browser-based multilingual translation

bergamot

Horizon 2020 Research and Innovation Action
Grant Agreement No. 825303

https://browser.mt

Deliverable 8.1: Data Management Plan

Lead author(s):	Ulrich Germann (UEDIN)
Contributing author(s):	Kenneth Heafield (UEDIN), Frédéric Blain (USFD), Ondřej Bojar (CUNI), Jindřich Libovický (CUNI), Mark Fishel (TARTU), Kelly Davis (MOZ)
Internal Reviewer(s):	Robin Hill (UEDIN)

Work Package:	8
Type of Deliverable:	Open Research Data Pilot (ORDP)
Due Date:	30 June 2019
Date of Submission:	29 June 2019
Current Version:	1.0

Document History

Version	Date	Changes
1.0	29 June 2019	Original Submission

Contents

Abstract

This document describes how data will be managed within the Bergamot project. It describes which data will be produced or re-used during the course of the project, where it will be stored, and how we will ensure that it is findable, accessible, interoperable, and reusable (FAIR).

1 Introduction

The Bergamot project aims to lower language barriers within Europe by

- adding and improving client-side machine translation (MT) in a web browser, enabling the user to translate text without submitting and exposing content to third-party providers in the cloud;
- providing trustworthy MT with a graphical user interface that shows automatic estimates of MT quality, indicating to the user where the MT output may be unreliable;
- enabling the user to interact with online forms in foreign languages by providing *outbound* MT, i.e., translating text that the user has produced;
- developing MT technology that dynamically adapts to the content that it is asked to translate;
- developing small and fast MT engines and models that can be deployed on an ordinary desktop.

The target audience are ordinary computer users that use a browser to access the internet.

These aims are reflected in the project's scientific and engineering work packages: User Experience (WP1), Quality Estimation (WP2), Outbound Translation (WP3), Dynamic Adaptation (WP4), Efficiency (WP5), and Browser Integration (WP6).

The Bergamot project is firmly committed to open research and open-source software. Most project outcomes with be public, and all public project outcomes will be made available through the project web site at https://browser.mt. The remainder of this document provides a detailed breakdown of how data management issues will be handled for specific aspects of the project: what data will be used, created and produced in the various work packages, how the data will be stored, and how we will ensure that the data will be FAIR: findable, accessible, interoperable, and reusable.

Section 2 addresses data management issues relating to the public project web site. Section 3 explains how and where software components and trained models MT and QE will be stored and distributed. Sections 4 to 8 address data management issues specific to individual work packages.

This deliverable follows the suggested format[1] for data management under the EC Open Data Research Pilot.[2] Each section first describes the data to be used, collected, or created, then addresses data FAIRness, and finally explains where data will be kept and how it will be kept secure.

2 Web Page Data Management

2.1 Data Summary (Web Page)

Bergamot's public project web page, https://browser.mt is the gateway to all public project outcomes of Bergamot. It is generated with the static web page generator Jekyll[3] from source code hosted in a private repository on https://github.com/browsermt/browsermt.github.io. The generated web page is currently hosted by GitHub Pages[4] under the address https://browser.mt. The web page is managed by Kelly Davis (`kdavis@mozilla.com`).

2.2 Data FAIRness (Web Page)

2.2.1 Findability (Web Page)

As a key feature of Bergamot's dissemination strategy, the web page will be advertised widely in all dissemination activities. The project's dissemination strategy is described in Deliverable 7.2. Since the web page's source code is under revision control, the web page's evolution over time is fully preserved. There are no plans for explicit semantic versioning of the web page.

2.2.2 Accessibility (Web Page)

The public web page is openly accessible without requiring user login.

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

[2] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

[3] https://en.wikipedia.org/wiki/Jekyll_(software)

[4] https://pages.github.com/

2.2.3 Interoperability (Web Page)

The web page is delivered as HTML code and can be parsed/rendered by any HTML parser/rendering engine.

2.2.4 Reusability (Web Page)

The web page is intended to be indexed by search engines, but not to be re-used in any other form.

2.3 Allocation of Resources (Web Page)

The web page is currently being hosted by Github Pages free of charge. It is maintained by Kelly Davis within his time allocation in the Bergamot project.

