



## Deliverable 3.1

# Data Management Plan (Initial Version)

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

**Dissemination Level:** Public

**Date:** 30 June 2015



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	30/06/2015
Actual date of delivery	30/06/2015
Deliverable number	D3.1
Deliverable title	Data Management Plan (Initial Version)
Type	Report
Status and version	Final (version 1.0)
Number of pages	20
WP leader	ATHENA RC
Task leader	ATHENA RC
Author(s)	Maria Koutsombogera (ATHENA RC) Stelios Piperidis (ATHENA RC)
Contributor(s)	Georg Rehm (DFKI) Nieves Sande (DFKI) Khalid Choukri (ELDA)
Internal reviewer(s)	Felix Sasaki (DFKI)
EC project officer	Pierre-Paul Sondag
The partners in CRACKER are:	<ul style="list-style-type: none"> <li>• Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany</li> <li>• Charles University in Prague (CUNI), Czech Republic</li> <li>• Evaluations and Language Resources Distribution Agency (ELDA), France</li> <li>• Fondazione Bruno Kessler (FBK), Italy</li> <li>• Athena Research and Innovation Center in Information, Communication and Knowledge Technologies (ATHENA RC), Greece</li> <li>• University of Edinburgh (UEDIN), UK</li> <li>• University of Sheffield (USFD), UK</li> </ul>

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](mailto: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>.

© 2015 CRACKER Consortium

## Contents

<b>History</b>	<b>4</b>
<b>1. Executive Summary</b>	<b>5</b>
<b>2. Background</b>	<b>6</b>
<b>3. The CRACKER DMP</b>	<b>7</b>
3.1 Introduction and Scope	7
3.2 Dataset Reference and Name	7
3.3 Dataset Description	8
3.4 Standards and Metadata	11
3.5 Data Sharing	12
3.6 Archiving and Preservation	12
<b>4. Collaboration with Other Projects and Initiatives</b>	<b>13</b>
<b>5. Recommendations for Harmonised DMPs for the ICT-17 Federation of Projects</b>	<b>14</b>
<b>5.1 Recommended Template of a DMP</b>	<b>15</b>
5.1.1 Introduction and Scope	15
5.1.2 Dataset Reference and Name	16
5.1.3 Dataset Description	16
5.1.4 Standards and Metadata	17
5.1.5 Data Sharing	17
5.1.6 Archiving and Preservation	17
<b>References</b>	<b>19</b>
<b>Appendix</b>	<b>20</b>

**History**

Version	Date	Status	Notes
0.5	20/04/2015	Internal	Working version
0.7	28/05/2015	Internal	Revisions following the Riga meeting with representatives of ICT-17 projects
0.9	16/06/2015	Internal	Document structure revised after consortium comments
1.0	30/06/2015	Public	First public version after internal review

## 1. Executive Summary

This document describes the Data Management Plan (DMP) to be adopted within CRACKER and provides information on CRACKER's data management policy and key information on all datasets to 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 first version includes the principles according to which the plan is structured and the standard practices for data management that will be implemented. Updates of the CRACKER DMP document will be provided in M18 (June 2016) and M36 (December 2017) respectively. In these next versions, more detailed information on the actual datasets and their management will be provided.

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<sup>1</sup>:

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.

---

<sup>1</sup> See details at:

[http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-data-mgt\\_en.pdf](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<sup>2</sup>.

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 3.0.1 version, released in January 2013. 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, will provide 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 will opt for a standard identification mechanism to be employed for each data set, in addition to the identifier used internally by META-SHARE itself. The

---

<sup>2</sup> [http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-pilot-guide\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf)

## D3.1: Data Management Plan (Initial Version)

options that will be addressed for the reference to the dataset ID are the use of either a PID (Persistent Identifier as a long-lasting reference to a dataset) or the ISLRN ([International Standard Language Resource Number](http://www.islrn.org/)), the most recent universal identification schema for LRs which provides LRs with unique names using a standardized nomenclature, ensuring that LRs are identified, and consequently recognized with proper references (cf. figures 1 and 2).

islrn


INTERNATIONAL STANDARD LANGUAGE RESOURCE NUMBER

[Home](#) | [Identify](#) | [Know More](#) | [Login](#) | [Sign up](#)

