

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 3.9

Data Management Plan (Final Version)

Authors: Kanella Pouli (Athena RC),
Stelios Piperidis (Athena RC)

Dissemination Level: Public

Date: 31/12/2017



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	31/12/2017
Actual date of delivery	31/12/2017
Deliverable number	D3.9
Deliverable title	Data Management Plan (Update)
Type	Report
Status and version	Final
Number of pages	33
WP leader	ATH
Task leader	ATH
Author(s)	Kanella Pouli (ATH), Stelios Piperidis (ATH)
Contributor(s)	DFKI, CUNI, FBK, UEDIN, USDF, ELDA
Internal reviewer(s)	Georg Rehm
EC project officer	Pierre-Paul Sondag (M1-M18), Susan Fraser (M19-M36)
The partners in CRACKER are:	<ul style="list-style-type: none"> • 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>.
 © 2017 CRACKER Consortium

Contents

1. Executive Summary	6
2. Background	7
3. The CRACKER DMP	8
3.1 Introduction and Scope	8
3.2 Dataset Reference and Name	8
3.3 Dataset Description	9
3.3.1 R#1 WMT 2015 Test Sets	10
3.3.2 R#2 WMT 2016 Test Sets	10
3.3.3 R#3 WMT 2017 Test Sets	11
3.3.4 R#4 WMT 2015 Translation Task Submissions	12
3.3.5 R#5 WMT 2016 Translation Task Submissions	12
3.3.6 R#6 WMT 2017 Translation Task Submissions	13
3.3.7 R#7 WMT 2015 Human Evaluations	13
3.3.8 R#8 WMT 2016 Human Evaluations	13
3.3.9 R#9 WMT 2017 Human Evaluations	14
3.3.10 R#10 WMT 2015 News Crawl	14
3.3.11 R#11 WMT 2016 News Crawl	15
3.3.12 R#12 WMT 2017 News Crawl	15
3.3.13 R#13 Quality Estimation Datasets	15
3.3.14 R#14 WMT 2016 Automatic Post-editing data set	16
3.3.15 R#15 WMT 2017 Automatic Post-editing data set	17
3.3.16 R#16 WMT 2018 Automatic Post-editing data set	18
3.3.17 R#17 QT21 Domain Specific Human Post-Edited data set	18
3.3.18 R#18 QT21 Domain Specific Human Error-Annotated data set	19
3.3.19 R#19 QT21 WMT17 Human Post-Edited data set	20
3.3.20 R#20 QT21 WMT17 Human Error Annotated data set	21
3.3.21 R#21 IWSLT 2015 Data Sets	22
3.3.22 R#22 IWSLT 2016 Data Sets	22
3.3.23 R#23 IWSLT 2017 Data Sets	23
3.3.24 R#24 IWSLT 2015 Human Post-Editing data	23
3.3.25 R#25 IWSLT 2016 Human Post-Editing data	24

3.3.26 R#26 IWSLT 2017 Human Post-Editing data	24
3.4 Standards and Metadata	25
3.5 Data Sharing	26
3.6 Archiving and Preservation	26
<u>4. Collaboration with Other Projects and Initiatives</u>	<u>27</u>
<u>5. Recommendations for Harmonised DMPs for the ICT-17 Federation of Projects</u>	<u>28</u>
5.1 Recommended Template of a DMP	29
5.1.1 Introduction and Scope	29
5.1.2 Dataset Reference and Name	29
5.1.3 Dataset Description	30
5.1.4 Standards and Metadata	31
5.1.5 Data Sharing	31
5.1.6 Archiving and Preservation	31
<u>References</u>	<u>32</u>
<u>Appendix</u>	<u>33</u>

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	22/12/2017	Public	Updated Final Version

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 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 final 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 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 DMP 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 categories¹:

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 [here](#).

3. The CRACKER DMP

3.1 Introduction and Scope

For its own datasets, CRACKER follows [META-SHARE](#)'s 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](#).