2.4 Data Security (Web Page)

Currently, the web page is hosted by a third-party provider (github.com) free of charge. Because of its simple design as a static web page, it can easily be self-hosted or moved to another provider, should the need ever arise. We do not host any private data on the website, so that there are no security concerns with respect to data privacy for data available through the web page.

Unique user tracking for the purpose of collecting dissemination statistics (cf. D-7.2) will be anonymised and in compliance with the GDPR.

3 Data Management of Software, Software Code, and MT/QE Models

3.1 Data Summary (Software and Models)

The outcomes of the Bergamot project that will be of the most value and and interest to the general public will be software for machine translation (MT) and quality estimation (QE), and their integration into the browser interface via a browser extension, along with associated models for MT and QE.

3.1.1 Software Code

For some components of the Bergamot project (e.g. Marian for MT, deepQuest and OpenKiwi for QE), we rely on an existing code base; the browser extension will be developed from scratch.

In addition to the code written by the software developers, software testing by developers and users will result in bug reports and subsequent online discussions. These bug reports and subsequent discussions will be filed and preserved in issue trackers associated with the respective software repositories.

3.1.2 Trained Models for MT and QE

In general, models for MT and QE will be developed from publicly available data sets, such as those provided for the shared task at the annual Conference on Machine Translation,[5] or hosted on public repositories of parallel data such as http://ParaCrawl.eu, or http://opus.nlpl.eu/. Additional data requirements for specific work packages and tasks are discussed in the respective sections later in this document.

3.2 Data FAIRness (Software and Models)

The Bergamot project is firmly commited to open research and open-source software. As mentioned earlier, all of the public project outcomes will be made available through the project web site at https://browser.mt. Software components will be licensed under permissive open-source software licenses (cf. Table 1); trained models for MT and QE will be licensed under Creative Commons license CC-BY-SA 4.0.[6]

All software components will be published as open-source software in publicly accessible online code repositories on github.com. Public code repositories hosted by these providers are usually indexed by all major internet search engines, so that Bergamot software components will be easy to find through an ordinary internet search.

Once the catalogue feature of the European Language Grid[7] is operational,[8] (links to) software components and models will also be listed there.

The Firefox extension that provides MT functionality to Firefox will be able to be installed just like any other Firefox extension, described for example by the section "How do I find and install add-ons?" on the Mozilla

[5] http://www.statmt.org/wmt19/index.html

[6] https://creativecommons.org/licenses/by-sa/4.0/

[7] https://european-language-grid.eu

[8] The first release of the ELG is scheduled for April 2020; the final release for Oct. 2021.

Table 1: Software repositories and licensing terms

.

Marian (existing toolkit that will be extended)	
Function:	Toolkit for training and applying neural MT models Contributions from WP5 (Efficiency) will be added to this code base.
Repository:	https://github.com/marian-nmt/marian
License:	MIT
deepQuest (existing toolkit that will be extended)	
Function:	Toolkit for quality estimation
Repository:	https://github.com/sheffieldnlp/deepQuest
License:	BSD 3
OpenKiwi (existing 3rd-party toolkit)	
Function:	Toolkit for quality estimation
Repository:	https://github.com/Unbabel/OpenKiwi
License:	GNU AGPL v3.0
Bergamot Extension for Firefox	
Function:	Browser integration of MT interface
Repository:	https://github.com/browsermt/Extension
License:	MPL
Dynamic Domain Adaptation	
Function:	Dynamic domain adaptation of MT engine
Repository:	https://github.com/TartuNLP/bergamot (eventually to be integrated into Marian)
License:	MIT
Outbound Translation (based on existing code)	
Function:	User Interface for outbound translation
Repository:	https://github.com/zouharvi/ptakopet (eventually to be transferred to https://github.com/browsermt)
License:	MIT

support page. In particular, the extension can be installed either by downloading it from the Firefox Add-ons site or installing it from a file.