**Resource: Coordination Annotation for the Penn Treebank**

Reference	Coordination Annotation for the Penn Treebank
Date of Submission	May 20, 2015, 12:06 p.m.
Status	accepted
ISLRN	060-785-139-403-2
Resource Type	Primary Text
Media Type	Text
Source	<a href="https://catalog.ldc.upenn.edu/LDC2015T08">https://catalog.ldc.upenn.edu/LDC2015T08</a>
Language	English
Format/MIME	text/plain
Type	
Size	19528 KB
Access Medium	Web Download
Description	<p><b>Introduction</b> Coordination Annotation for the Penn Treebank is a stand-off annotation for the Wall Street Journal portion of Treebank-3 (PTB3) (LDC99T42) developed by researchers at the University of Düsseldorf and Indiana University. It marks all tokens that have a coordinating function (potentially among other functions). Coordination is a syntactic structure that links together two or more elements known as conjuncts or conjoints. The presence of coordination is often signaled by the appearance of a coordinator (coordinating conjunction), such as and, or, but in English. Penn Coordination Annotation is available at no cost to all licensees of PTB3 and appears in their download queue associated with LDC99T42 as penn_coordination_anno_LDC2015T08.tgz.</p> <p><b>Data</b> This annotation is presented in a single UTF-8 plain text tab file with columns as follows: section: Penn Treebank WSJ section number file: Number of file within section sentence: Number of sentence (starting with 0) token: Number of token (starting with 0) annotation: "P" if the token is a coordinating punctuation, "O" otherwise</p>
Version	1.0
Creator	Sandra Kübler, Wolfgang Maier, Erhard Hinrichs
Distributor	Linguistic Data Consortium

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/>.



European Language Resources Association

Home > [Language Resources](#) > [Multilingual lexicons](#)

Language Resources

Spoken Resources

Written Resources

Terminological Resources

Multimedia/Multimedia Resources

Blog reports

Send us your blog reports

Search Catalogue

Use keywords to find the product you are looking for. Advanced Search

Language(s): English


ISLRN: 952-246-779-755-9

Language(s): English <<<<<< Russian

ISLRN: 009-573-372-552-7

Language(s): English >>>>> Italian

ISLRN: 138-251-030-365-3



Linguistic Data Consortium

Home > [Language Resources](#) > [Data](#)

ABOUT

MEMBERS

COMMUNICATION

LANGUAGE RESOURCES

Log in or Register

The New York Times Annotated Corpus

Item Name: The New York Times Annotated Corpus

Author(s): Evan Sandhaus

LDC Catalog No.: LDC2008T19

ISBN: 1-55583-488-7

ISLRN: 429-488-225-160-9

Release Date: October 17, 2008

Member Year(s): 2008

DCMI Type(s): Text

Data Source(s): newswire

Application(s): summarization, metadata extraction, information retrieval, information extraction

Language(s): English

Language ID(s): eng

License(s): The New York Times Annotated Corpus Agreement

Online Documentation: LDC2008T19 Documents

Licensing Instructions: Subscription & Standard Members, and Non-Members

Citation: Sandhaus, Evan. The New York Times Annotated Corpus LDC2008T19. DVD. Philadelphia: Linguistic Data Consortium, 2008.

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 will address 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 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 preliminary 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.

R#1

<b>Resource Name</b>	WMT Test Sets
<b>Resource Type</b>	Corpus
<b>Media Type</b>	Text
<b>Language(s)</b>	The core languages are German-English and Czech-English; other guest language pairs will be introduced in each year.
<b>License</b>	The source data are crawled from online news sites and carry the respective licensing conditions.
<b>Distribution Medium</b>	Downloadable
<b>Usage</b>	For tuning and testing MT systems.
<b>Size</b>	3000 sentences per language pair, per year. We typically have 5 language pairs (not all funded by cracker).
<b>Description</b>	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. They are made available from the appropriate WMT website (i.e. <a href="http://www.statmt.org/wmt15/">http://www.statmt.org/wmt15/</a> for 2015)

R#2

<b>Resource Name</b>	WMT Translation Task Submissions
<b>Resource Type</b>	Corpus
<b>Media Type</b>	Text
<b>Language(s)</b>	They match the languages of the test sets.