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 3.1.1 version, released in December 2016. It currently features 28 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 has built upon and extended the existing META-SHARE resource infrastructure, its specific [MT-dedicated repository](#) 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

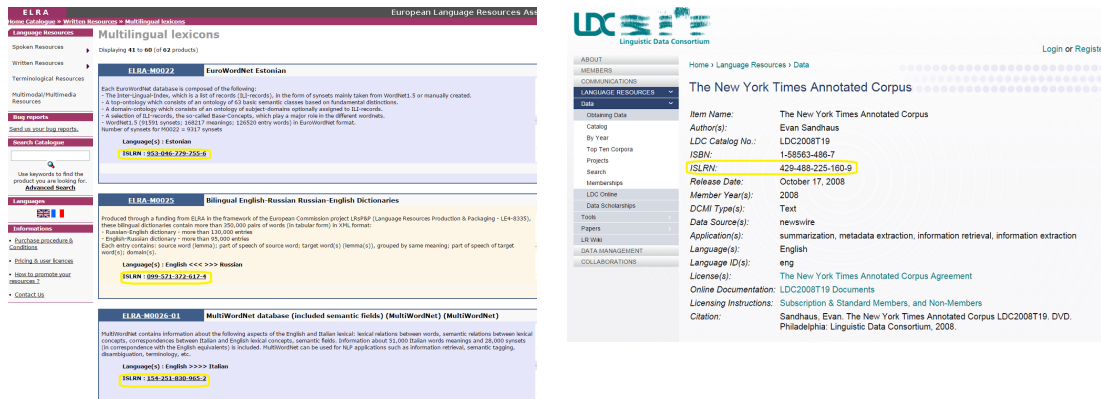
CRACKER opts for a standard identification mechanism to be employed for each data set, in addition to the identifier used internally by META-SHARE itself. Reference to a dataset ID can be optionally made with the use of an ISLRN ([International Standard Language Resource Number](#)), the most recent 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).



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	https://catalog.ldc.upenn.edu/LDC2015T08
Language	English
Format/MIME Type	text/plain
Size	19528 KB
Access Medium	Web Download
Description	<p>Introduction</p> <p>Coordination Annotation for the Penn Treebank is a stand-off annotation for the Wall Street Journal portion of Treebank-3 (PTB3) (LDC99742) developed by researchers at the University of Dusseldorf 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 conjoins. The presence of coordination is often signaled by the appearance of a coordinator (coordinating conjunction), such as and, or, but in English.</p> <p>Penn Coordination Annotation is available at no cost to all licensees of PTB3 and appears in their download queue associated with LDC99742 as penn_coordination_anno_LDC2015T08.tgz.</p> <p>Data</p> <p>This annotation is presented in a single UTF-8 plain text tsv file with columns as follows:</p> <p>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.



The left screenshot shows the ELRA website with three resource entries:

- ELRA-M0022 EuroWordNet Estonian**: A top-ontology which consists of an ontology of 43,144 semantic classes based on fundamental distinctions. ISLRN: 353-016-729-255-6
- ELRA-M0025 Bilingual English-Russian Russian-English Dictionaries**: Produced through a funding from ELRA in the framework of the European Commission project LDC4P (Language Resources Production & Packaging - LE4-0335). ISLRN: 099-571-322-617-1
- ELRA-M0026-01 MultiWordNet database (included semantic fields) (MultiWordNet) (MultiWordNet)**: MultiWordNet contains information about the following aspects of the English and Italian lexical, lexicol-relations between words, semantic relations between lexical concepts, correspondences between Italian and English lexical concepts, semantic fields. ISLRN: 154-251-830-955-2

The right screenshot shows the LDC website entry for **The New York Times Annotated Corpus** with the following details:

- Item Name: The New York Times Annotated Corpus
- Author(s): Evan Santhaus
- LDC Catalog No.: LDC2008T19
- ISBN: 1-58563-486-7
- 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 ontology, CRACKER has been 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 produced (test data, training data) by the WMT, IWSLT and QT Marathon events and extended with information on the results of their respective evaluation and benchmarking campaigns (documentation, performance of the systems etc.) are documented and made available through META-SHARE.