The MT and QE models will initially be hosted by the University of Edinburgh's Machine Translation group[9] and also be made available through the European Language Grid once it is operational. Eventually, models will also be stored as versioned binary "assets"[10] with the respective software releases on github.com's infrastructure. Versioning of models will facilitate bug tracking and backwards compatibility as the project evolves, and give users greater choice with respect to trade-offs between model size and performance.

Users of the Firefox Browser Extension will be able to download models directly through the extension.

Issue trackers for open-access software will be accessible for reading and filing through links in the respective code repository, following established software development procedures.

3.3 Allocation of Resources (Software and Models)

Cloud-based public software code repository providers such as https://github.com tend to host open-source software project at no charge. All academic project partners rely on https://github.com for public hosting their projects, including data security. As an added measure of security against data loss, we will keep local copies of the software repositories on our institutions' computing infrastructure. The hardware footprint of these copies is negligible (on the order of gigabytes of disk space per repository).

3.3.1 Data Management at Mozilla

As for data management at Mozilla, the ultimate responsibility for its data management is held by Mozilla's project director. With respect to the Bergamot project, Mozilla distinguishes the following data management roles, whereby different roles may be filled by the same person:

- The **Data Collector** is responsible for obtaining the training data.
- The **Metadata Generator** is responsible for producing data about the data that is of use for various downstream tasks.
- The **Data Analyzer** is responsible for training models on the data.
- **IT Support** is responsible for maintaining the cluster where the training data is hosted.

[9] http://data.statmt.org/

[10] https://developer.github.com/v3/repos/releases/#upload-a-release-asset

Generally, Mozilla will not be concerned with data collection for model training and relies on available resources (e.g., from ParaCrawl). As for the Metadata Generator and Data Analyzer, generally we view these roles as being executed by a single person who, through iterative training, is able to derive relevant metadata to tag the data with. Finally, IT support will help ensure that the internal cluster hosting the data, e.g. a copy of the ParaCrawl data, remains healthy.

3.4 Data Security (Software and Models)

The University of Edinburgh's Machine Translation Group uses Redundant Arrays of Independent Disks (RAID)[11] for data storage, which protect the data against ordinary disk failure. In addition, the group maintains back-ups in a secondary, independent location within the University of Edinburgh's computing infrastructure.

Mozilla's Machine Learning Group's servers also store all data on a RAID infrastructure to protect data against disk failure. In addition, all production models and codes are backed-up by an external 3rd-party service which also keeps redundant copies of these models and codes.

For protection against unauthorized access, Mozilla's Machine Learning Group's restricted data is stored on a private network that requires LDAP access. This LDAP access mechanism is maintained and hardened by Mozilla's IT security team. In addition, to get access to Mozilla's servers, even from within the private network, one has to be white-listed for cluster access. Once one is within the private network and white-listed for cluster access, for each data set one is given access to one must be further white-listed for that particular data set. This setup provides multiple layers of security to keep any restricted data restricted.

4 Data Management Plan for User Experience Data (WP1)

The data management plan for data created within the User Experience work package (WP1) was already described in the Bergamot Ethics Plan (D-1.1); this section repeats the plan set out there. Data created within WP1 is managed by Robin Hill (`r.l.hill@ed.ac.uk`).

4.1 Data Summary (User Experience Data)

The purpose of this work package is to evaluate human interaction with the Bergamot system as it is being developed. This involves monitoring per-

[11] https://en.wikipedia.org/wiki/RAID

formance during simulated tasks and during more open exploration (e.g. "think aloud" protocol). As well as implicit measures (biometric monitoring), explicit measures such as task success rates and completion times will be recorded. Participants will also provide subjective measures of usability and complete standard usability (HCI) questionnaires. To ensure an understanding of the sample of the population taking part in our experiments, basic demographic information will be requested, along with questions related to their current use of translation technologies and future expectations. Initial requirement gathering surveys will be conducted. Interviews with expert users, stakeholders and potential end-users will be carried out and transcribed.

