License	The source data are crawled from online news sites and carry the respective licensing conditions.
Distribution Medium	Downloadable
Usage	For tuning and testing MT systems.
Size	3000 sentences per language pair, per year.
Description	These are the test sets for the WMT shared translation task. They are small parallel data sets used for testing MT systems, and are typically created by translating a selection of crawled articles from online news sites. WMT15 test sets are available at http://www.statmt.org/wmt15/ WMT16 test sets are available at http://data.statmt.org/wmt16/translation-task/test.tgz

3.3.2 R#2 WMT Translation Task Submissions

Resource Name	WMT Translation Task Submissions
Resource Type	Corpus
Media Type	Text
Language(s)	They match the languages of the test sets.
License	Preferably CC BY 4.0.
Distribution Medium	Downloadable
Usage	Research into MT evaluation. MT error analysis.
Size	The 2015 tarball is 25M The 2016 tarball is 44M
Description	These are the submissions to the WMT translation task from all teams. We create a tarball for use in the metrics task, but it is available for future research in MT evaluation. The WMT15 version is available at http://www.statmt.org/wmt15/ The WMT16 version is available at http://data.statmt.org/wmt16/translation-task/wmt16-submitted-data-v2.tgz

3.3.3 R#3 WMT Human Evaluations

Resource Name	WMT Human Evaluations
Resource Type	Pairwise rankings of MT output.
Media Type	Numerical data (in csv)
Language(s)	N/a
License	Preferably CC BY 4.0
Distribution Medium	Downloadable
Usage	In conjunction with the WMT Translation Task Submissions, this can be used for research into MT evaluation.
Size	For 2014, it was 0.5MB
Description	These are the pairwise rankings of the translation task submissions. The WMT15 versions are available at http://www.statmt.org/wmt15/ The WMT16 versions will be available at http://www.statmt.org/wmt16/. They will be made available in time for the workshop in August 2016.

3.3.4 R#4 WMT News Crawl

Resource Name	WMT News Crawl
Resource Type	Corpus
Media Type	Text
Language(s)	English, German, Czech plus variable guest languages.
License	The source data are crawled from online news sites and carry the respective licensing conditions.
Distribution Medium	Downloadable
Usage	Building MT systems
Size	For 2014, it was 5.3G (compressed) The WMT16 version was 4.8G
Description	This data set consists of text crawled from online news, with the html stripped out and sentences shuffled. For WMT15 it is available at http://www.statmt.org/wmt15/ For WMT16 it is available at http://data.statmt.org/wmt16/translation-task/training-monolingual-news-crawl.tgz

3.3.5 R#5 Quality Estimation Datasets

Resource Name	Quality Estimation Datasets
Resource Type	Bilingual corpora labelled for quality at phrase-level
Media Type	Text
Language(s)	German-English, English-German and one of the challenging language pairs addressed in WMT (either Romanian or Latvian)
License	Creative Commons
Distribution Medium	Downloadable
Usage	Other researchers working on quality estimation or evaluation of machine translation
Size	At least 1,000 machine translations will be annotated for quality to train and test quality estimation systems for each language pair.
Description	The corpus will consist of source segments in English, their machine translation, a segmentation of these translations into phrases and a binary score given by humans indicating the quality of these phrases.

3.3.6 R#6 WMT 2016 Automatic Post-editing data set

Resource Name	WMT 2016 Automatic Post-editing data set
Resource Type	corpus
Media Type	text
Language(s)	English to German
License	TAUS Terms of Use (https://lindat.mff.cuni.cz/repository/xmlui/page/licence-TAUS_QT21). TAUS grants to QT21 User access to the WMT Data Set with the following rights: i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WM T Data Set as if it is its own new translation; ii) the right to make Derivative Works; and iii) the right to use or resell such Derivative Works commercially and for the following goals: i) research and benchmarking; ii) piloting new solutions; and iii) testing of new commercial services.
Distribution Medium	downloadable
Usage	Training of Automatic Post-editing and Quality Estimation components
Size	1294 kb
Description	Training, development and text data (the same used for the Sentence-level Quality Estimation task) consist of English-German triplets (*source*, *target* and *post-edit*) belonging to the Information Technology domain and already tokenized. Training and development respectively contain 12,000 and 1,000 triplets, while the test set contains 2,000 instances. Target sentences are machine-translated with the KIT system. Post-edits are collected by Text&Form from professional translators. All data is provided by the EU project QT21 (http://www.qt21.eu/).