A brief description of all the resources generated by CRACKER, or with the support of CRACKER, and in coordination with project QT21, is provided below.

3.3.1 R#1 WMT 2015 Test Sets

Resource Name	WMT 2015 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/

3.3.2 R#2 WMT 2016 Test Sets

Resource Name	WMT 2016 Test Sets
Resource Type	Corpus
Media Type	Text

Language(s)	<p>Cracker has contributed to the German-English and Czech-English test sets from 2015 to 2018², as well as a different guest language in each of these years.</p> <p>The guest language pairs for 2016 were Romanian-English.</p> <p>We also included Russian, Turkish, Chinese, Estonian and Kazakh with funding from other sources, as well as Finnish in 2016.</p>
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	<p>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.</p> <p>WMT16 test sets are available at http://data.statmt.org/wmt16/translation-task/test.tgz</p>

3.3.3 R#3 WMT 2017 Test Sets

Resource Name	WMT 2017 Test Sets
Resource Type	Corpus
Media Type	Text
Language(s)	<p>Cracker has contributed to the German-English and Czech-English test sets from 2015 to 2018³, as well as a different guest language in each of these years. The guest language pairs for 2017 were Latvian-English (2017).</p> <p>We also included Russian, Turkish, Chinese, Estonian and Kazakh with funding from other sources, as well as Finnish in 2017.</p>
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

² The 2018 test sets have not yet been made available.

³ The 2018 test sets have not yet been made available.

	<p>created by translating a selection of crawled articles from online news sites.</p> <p>WMT17 test sets are at http://data.statmt.org/wmt17/translation-task/test.tgz</p>
--	---

3.3.4 R#4 WMT 2015 Translation Task Submissions

Resource Name	WMT 2015 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	25M (compressed text)
Description	<p>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.</p> <p>The WMT15 version is available at http://www.statmt.org/wmt15/wmt15-submitted-data.tgz</p>

3.3.5 R#5 WMT 2016 Translation Task Submissions

Resource Name	WMT 2016 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	44M (compressed text)
Description	<p>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.</p> <p>The WMT16 version is available at http://data.statmt.org/wmt16/translation-task/wmt16-submitted-data-v2.tgz</p>

3.3.6 R#6 WMT 2017 Translation Task Submissions

Resource Name	WMT 2017 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	46M (compressed text)
Description	<p>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.</p> <p>The WMT17 version is at http://data.statmt.org/wmt17/translation-task/wmt17-submitted-data-v1.0.tgz</p>

3.3.7 R#7 WMT 2015 Human Evaluations

Resource Name	WMT 2015 Human Evaluations
Resource Type	Pairwise rankings of MT output (2015-2016), and direct assessments (i.e., adequacy and fluency) (2016-2017)
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	50M
Description	<p>Data available here: 2015 – http://www.statmt.org/wmt15/translation-judgements.zip</p>

3.3.8 R#8 WMT 2016 Human Evaluations

Resource Name	WMT 2016 Human Evaluations
Resource Type	Pairwise rankings of MT output (2015-2016), and direct assessments (i.e., adequacy and fluency) (2016-2017)
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	50M (gzipped).
Description	Data available here: 2016 – http://data.statmt.org/wmt16/translation-task/wmt16-translation-judgements.zip 2016 – http://computing.dcu.ie/~ygraham/da-human-judgments.tar.gz

3.3.9 R#9 WMT 2017 Human Evaluations

Resource Name	WMT 2017 Human Evaluations
Resource Type	Pairwise rankings of MT output (2015-2016), and direct assessments (i.e., adequacy and fluency) (2016-2017)
Media Type	Numerical data (in csv); 2017 with full output (texts).
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	60MB (gzipped).
Description	Data available here: http://computing.dcu.ie/~ygraham/newstest2017-system-level-human.tar.gz http://www.statmt.org/wmt17/results.html

3.3.10 R#10 WMT 2015 News Crawl

Resource Name	WMT 2015 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	5.2Gb