Laboratory-based usability testing will involve biometric monitoring: eye-tracking, facial expression analysis, pupillometry, heart rate and galvanic skin response. The current laboratory software combines all these sensor dataand synchronises with the stimuli that was being displayed or interacted with to produce a raw data file for each participant. An anonymised processed file is then produced for each individual and these are collated into a single analysis file for each experiment. These are output as documented CSV or tab-delimited alphanumeric files suitable for import into any statistical package or data analysis software. All questionnaires, surveys and other response data will also be transformed into a generic spreadsheet format which only contain anonymised information. Participants will be recruited and interviewed to confirm suitability. They will also be fully briefed and informed before they are permitted to take part in the study.

Storage will take place on secure servers based at the University of Edinburgh. The University operates a dedicated service for this purpose: the Edinburgh Datastore.[12] Depending on sample rate and resolution, the initial lab-based raw sensor data can be quite large (10 to 15 MB/min), but the processed data are simple ASCII text files.

The primary beneficiaries are the Bergamot developers as they are adopting a user-centred and more agile design and development process. Work Package components and holistic prototype iterations will be tested throughout the duration of the project. The aim is therefore to ensure the optimal layout, information delivery and data visualisation in the browser interface. Studies will verify whether quality estimates of machine output and algorithmic levels of confidence are communicated to users in an accessible and usable manner. Human levels of tolerance to response latency, other delays and error rates will be investigated leading to an understanding of the upper and lower bounds of speed-accuracy trade-off that the Bergamot system must try to operate within. In addition to validating the overall quality of machine translation by human perception, the predicted levels of confidence automatically calculated will also be tested for correlation with human judgements to ensure an accurate reflection. However, there will also be wider theory-testing aspects to the empiricial research which

[12] https://www.ed.ac.uk/information-services/research-support/research-data-service/working-with-data/data-storage

will appeal to a wider scientific audience. For example, the studies on error detection by humans will have a broader impact for visual processing, affective computing and cognitive psychology.

4.2 Data FAIRness (User Experience Data)

4.2.1 Findability (User Experience Data)

While any appropriate data access request will be honoured (e.g. for replication purposes, GDPR regulations, or transparency during peer review) it is anticipated that the more public release of the user experience dataset will only occur at the end of the project. For future archiving and wider accessibility, the University provides a long-term repository and data vault, which includes providing searchable metadata and a persistent Digital Object Identifier (DOI).[13] Current recommended convention is to store data for 10 years after the last access.

4.2.2 Accessibility (User Experience Data)

The processed, anonymised and collated data from the laboratory experiments, field studies and interviews will be made openly available via the dedicated University of Edinburgh Information Services facility described above. The raw data will not be made available as this contains potentially identifiable information.

Blank copies of all questionnaires and surveys will be stored with the data. All files will have human-readable file names as well as internal documentation and/or readme.txt files. Directories will be indexed. All variables and column headings will be catalogued and described.

4.2.3 Interoperability (User Experience Data)

The files themselves will contain minimal formatting (essentially open spreadsheets) but will be accompanied by files describing the details of any experimental procedure, sample size, dates, etc. as well as the column headings and variables. The files themselves will have appropriate metadata, following the guidelines and requirements proposed by the University of Edinburgh.

4.2.4 Reusability (User Experience Data)

Our consent forms and ethical procedure covers re-use for research and academic purposes, including dissemination via journal or conference pa-

[13] https://www.ed.ac.uk/information-services/research-support/research-data-service/after

pers.

4.3 Allocation of Resources (User Experience Data)

We do not anticipate the need for data storage beyond the default provided by the University of Edinburgh for all official University projects. If our requirements and demands change over the course of the project then the current fees for additional storage are £175 per TB per year for DataStore and £500 per TB for 10 years of long-term DataVault storage.

4.4 Data Security (User Experience Data)

The Edinburgh University servers are backed-up daily. Data on laboratory machines are backed-up onto the University network at the end of each day of testing and transferred off the lab machines at the end of each run of an experiment. This is high quality, enterprise-class storage with guaranteed backup and resilience. The data is automatically replicated to an off-site disaster facility and also backed up with a 60-day retention period, with 10 days of file history visible online.[14]

4.5 Ethical Aspects (User Experience Data)

The Bergamot proposal has passed through the School of Informatics Ethical Procedure at the University of Edinburgh (separate documentation available along with this DMP as part of Deliverable D1.1).