<b>License</b>	Preferably CC BY 4.0.
<b>Distribution Medium</b>	Downloadable
<b>Usage</b>	Research into MT evaluation. MT error analysis.
<b>Size</b>	The 2015 tarball is 25M
<b>Description</b>	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. Again it is available from the WMT website ( <a href="http://www.statmt.org/wmt15/">http://www.statmt.org/wmt15/</a> )

## R#3

<b>Resource Name</b>	WMT Human Evaluations
<b>Resource Type</b>	Pairwise rankings of MT output.
<b>Media Type</b>	Numerical data (in csv)
<b>Language(s)</b>	N/a
<b>License</b>	Preferably CC BY 4.0
<b>Distribution Medium</b>	Downloadable
<b>Usage</b>	In conjunction with the WMT Translation Task Submissions, this can be used for research into MT evaluation.
<b>Size</b>	For 2014, it was 0.5MB
<b>Description</b>	These are the pairwise rankings of the translation task submissions. They will also be available from the WMT website (e.g., <a href="http://www.statmt.org/wmt15/">http://www.statmt.org/wmt15/</a> )

## R#4

<b>Resource Name</b>	WMT News Crawl
<b>Resource Type</b>	Corpus
<b>Media Type</b>	Text
<b>Language(s)</b>	English, German, Czech plus variable guest languages.
<b>License</b>	The source data are crawled from online news sites and carry the respective licensing conditions.
<b>Distribution Medium</b>	Downloadable
<b>Usage</b>	Building MT systems
<b>Size</b>	For 2014, it was 5.3G (compressed)

<b>Description</b>	<p>This data sets consists of text crawled from online news, with the html stripped out and sentences shuffled.</p> <p>They will also be available from the WMT website (e.g., <a href="http://www.statmt.org/wmt15/">http://www.statmt.org/wmt15/</a>)</p>
--------------------	---

### 3.4 Standards and Metadata

CRACKER will follow 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 schemas with simple structures (for ease of use) but also extensive, detailed schemas (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, 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. 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; accessible via the Linghub portal (<http://linghub.lider-project.eu>).

Included in the conversion process to OWL<sup>3</sup> was the legal rights module of the META-SHARE schema, taking into account the ODRL model & vocabulary v.2.1 (<https://www.w3.org/community/odrl/model/2.1/>).

---

<sup>3</sup> <https://github.com/ld4lt/metashare>

### 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<sup>4</sup>, helping rights holders in defining and selecting the appropriate license under which 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.

---

<sup>4</sup> <http://wizard.elra.info/>

---

## 4. Collaboration with Other Projects and Initiatives

CRACKER will pursue close collaboration with the Coordination and Support Action project LT-Observatory in coordinating their respective activities regarding documentation, sharing, annotation and filtering of machine translation related language resources.

The two projects have planned to use the META-SHARE and CLARIN infrastructures respectively. META-SHARE/META-NET and CLARIN have a long standing Collaboration Agreement, which was initially realised in terms of building bridges and mapping services between their metadata models, the META-SHARE MD schema<sup>5</sup> and the CLARIN CMDI<sup>6</sup>. Furthermore, the two infrastructures can now engage in mutual harvesting of their metadata inventories using standard protocols that have now been implemented by both of them.

In parallel, the two-year service contract CEF.AT, which aims at the collection of data produced by public sector bodies in the EU for the CEF Automated Translation Digital Infrastructure is another excellent opportunity for collaboration with CRACKER. CRACKER will discuss the possibility of storing or providing links to and curating the open datasets that will be collected within CEF.AT.

---

<sup>5</sup> <http://www.meta-net.eu/meta-share/metadata-schema/>

<sup>6</sup> <http://www.clarin.eu/content/component-metadata>

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

One of CRACKER's main goals is to bring together all actions also funded through H2020-ICT17 ([QT21](#), [HimL](#), [TraMOOC](#), [MMT](#), [LT\\_Observatory](#)), including the FP7 project [QT-Leap](#) and related other projects (the “Cracking the language barrier” federation of projects), and to find synergies and establish information channels between them, including a suggested approach towards harmonised Data Management Plans that share the same set of key principles.

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 will be 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<sup>7</sup> for documenting their resources, even if they do not wish to link to the network.

---

<sup>7</sup> <https://github.com/metashare/META-SHARE>

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 *Multilateral Memorandum of Understanding, “Cracking the Language Barrier”*. This MoU document was initiated by CRACKER upon the decision of the representatives of all European projects funded through Horizon 2020, ICT-17, in Riga, April 2015. All projects have been invited to sign the MoU, whose goal is to establish a federation that contributes to the overall strategic objective of “cracking the language barrier”. Participation in one or more of the potential areas of collaboration in this joint community activity, is optional.