Description	This data set consists of text crawled from online news, with the html stripped out and sentences shuffled. 2015 – http://www.statmt.org/wmt15/training-monolingual-news-2014.v2.tgz
--------------------	---

3.3.11 R#11 WMT 2016 News Crawl

Resource Name	WMT 2016 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	4.8Gb
Description	This data set consists of text crawled from online news, with the html stripped out and sentences shuffled. 2016 – http://data.statmt.org/wmt16/translation-task/training-monolingual-news-crawl.tgz

3.3.12 R#12 WMT 2017 News Crawl

Resource Name	WMT 2017 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	3.7Gb
Description	This data set consists of text crawled from online news, with the html stripped out and sentences shuffled. 2017 – http://data.statmt.org/wmt17/translation-task/training-monolingual-news-crawl.tgz

3.3.13 R#13 Quality Estimation Datasets

Resource Name	WMT 2017 Quality Estimation Datasets – phrase-level
----------------------	---

Resource Type	Bilingual corpora labelled for quality at phrase-level
Media Type	Text
Language(s)	German-English
License	<p>TAUS Terms of Use (https://lindat.mff.cuni.cz/repository/xmlui/page/licence-TAUS_QT21).</p> <p>TAUS grants to QT21 User access to the WMT Data Set with the following rights:</p> <ul style="list-style-type: none"> i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WMT 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: <ul style="list-style-type: none"> i) research and benchmarking; ii) piloting new solutions; and iii) testing of new commercial services.
Distribution Medium	Downloadable
Usage	Other researchers working on quality estimation or evaluation of machine translation
Size	7,500 machine translations annotated for quality with binary labels (good/bad) at the phrase-level (67,817 phrases). To be used to train and test quality estimation systems.
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.

3.3.14 R#14 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	<p>TAUS Terms of Use</p> <p>TAUS grants to QT21 User access to the WMT Data Set with the following rights:</p> <ul style="list-style-type: none"> i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WMT 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: <ul style="list-style-type: none"> 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 consist of English-German triplets (<i>source</i> , <i>target</i> and <i>post-edit</i>) 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).

3.3.15 R#15 WMT 2017 Automatic Post-editing data set

Resource Name	WMT 2017 Automatic Post-editing data set
Resource Type	Corpus
Media Type	text
Language(s)	English to German
License	<p>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:</p> <ul style="list-style-type: none"> i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WMT 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: <ul style="list-style-type: none"> 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	<p>For WMT 2017, 11.000 segments have been added to the WMT16 training set (En-De) together with a new test (for 2017) made of 2.000 segments (En-De). Also in 2017, a new language pair has been added: De-En with 25k segments for training, 1k segments for dev, 2k segments for test. Adding the 2016 and 2017 Auto PE data together, we obtain for each language pair a total of 28k segments each, split in: En-De: training set = 23 k, dev set = 1k, test-set16 = 2k, test-set17 = 2k, De-En: training set: 25k, dev-set = 1k, test-set17= 2k</p> <p>Training, development and text data consist of English-German triplets (<i>source</i>, <i>target</i> and <i>post-edit</i>) 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</p>

	system. Post-edits are collected by Text & Form from professional translators. All data is provided by the EU project QT21 (http://www.qt21.eu/).
--	---

3.3.16 R#16 WMT 2018 Automatic Post-editing data set

Resource Name	WMT 2018 Automatic Post-editing data set
Resource Type	Corpus
Media Type	text
Language(s)	English to German
License	<p>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:</p> <ul style="list-style-type: none"> i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WMT 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: <ul style="list-style-type: none"> 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	<p>For WMT2018 we have added a new test set of 2.000 segments for each of the 2 language pairs from 2017 (en-de and de-en). Each language pair covers 30k segments. The split is: En-De: training set = 23 k, dev set = 1k, test-set16 = 2k, test-set17 = 2k, test-set18= 2k, De-En: training set: 25k, dev-set = 1k, test-set17= 2k, test-set18 = 2k.</p> <p>Training, development and text data 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/).</p>