3.4 Standards and Metadata

CRACKER follows META-SHARE's best practices for data documentation. The basic design principles of the META-SHARE model have been formulated according to specific needs identified, namely: (a) a typology for language resources (LR) identifying and defining all types of LRs and the relations between them; (b) a common terminology with as clear semantics as possible; (c) minimal schema with simple structures (for ease of use) but also extensive, detailed schema (for exhaustive description of LRs); (d) interoperability between descriptions of LRs and associated software across repositories.

In answer to these needs, the following design principles were formulated:

- expressiveness, i.e., cover any type of resource;

- extensibility, allowing for future extensions and catering for combinations of LR types for the creation of complex resources;
- semantic clarity, through a bundle of information accompanying each schema element;
- flexibility, by employing both exhaustive and minimal descriptions;
- interoperability, through mappings to widely used schemas (DC, Clarin Concept Registry (which has taken over the ISOcat DCR)).

The central entity of the META-SHARE ontology is the Language Resource. In parallel, LRs are linked to other satellite entities through relations, represented as basic elements. The interconnection between the LR and these satellite entities pictures the LR's lifecycle from production to use: reference documents related to the LR (papers, reports, manuals etc.), persons/organizations involved in its creation and use (creators, distributors etc.), related projects and activities (funding projects, activities of usage etc.), accompanying licenses, etc. CRACKER will follow these standard practices for data documentation, in line with their design principles of expressiveness, extensibility, semantic clarity, flexibility and interoperability.

The META-SHARE metadata can also be represented as linked data following the work being done in Task 3.3 of the CRACKER project, the LD4LT group (https://www.w3.org/community/ld4lt/), and the LIDER project, which has produced an OWL version of the META-SHARE metadata schema (http://purl.org/net/def/metashare). Such representation can be generated by the mapping process initiated by the above tasks and initiatives.

As an example, a subset of the META-SHARE metadata records has been converted to Linked Data and is accessible via the Linghub portal (http://linghub.lider-project.eu).

Included in the conversion process to OWL was the legal rights module of the META-SHARE schema (http://purl.org/NET/ms-rights), taking into account the ODRL model & vocabulary v.2.1 (https://www.w3.org/community/odrl/model/2.1/).

3.5 Data Sharing

As said, resource sharing will build upon META-SHARE. CRACKER will maintain and release an improved version of the META-SHARE software.

For its own data sets, CRACKER will continue to apply, whenever possible, the permissive licensing and open sharing culture which has been one of the key components of META-SHARE for handling research data in the digital age.

Consequently, for the MT/LT research and user communities, sharing of all CRACKER data sets will be organised through META-SHARE. The metadata schema provides components and elements that address copyright and Intellectual Property Rights (IPR) issues, restrictions imposed on data sharing and also IPR holders. These together with an existing licensing toolkit can serve as guidance for the selection of the appropriate licensing solution and creating the respective metadata. In parallel, ELRA/ELDA has recently implemented a licensing wizard[3], helping rights holders in defining and selecting the appropriate license under which

[3] http://wizard.elra.info/

they can distribute their resources. The wizard will be possibly integrated or linked to META-SHARE.

3.6 Archiving and Preservation

All datasets produced will be provided and made sustainable through the existing META-SHARE repositories, or new repositories that partners may choose to set up and link to the META-SHARE network. Datasets will be locally stored in the repositories' storage layer in compressed format.

4. Collaboration with Other Projects and Initiatives

CRACKER has created an umbrella initiative that includes all currently running and recently completed EU-supported projects working on technologies for a multilingual Europe, namely the Cracking the Language Barrier initiative[4]. This federation of projects is set up around a short multi-lateral Memorandum of Understanding (MoU)[5].

The MoU contains a non-exhaustive list of general areas of collaboration, and all projects and organisations that sign this document are invited to participate in these collaborative activities.

At the time of writing (June 2016), the MoU has been signed by 10 organisations and 23 projects (including service contracts):

- *Organisations:* CITIA, CLARIN, ELEN, EFNIL, GALA, LT-Innovate, META-NET, NPLD, TAUS, W3C.
- *Projects:* ABUMATRAN, CRACKER, DLDP, ELRC, EUMSSI, EXPERT, Falcon, FREME, HimL, KConnect, KRISTINA, LIDER, LT_Observatory, MixedEmotions, MLi, MMT, MultiJEDI, MultiSensor, Pheme, QT21, QTLeap, SUMMA, XLiMe

Additional organisations and projects have been approached for participation in the initiative. The group of members is constantly growing.

[4] http://www.cracking-the-language-barrier.eu

[5] http://www.cracking-the-language-barrier.eu/download/mou/

5. Recommendations for Harmonised DMPs for the ICT-17 Federation of Projects

One of the areas of collaboration included in the CRACKER MoU refers to the data management and repositories for data, tools and technologies; thus, all projects and organisations participating in the initiative are invited to join forces and to collaborate on harmonising data management plans (metadata, best practices etc.) as well as data, tools and technologies distribution through open repositories.