Participants provide their own data with full consent. There is no collection from 3rd parties or other agencies about participants. Similarly, no personal information is transferred to any 3rd party.

Any data not handled by the researchers directly involved in the data collection will be fully anonymised. Any data that could potentially lead to identification of individuals (e.g. video recording of participants or audio files of interviews) will not be part of any database/corpus release outwith the Bergamot consortium. The research team is highly experienced and qualified in handling human experiment data.

5 Data Management for Quality Estimation (WP2)

The data manager for data related to research on Quality Estimation is Fred Blain (`f.blain@sheffield.ac.uk`).

[14] http://www.ed.ac.uk/is/data-management

5.1 Data Summary (Quality Estimation Data)

The purpose of this work package is the design and delivery of performing QE models for all the languages covered in the project. Our aim is to produce accurate quality estimates on our MT output, to inform Bergamot end users. Those quality indicators will be shared with the developers of both the User Experience and the Outbound Translation work packages (WP1 and WP3).

Quality Estimation, framed as a supervised machine learning problem, requires by definition labelled data for the QE models to be trained on. Several sources of data will be considered throughout the project:
During the initial phase of Bergamot, QE models will be trained on publicly available datasets, such as those released by the QE shared task at the annual Conference on Machine Translation (WMT). Those datasets consist of plain text files of source and target sentences, along with their quality labels.
To provide quality estimates for all the target languages covered in Bergamot, we will investigate training our QE models under weak or partial supervision. For this purpose, will be considered parallel with automatically generated “quality” indicators, such as scores from automatic metrics for MT, or human direct assessment (DA). Such data will be collated and processed from the freely available dataset released by both the MT and the Metric shared tasks at the WMT. Those datasets consist of plain text files of source sentences, machine and reference translations, and plain text files with numerical values derived from DA.
Finally, work will be conducted alongside the Bergamot developers on User Experience Data (WP1) to determine the more effective feedback from end users and the best approach to collect it. Such data will be used to support work on adaptive QE, to improve our QE models through time.

5.2 Data FAIRness (Quality Estimation Data)

5.2.1 Findability (Quality Estimation Data)

To share, archive and globally wider accessibility, the University of Sheffield provides a long-term research data catalogue and repository, which includes providing searchable metadata. Each deposited research data may be accessed in an open or controlled manner. Preserved for at least 10 years, each deposited data will also be given a DataCite Digital Object Identifier (DOI).[15]

[15] https://www.sheffield.ac.uk/library/rdm/orda

5.2.2 Accessibility (Quality Estimation Data)

As mentioned above, trained QE models will initial be hosted by the University of Edinburgh's Machine Translation group, and also be made available through the European Language Grid.

The collated and processed QE data from the laboratory experiments and field studies will be made openly available via the dedicated University of Sheffield Research Data Catalogue and Repository described above.

5.2.3 Interoperability (Quality Estimation Data)

QE models released during the project will be accompanied by files describing in details how to use them to generate quality estimates. This implies that those files will also provide links and information regarding the source code to be use along with the models.

All QE data collated during the project will contain a minimal formatting (essentially plain text), and will be accompanied by files describing the details of its origins, languages, sample size, dates, appropriate citations, contact information, etc. The files themselves will have appropriate metadata, following the guidelines and requirements proposed by the University of Sheffield. [16]

5.2.4 Reusability (Quality Estimation Data)

All data coming from publicly available resources such as those released by shared task organisers at WMT, and used to train our QE models, will by definition be reusable by third parties.

Source code for the training of QE models will be available on hosting services with version control such as GitHub, and published under a permissive license.

Released QE models will be made available along with documentation comprising training and tutorials, configuration files, benchmarks, etc. in order to assure reusability and reproducibility of our results.

5.3 Allocation of Resources (Quality Estimation Data)

The University of Sheffield provides data storage solutions to each of its research groups, to store materials during the life of a project. We do not anticipate the need for additional storage capacity beyond the default offering. If our requirements and demands change over the course of the project then the current fees for additional storage are £100 a year per TB per copy.