3.3.17 R#17 QT21 Domain Specific Human Post-Edited data set

Resource Name	QT21 Domain Specific Human Post-Edited data set
Resource Type	corpus
Media Type	text

Language(s)	English to German, English to Czech, English to Latvian, German to English
License	<p>QT21-TAUS Terms of Use (https://lindat.mff.cuni.cz/repository/xmlui/page/licence-TAUS_QT21).</p> <p>TAUS grants to QT21 User access to the WMT Data Set with the following rights:</p> <ul style="list-style-type: none"> i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WMT 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: <ul style="list-style-type: none"> 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 / Quality Estimation / Error Analysis
Size	70 MB
Description	<p>Set of 165,000 domain specific Human Post Edited (HPE) triplets for 4 language pairs and 6 translation engines. Each triplet consists in (source, reference, HPE). The domain for En-De and En-Cz is IT, the domain for En-Lv and De-En is Pharma. A total of 6 translation engines have been used to produce the targets that have been post edited: PBMT and NMT from KIT for En-De, PBMT from KIT for De-En, PBMT from CUNI for En-Cz and both PBMT and NMT system from Tilde for En-Lv. For each language pair, one unique set of source segments has been used as input to the different translation engines. Each translation engine has provided 30,000 target segments except for the two En-Lv engines which have provided 22,500 target segments each. En-De and De-En HPEs have been collected by professional translators from Text&Form. En-Lv HPEs have been collected by professional translators from Tilde. En-Cz HPEs have been collected by professional translators from Traductera. All data is provided by the EU project QT21 (http://www.qt21.eu).</p>

3.3.18 R#18 QT21 Domain Specific Human Error-Annotated data set

Resource Name	QT21 Domain Specific Human Error Annotated data set
Resource Type	corpus
Media Type	text
Language(s)	English to German, English to Czech, English to Latvian, German to English
License	<p>QT21-TAUS Terms of Use (https://lindat.mff.cuni.cz/repository/xmlui/page/licence-TAUS_QT21).</p> <p>TAUS grants to QT21 User access to the WMT Data Set with the following rights:</p> <ul style="list-style-type: none"> i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WMT Data Set as if it is its own new translation;

	<ul style="list-style-type: none"> ii) the right to make Derivative Works; and iii) the right to use or resell such Derivative Works commercially and for the following goals: <ul style="list-style-type: none"> 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 / Quality Estimation / Error Analysis
Size	39 MB
Description	<p>Set of 14,000 domain specific Human Error Annotated (HEA) quadruplets for 4 language pairs and 6 translation engines. Each quadruplet consists in (source, reference, HPE, HEA). The domain for En-De and En-Cz is IT, the domain for En-Lv and De-En is Pharma. This HEA data set is based on the HPE in Section 3.3.15. A total of 6 translation engines have been used to produce the targets that have been post-edited: PBMT and NMT from KIT for En-De, PBMT from KIT for De-En, PBMT from CUNI for En-Cz and both PBMT and NMT system from Tilde for En-Lv. For each language pair, one unique set of source segments has been used as input to the different translation engines. From each translation engine, 2.000 target segments have been error-annotated. From each subset of 2.000 HEA segments, 200 are annotated by 2 different professional translator. En-De and De-En HEAs have been collected by professional translators from Text & Form. En-Lv HEAs have been collected by professional translators from Tilde. En-Cz HEAs have been collected by professional translators from Aspena. All data is provided by the EU project QT21 (http://www.qt21.eu/).</p>