At the kick-off meeting of the ICT-17 group of projects on April 28, 2015, CRACKER offered support to the "Cracking the language barrier" federation of projects by proposing a Data Management Plan template with shared key principles that can be applied, if deemed helpful, by all projects, again, advocating an open sharing approach whenever possible (also see D1.2). This plan has been included in the overall communication plan and it will inform the working group that will maintain and update the roadmap for European MT research.

In future face-to-face or virtual meetings of the federation, we propose to discuss the details about metadata standards, licenses, or publication types. Our goal is to prepare a list of planned tangible outcomes of all projects, i.e., all datasets, publications, software packages and any other results, including technical aspects such as data formats. We would like to stress that the intention is not to provide the primary distribution channel for all projects' data sets but to provide, in addition to the channels foreseen in the projects' respective Descriptions of Actions, one additional, alternative common distribution platform and approach for metadata description for all data sets produced by the "Cracking the language barrier" federation of projects.

In this respect, the activities that the participating projects may optionally undertake are the following:

1. Participating projects may consider using META-SHARE as an additional, alternative distribution channel for their tools or data sets, using one of the following options:
   a. projects may set up a project or partner specific META-SHARE repository, and use either open or even restrictive licences;
   b. projects may join forces and set up one dedicated "Cracking the language barrier" META-SHARE repository to host the resources developed by all participating projects, and use either open or even restrictive licences (as appropriate).
2. Participating projects may wish to use the META-SHARE repository software[6] for documenting their resources, even if they do not wish to link to the network.

[6] https://github.com/metashare/META-SHARE

As mentioned above, the collaboration in terms of harmonizing data management plans and recommending distribution through open repositories forms one of the six areas of collaboration indicated in the *"Cracking the Language Barrier" MoU*. Participation in one or more of the potential areas of collaboration in this joint community activity, is optional.

An example of harmonized DMP is that of the FREME project. FREME signed the corresponding Memorandum of Understanding and is participating in this initiative. As part of the effort, FREME will make available its metadata from existing datasets that are used by FREME, using a combined metadata scheme: this covers both the META-SHARE template provided by CRACKER, as well as the DataID schema[7]. FREME will follow both META-SHARE and DataID 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. Further details as well as the actual dataset descriptions have been documented in the FREME Data management Plan[8]. See section 3.1.2 of that plan for an example of the combined approach.

5.1 Recommended Template of a DMP

As pointed out already, the collaboration in terms of harmonizing data management plans is considered an important aspect of convergence within the groups of projects. In this respect, any project that is interested in and intends to collaborate towards a joint approach for a DMP may follow the proposed structure of a DMP template. The following section describes a recommended template, while the previous section (3) has provided a concrete example of such an implementation, i.e. the CRACKER DMP. It is, of course, expected that any participating project may accommodate its DMP content according to project-specific aspects and scope. These DMPs are also expected to be gradually completed as the project(s) progress into their implementation.

I. The ABC Project DMP

- **i. Introduction/ Scope**
- **ii. Data description**
- **iii. Identification mechanism**
- **iv. Standards and Metadata**
- **v. Data Sharing**
- **vi. Archiving and preservation**

[7] See http://wiki.dbpedia.org/projects/dbpedia-dataid

[8] http://www.freme-project.eu/resources/D7.5_DMP_V.3_FINAL.pdf

Figure 3. The recommended template for the implementation and structuring of a DMP.

5.1.1 Introduction and Scope

Overview and approach on the resource sharing activities underpinning the language technology and machine translation research and development within each participating project and as part of the "Cracking the language barrier" initiative of projects.

5.1.2 Dataset Reference and Name

It is recommended that a standard identification mechanism should be employed for each data set, e.g., (a) a PID (Persistent Identifier as a long-lasting reference to a dataset) or (b) ISLRN (International Standard Language Resource Number).