[16] https://www.sheffield.ac.uk/library/rdm/index

5.4 Data Security (Quality Estimation Data)

All spaces of the University of Sheffield's research data storage get regular snapshots, which are copies of the storage area at a specific point in time. Snapshots of each area are taken several times a day, and are retained for 7 days.[17]

6 Data Management for Research on Outbound Translation (WP3)

Data manager for data related to research on outbound translation is Ondrej Bojar (bojar@ufal.mff.cuni.cz).

In this section, we specifically address data management for Outbound Translation. Unlike other WPs, Outbound Translation does not require specific data resources and mostly combines models from other work packages (MT, quality estimation).

The only exception is modeling problematic source words. These models will be based on publicly available data for automatic MT post-editing and on synthetic data generated by MT systems. The software both for generating the synthetic training data and training the models will be open-sourced under a permissive license.

Logs from the user interaction with the interface will be collected and used for improving the user experience. Users (both during user testing and during the experimental deployment of the system) will be explicitly informed what information is collected from the interface and instructed to avoid inserting information that might be considered sensitive.

Based on an analysis of the data collected during user testing, we might use the logs to mine data for quality estimation or automatic MT post-editing. In that case, the data will be released through LINDAT/CLARIN.[18]

Terms of Use of the repository already satisfy the FAIRness requirements.

6.1 Data Summary (Outbound Translation)

Outbound translation will reuse models for MT and Quality estimation from other WPs of the project.

Models for detecting problematic words on the source text will be trained using publicly available data for automatic MT post-editing and synthetic training data generated using already trained translation models. Detection of problematic words strongly depends on the underlying MT system used

[17] https://www.sheffield.ac.uk/cics/research-storage/standard-storage

[18] https://lindat.mff.cuni.cz

in Outbound Translation and so do the synthetic training data. Because of that, we do not consider this data re-usable for other purposes and thus will only publish the software that can be used to generate the synthetic data for a particular translation system.

During user testing and experimental deployment of the Outbound Translation system, detailed logs will be collected. We believe it will be possible to use the logs to compile datasets which might be useful for quality estimation of automatic MT post-editing. In that case, the dataset will be anonymized and published in the LINDAT/CLARIN[19] repository.

6.2 Data FAIRness (Outbound Translation)

6.2.1 Accessibility (Outbound Translation)

The LINDAT/CLARIN repository allows all registered user agreeing with the respective data license to download the stored datasets. For searching the repository, no registration is needed.

Datasets stored in LINDAT/CLARIN are indexed by standard search engines. In addition, it provides a proprietary search engine that allows searching for dataset based on various criteria including keyword and metadata search.

6.2.2 Interoperability (Outbound Translation)

If any dataset will be compiled from the outbound translation logs, it will be distributed in a simplest possible format, either as plain text or in JSON format, to keep technological barriers as low as possible.

6.2.3 Reusability (Outbound Translation)

All datasets compiled from the Outbound Translation logs will be published under a permissive variant of the Creative Common Licence.

Source code for the data generation and models for detection of problematic words will be available on github.com and published under a permissive open-source software license.

6.3 Allocation of Resources (Outbound Translation)

The source-code repositories are hosted by github.com free of charge.

[19] https://lindat.mff.cuni.cz/en

The LINDAT/CLARIN repository hosts all data free of charge. Its operation is funded by the Ministry of Education, Youth and Sports of the Czech Republic.

6.4 Data Security (Outbound Translation)

Since the logs (both from the controlled user studies and experimental deployment of the outbound translation system) will be anonymised, we do not expect to work with sensitive data of any kind. The users will be instructed not to translate any sensitive information. In spite of that, if any public dataset will be compiled from the logs, the data will be anonymised by removing or masking automatically detected entities (addresses, phone numbers, etc.)

LINDAT/CLARIN is committed to the long-term care of items deposited in the repository, to preserve the research, and to help keep research replicable. It strives to adopt the current best practice in digital preservation. The data is regularly backed up with offline backups in physically distinct locations. In the case of a withdrawal of funding, the repositories content would be transferred to another CLARIN centre. While the legal aspects of the process of relocating data to another institution are underway the hosting institute (UFAL, CUNI) offers a timeframe of at least 10 years, in which it will provide access to the data.