3.3.19 R#19 QT21 WMT17 Human Post-Edited data set

Resource Name	QT21 WMT Human Post-Edited data set
Resource Type	corpus
Media Type	text
Language(s)	English to German, English to Czech, English to Latvian
License	<p>QT21-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:</p> <ul style="list-style-type: none"> i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WMT 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: <ul style="list-style-type: none"> 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 / Quality Estimation / Error Analysis
Size	10,800 Human Post Edited (HPE) triplets (for 3 language pairs)
Description	Set of 10,800 Human Post Edited (HPE) triplets for 3 language pairs on WMT17 news task data. Each triplet consists in (source, reference, HPE). For each language pair, the target segments have been produced on the WMT17 news task by the 3 best WMT17 systems in their respective language pair. Each translation engine has provided 1,200 segments. Translations (targets) have been generated using, “1 62.0 0.308 uedin-nmt”, “3 55.9 0.111 limsi-factored-norm”, “54.1 0.050 CU-Chimera” for En-Cz, “69.8 0.139 uedin-nmt”, “66.7 0.022 KIT”, “66.0 0.003 RWTH-nmt-ensem” for En-De and “54.4 0.196 tilde-nc-nmt-smt”, “50.8 0.075 limsi-fact-norm”, “50.0 0.058 usfd-cons-qt21” for En-Lv. HPEs for En-De have been collected by professional translators from Text&Form. En-Lv HPEs have been collected by professional translators from Tilde. En-Cz HPEs have been collected by professional translators from Traductera. All data is provided by the EU project QT21 (http://www.qt21.eu/).

3.3.20 R#20 QT21 WMT17 Human Error Annotated data set

Resource Name	QT21 WMT Human Error Annotated data set
Resource Type	corpus
Media Type	text
Language(s)	English to German, English to Czech, English to Latvian
License	<p>QT21-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:</p> <ul style="list-style-type: none"> i) the right to use the target side of the translation units into a commercial product, provided that QT21 User may not resell the WMT 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: <ul style="list-style-type: none"> 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 / Quality Estimation / Error Analysis
Size	3,600 quadruplets (for 3 language pairs)
Description	Set of 3,600 WMT17 Human Error Annotated (HEA) quadruplets for 3 language pairs and 9 translation engines. Each quadruplet consists in (source, reference, HPE, HEA). The source data comes from the WMT17 news task. A total of 9 translation engines have been used to produce the targets that have been post edited: Translations (targets) have been generated using, “1 62.0 0.308 uedin-nmt”, “3 55.9 0.111 limsi-factored-norm”, “54.1 0.050 CU-Chimera” for En-Cz, “69.8 0.139

	uedin-nmt”, “66.7 0.022 KIT”, “66.0 0.003 RWTH-nmt-ensem” for En-De and “54.4 0.196 tilde-nc-nmt-smt”, “50.8 0.075 limsi-fact-norm”, “50.0 0.058 usfd-cons-qt21” for En-Lv. From each translation engine, 200 target segments have been post edited which further have been error annotated by 2 different professional translator. En-De HEAs have been collected by professional translators from Text&Form. En-Lv HEAs have been collected by professional translators from Tilde. En-Cz HEAs have been collected by professional translators from Aspensa. All data is provided by the EU project QT21 (http://www.qt21.eu/).
--	---

3.3.21 R#21 IWSLT 2015 Data Sets

Resource Name	IWSLT 2015 Data Sets
Resource Type	Corpus
Media Type	Text
Language(s)	IWSLT 2015: from/to English to/from French, German, Chinese, Thai, Vietnamese, Czech
License	Data are crawled from the TED website and carry the respective licensing conditions.
Distribution Medium	Downloadable
Usage	For training, tuning and testing MT systems.
Size	Approximately, for each language pair, training sets include 2,000 talks, 200K sentences and 4M tokens per side, while each dev and test sets 10-15 talks, 1.0K-1.5K sentences and 20K-30K tokens per side. In each edition, the training sets of previous editions are re-used and updated with new talks added to the TED repository in the meanwhile.
Description	These are the data sets for the MT tasks of the evaluation campaigns of IWSLT. They are parallel data sets used for building and testing MT systems. They are publicly available through the WIT3 website http://wit3.fbk.eu , see release: 2015-01

3.3.22 R#22 IWSLT 2016 Data Sets

Resource Name	IWSLT 2016 Data Sets
Resource Type	Corpus
Media Type	Text
Language(s)	IWSLT 2016: from/to English to/from Arabic, Czech, French, German
License	Data are crawled from the TED website and carry the respective licensing conditions.
Distribution Medium	Downloadable
Usage	For training, tuning and testing MT systems.