5.1.3 Dataset Description

It is recommended that the following resource and media types are addressed:

- **corpora** (text, audio, video, multimodal/multimedia corpora, n-gram resources),
- **lexical/conceptual resources** (e.g., computational lexicons, ontologies, machine-readable dictionaries, terminological resources, thesauri, multimodal/ multimedia lexicons and dictionaries, etc.)
- **language descriptions** (e.g., computational grammars)
- **technologies** (tools/services) that can be used for the processing of data resources

In relation to the resource identification of the "Cracking the language barrier" initiative and to have a first rough estimation of their number, coverage and other core characteristics, CRACKER will circulate two templates dedicated to datasets and associated tools and services respectively. Projects that wish and decide to participate in this uniform cataloguing are invited to fill in these templates with brief descriptions of the resources they estimate to be produced and/or collected. The templates are as follows (also in the Appendix):

Resource Name	Complete title of the resource
Resource Type	Choose one of the following values: Lexical/conceptual resource, corpus, language description (missing values can be discussed and agreed upon with CRACKER)
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 LR 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, hard copy 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 resource (including url, if any)

Table 1. Template for datasets description

Technology Name	Complete title of the tool/service/technology
Technology Type	Tool, service, infrastructure, platform, etc.
Technology Type	The function of the tool or service, e.g., parser, tagger, annotator, corpus workbench etc.
Media Type	The physical medium of the content representation, e.g., video, image, text, numerical data, n-grams, etc.
Language(s)	The language(s) that the tool/service operates on
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 tool/service, e.g., accessible through interface, downloadable, CD/DVD, etc.
Usage	Foreseen use of the tool/service for which it has been produced
Description	A brief description of the main features of the tool/service

Table 2. Template for technologies description

5.1.4 Standards and Metadata

Participating projects are recommended to deploy the META-SHARE metadata schema for the description of their resources and provide all details regarding their name, identification, format, etc.

Providers of resources wishing to participate in the initiative will be able to request and get assistance through dedicated helpdesks on questions concerning (a) the metadata based LR documentation at helpdesk-metadata@meta-share.eu (b) the use of licences, rights of use, IPR issues, etc. at helpdesk-legal@meta-share.eu and (c) the repository installation and use at helpdesk-technical@meta-share.eu.

5.1.5 Data Sharing

It is recommended that all datasets (including all relevant metadata records) to be produced by the participating projects will be made available under licenses, which are as open and as standardised as possible, as well as established as best practice. as Any interested provider can consult the META-SHARE licensing options and pose related questions to the respective helpdesk.

5.1.6 Archiving and Preservation

As regards the procedures for long-term preservation of the datasets, two options may be considered:

1. As part of the further development and maintenance of the META-SHARE infrastructure, a project that participates in the “Cracking the language barrier” initiative may opt to set up its own project or partner specific META-SHARE repository and link to the META-SHARE network, with CRACKER providing all support necessary in the installation, configuration and set up process.
2. Alternatively, one dedicated “Cracking the language barrier” META-SHARE repository can be set up to host the resources developed by all participating projects, with CRACKER catering for procedures and mechanisms enabling long-term preservation of the datasets.

It should be repeated at this point that following the META-SHARE principles, the curation and preservation of the datasets, together with the rights of their use and possible restrictions, are under the sole control and responsibility of the data providers.

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

McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodriguez Doncel, V. and Cimiano, P. (2015) One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. 4th Workshop on the Multilingual Semantic Web, (accepted).

Gavrilidou, M., Labropoulou, E., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., Mapelli, V. (2012) The META-SHARE Metadata Schema for the Description of Language Resources. In Calzolari et al. (ed.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA).

Labropoulou, P., Desipri, E. (ed.) March 2012. Documentation and User Manual of the META-SHARE Metadata Model. Available at: http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf

Piperidis, S., Papageorgiou, H. Spurk, C., Rehm, G., Choukri, K., Hamon, O., Calzolari, N., Del Gratta, R., Magnini, B., Girardi, C. (2014) META-SHARE: One Year After. In Calzolari et al. (ed.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA).

Piperidis, S. (2012) The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Calzolari et al. (ed.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA).

Rehm, G., Hajic, J., van Genabith, J., Vasiljevs, A. (2016) Fostering the Next Generation of European Language Technology: Recent Developments — Emerging Initiatives — Challenges and Opportunities. In Calzolari et al. (ed.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA).

Appendix

Recommended templates for the description of the resources to be collected (to be filled in by each participating project).

Template for Datasets

Resource Name	Complete title of the resource
Resource Type	Choose one of the following values: Lexical/conceptual resource, corpus, language description (missing values can be discussed and agreed upon with CRACKER)
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 LR 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, hard copy 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 resource (including url, if any)

Template for Tools/Services

Technology Name	Complete title of the tool/service/technology
Technology Type	Tool, service, infrastructure, platform, etc.
Technology Type	The function of the tool or service, e.g., parser, tagger, annotator, corpus workbench etc.
Media Type	The physical medium of the content representation, e.g., video, image, text, numerical data, n-grams, etc.
Language(s)	The language(s) that the tool/service operates on
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 tool/service, e.g., accessible through interface, downloadable, CD/DVD, etc.
Usage	Foreseen use of the tool/service for which it has been produced
Description	A brief description of the main features of the tool/service