### 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.

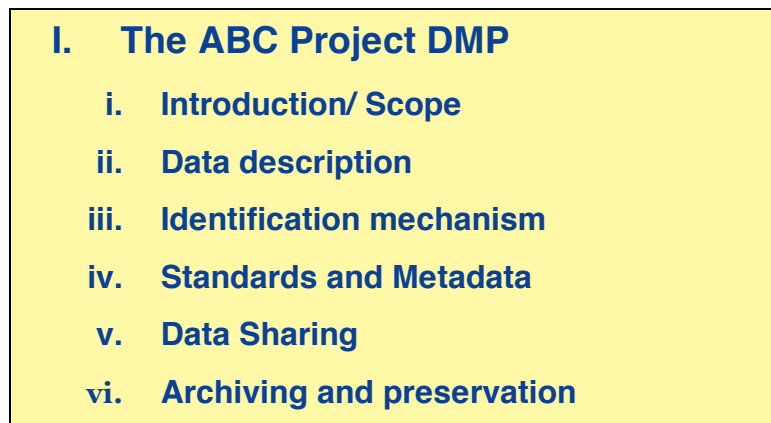


Figure 3. The recommended template for the implementation and structuring of a DMP.

#### 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) [ISLRLN](#) (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):

<b>Resource Name</b>	Complete title of the resource
<b>Resource Type</b>	Choose one of the following values: Lexical/conceptual resource, corpus, language description (missing values can be discussed and agreed upon with CRACKER)
<b>Media Type</b>	The physical medium of the content representation, e.g., video, image, text, numerical data, n-grams, etc.
<b>Language(s)</b>	The language(s) of the resource content
<b>License</b>	The licensing terms and conditions under which the LR can be used
<b>Distribution Medium</b>	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.
<b>Usage</b>	Foreseen use of the resource for which it has been produced
<b>Size</b>	Size of the resource with regard to a specific size unit measurement in form of a number
<b>Description</b>	A brief description of the main features of the resource (including url, if any)

**Table 1. Template for datasets description**



<b>Technology Name</b>	Complete title of the tool/service/technology
<b>Technology Type</b>	Tool, service, infrastructure, platform, etc.
<b>Technology Type</b>	The function of the tool or service, e.g., parser, tagger, annotator, corpus workbench etc.
<b>Media Type</b>	The physical medium of the content representation, e.g., video, image, text, numerical data, n-grams, etc.
<b>Language(s)</b>	The language(s) that the tool/service operates on
<b>License</b>	The licensing terms and conditions under which the tool/service can be used
<b>Distribution Medium</b>	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.
<b>Usage</b>	Foreseen use of the tool/service for which it has been produced
<b>Description</b>	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](mailto:helpdesk-metadata@meta-share.eu) (b) the use of licences, rights of use, IPR issues, etc. at [helpdesk-legal@meta-share.eu](mailto:helpdesk-legal@meta-share.eu) and (c) the repository installation and use at [helpdesk-technical@meta-share.eu](mailto: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. 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](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](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).

## Appendix

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

### Template for Datasets

<b>Resource Name</b>	Complete title of the resource
<b>Resource Type</b>	Choose one of the following values: Lexical/conceptual resource, corpus, language description (missing values can be discussed and agreed upon with CRACKER)
<b>Media Type</b>	The physical medium of the content representation, e.g., video, image, text, numerical data, n-grams, etc.
<b>Language(s)</b>	The language(s) of the resource content
<b>License</b>	The licensing terms and conditions under which the LR can be used
<b>Distribution Medium</b>	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.
<b>Usage</b>	Foreseen use of the resource for which it has been produced
<b>Size</b>	Size of the resource with regard to a specific size unit measurement in form of a number
<b>Description</b>	A brief description of the main features of the resource (including url, if any)

### Template for Tools/Services

<b>Technology Name</b>	Complete title of the tool/service/technology
<b>Technology Type</b>	Tool, service, infrastructure, platform, etc.
<b>Technology Type</b>	The function of the tool or service, e.g., parser, tagger, annotator, corpus workbench etc.
<b>Media Type</b>	The physical medium of the content representation, e.g., video, image, text, numerical data, n-grams, etc.
<b>Language(s)</b>	The language(s) that the tool/service operates on
<b>License</b>	The licensing terms and conditions under which the tool/service can be used
<b>Distribution Medium</b>	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.
<b>Usage</b>	Foreseen use of the tool/service for which it has been produced
<b>Description</b>	A brief description of the main features of the tool/service