Size	Approximately, for each language pair, training sets include 2,000 talks, 200K sentences and 4M tokens per side, while each dev and test sets 10-15 talks, 1.0K-1.5K sentences and 20K-30K tokens per side. In each edition, the training sets of previous editions are re-used and updated with new talks added to the TED repository in the meanwhile.
Description	These are the data sets for the MT tasks of the evaluation campaigns of IWSLT. They are parallel data sets used for building and testing MT systems. They are publicly available through the WIT3 website http://wit3.fbk.eu , see release: 2016-01

3.3.23 R#23 IWSLT 2017 Data Sets

Resource Name	IWSLT 2017 Data Sets
Resource Type	Corpus
Media Type	Text
Language(s)	IWSLT 2017: <ul style="list-style-type: none"> • multilingual: German, English, Italian, Dutch, Romanian • bilingual: from/to English to/from Arabic, German, French, Japanese, Korean, Chinese
License	Data are crawled from the TED website and carry the respective licensing conditions.
Distribution Medium	Downloadable
Usage	For training, tuning and testing MT systems.
Size	Approximately, for each language pair, training sets include 2,000 talks, 200K sentences and 4M tokens per side, while each dev and test sets 10-15 talks, 1.0K-1.5K sentences and 20K-30K tokens per side. In each edition, the training sets of previous editions are re-used and updated with new talks added to the TED repository in the meanwhile.
Description	These are the data sets for the MT tasks of the evaluation campaigns of IWSLT. They are parallel data sets used for building and testing MT systems. They are publicly available through the WIT3 website http://wit3.fbk.eu , see release: 2017-01

3.3.24 R#24 IWSLT 2015 Human Post-Editing data

Resource Name	IWSLT 2015 Human Post-Editing data
Resource Type	corpus
Media Type	text
Language(s)	English to German (EnDe) and Vietnamese to English (ViEn)
License	Post-edits are released under a Creative Commons Attribution (CC-BY) 4.0 International License.
Distribution Medium	downloadable

Usage	Analysis of MT quality and Quality Estimation components
Size	600 segments for EnDe and 500 segments for ViEn (10K tokens each). 5 different automatic translations post-edited by professional translators
Description	<p>The human evaluation (HE) dataset created for EnDe and ViEn MT tasks was a subset of the official test set of the IWSLT 2015 evaluation campaign. The resulting HE sets are composed of 600 segments for EnDe and 500 segments for EnFr, each corresponding to around 10,000 words. Human evaluation was based on Post-Editing, i.e., the manual correction of the MT system output, which was carried out by professional translators. Five primary runs submitted to the evaluation campaign were post-edited for each of the two tasks.</p> <p>Data are publicly available through the WIT3 website http://wit3.fbk.eu, at this page.</p>

3.3.25 R#25 IWSLT 2016 Human Post-Editing data

Resource Name	IWSLT 2016 Human Post-Editing data
Resource Type	corpus
Media Type	text
Language(s)	English to German (EnDe) and English to French (EnFr)
License	Post-edits are released under a Creative Commons Attribution (CC-BY) 4.0 International License.
Distribution Medium	downloadable
Usage	Analysis of MT quality and Quality Estimation components
Size	600 segments for both EnDe and EnFr (10K tokens each). Respectively, 9 and 5 different automatic translations post-edited by professional translators
Description	<p>The human evaluation (HE) dataset created for EnDe and EnFr MT tasks was a subset of one of the official test sets of the IWSLT 2016 evaluation campaign. The resulting HE sets are composed of 600 segments for both EnDe and EnFr, each corresponding to around 10,000 words. Human evaluation was based on Post-Editing, i.e., the manual correction of the MT system output, which was carried out by professional translators. Nine and five primary runs submitted to the evaluation campaign were post-edited for the two tasks, respectively.</p> <p>Data are publicly available through the WIT3 website http://wit3.fbk.eu, at this page.</p>