7 Data Management for Dynamic Adaptation (WP 4)

Data manager for data related to research on dynamic adaptation of MT engines is Mark Fishel (`fishel@ut.ee`).

In this section, we specifically need to address data management for Task 4.1: Data Collection for DA. All of the work in WP 4 involves public domain data and creating the necessary specific resources based on it, without any user-specific data or modelling involved.

7.1 Data Summary (Dynamic Adaptation Data)

The main purpose of the data to be collected in Task 4.1 of this WP is to support the experiments in Tasks 4.2, 4.3 and 4.4, which consist of research on machine translation models that can dynamically adapt to the text domain and/or document context at hand.

The data will consist of parallel and monolingual texts with additional meta-information such as the URL of the web page source, HTML metadata, manually and automatically identified text domain and/or domain clusters. We will base our text corpora on ParaCrawl,[20] which is a massive set of crawled web texts. Mainly we will analyze and supplement these texts with the additional information.

Speaking broadly, the resulting data collection will be of use to anyone working on integrating additional information into MT models. In particular, this means methods of adaptation to text domain and overall context.

7.2 Data FAIRness (Dynamic Adaptation Data)

7.2.1 Findability and Accessibility (Dynamic Adaptation Data)

The main channel for releasing data sets created under WP 4 is the Center of Estonian Language Resources (CELR)[21] on the META-SHARE network. CELR is CoreTrustSeal-certified and as such satisfies the FAIR data principles.

The CELR META-SHARE repository allows all registered user agreeing with the respective data license to download the stored datasets. For searching the repository, no registration is needed.

Datasets stored in the CELR META-SHARE are indexed by standard search engines. In addition, it provides a proprietary search engine that allows searching for dataset based on various criteria including keyword and metadata search.

7.2.2 Interoperability (Dynamic Adaptation Data)

All datasets compiled as part of WP 4 will be distributed in the simplest possible format as plain-text or JSON, to keep technological barriers as low as possible.

7.2.3 Reusability (Dynamic Adaptation Data)

The data to be collected as part of WP 4 is general and thus highly reusable. This data will be released under a permissive open-source license and disseminated via open access conference papers and journal articles.

7.3 Allocation of Resources (Dynamic Adaptation Data)

The source-code repositories are hosted by github.com free of charge. The CELR META-SHARE repository hosts all data free of charge; its operation is funded by the Estonian Ministry of Education and Research.

[20] https://paracrawl.eu

[21] https://www.keeleressursid.ee/en/, https://metashare.ut.ee/

7.4 Data Security (Dynamic Adaptation Data)

All storage space used by CELR META-SHARE is provided by the University of Tartu's High-Performance Computing Center (HPC) and gets regular daily snapshots and backups.

The base data on which this WP is based is public domain and does not involve any personal information, thus removeing any risk in this regard.

8 Browser Integration (WP6)

8.1 Data Summary (Browser Integration)

This work package is mostly concerned with software engineering. The outward-facing outcome will be software code, which is covered in Section 3. User testing during software development will result in bug reports, which will be stored in an issue tracker maintained and controlled by the Mozilla corporation. Data manager for data produced in this WP is Kelly Davis (`kdavis@mozilla.com`).

8.2 Data FAIRness (Browser Integration)

8.2.1 Findability (Browser Integration)

The aforementioned issue tracker will be linked from the code repository, which we anticipate to be indexed by most major search engines.

8.2.2 Accessibility (Browser Integration)

In adherence with the Accessibility pillar of FAIR, the browser extension code and associated models will be accessible. All release versions of both will be stored on github.com. The sources and their documentation will be stored in a normal GitHub repository and the models will be bundled as "assets" with each release. So, one can obtain the browser extension code using the well documented `git` command, which itself is open source, and one can obtain the models through a simple `GET` request, using for example the open source browser Firefox.

All users of the software will be able to file issues; only members of the development team will be able to manage issues once filed.

8.2.3 Interoperability and Reusability (Browser Integration)

Since bug reports are specific to particular pieces of software, they are not expected to be interoperable or reusable.