3.3.26 R#26 IWSLT 2017 Human Post-Editing data

Resource Name	IWSLT 2017 Human Post-Editing data
Resource Type	corpus

Media Type	text
Language(s)	Dutch to German (NIDe) and Romanian to Italian (Rolt)
License	Post-edits will be released under a Creative Commons Attribution (CC-BY) 4.0 International License.
Distribution Medium	will be downloadable
Usage	Analysis of MT quality and Quality Estimation components
Size	603 segments for both NIDe and Rolt (10K tokens each). For each direction, 9 different automatic translations post-edited by professional translators
Description	<p>The human evaluation (HE) dataset created for NIDe and Rolt MT tasks was a subset of the official test set of the IWSLT 2017 evaluation campaign. The resulting HE sets are composed of 603 segments for both NIDe and Rolt, each corresponding to around 10,000 words. Human evaluation was based on Post-Editing, i.e., the manual correction of the MT system output, which was carried out by professional translators. Nine primary runs submitted to the evaluation campaign with engines trained on constrained data conditions and in bilingual/multilingual/zero-shot mode, were post-edited for each of the two tasks.</p> <p>Data will be publicly available through the WIT3 website http://wit3.fbk.eu.</p>

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 has followed 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](#) and the LIDER project, which has produced an [OWL version](#) of the META-SHARE metadata schema. 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.

Included in the conversion process to OWL was the [legal rights](#) module of the META-SHARE schema, taking into account the [ODRL](#) model & vocabulary v.2.1.

3.5 Data Sharing

As said, resource sharing has built upon META-SHARE. CRACKER maintained and released an improved version of the META-SHARE software.

For its own data sets, CRACKER has applied, 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 has been 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 has served as guidance for the selection of the appropriate licensing solution and creation of the respective metadata. In parallel, ELRA/ELDA has implemented a [licensing wizard](#), helping rights holders in defining and selecting the appropriate license under which they can distribute their resources.

3.6 Archiving and Preservation

All datasets produced are 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 are locally stored in the repositories' storage layer in compressed format.

4. Collaboration with Other Projects and Initiatives

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

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 (December 2017), the MoU has been signed by 12 organisations and 25 projects (including service contracts):

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

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

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 Deliverable 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 has been 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.

In this respect, the activities that the participating projects may optionally undertake in the future 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](#) for documenting their resources, even if they do not wish to link to the network.

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](#). 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](#). 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 DMPs 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 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.

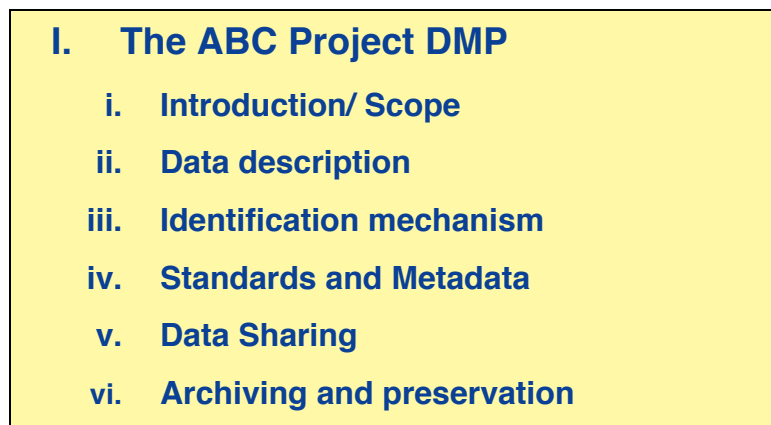
- 
- I. The ABC Project DMP**
 - i. Introduction/ Scope**
 - ii. Data description**
 - iii. Identification mechanism**
 - iv. Standards and Metadata**
 - v. Data Sharing**
 - vi. Archiving and preservation**

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.

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 has circulated two templates dedicated to datasets and associated tools and services respectively. Projects that wished and decided to participate in this uniform cataloguing were 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 have been 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 was recommended that all datasets (including all relevant metadata records) produced by the participating projects would be made available under licenses, which are as open and as standardised as possible, as well as established as best practices. 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 long-term preservation, 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