Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
First Periodic Technical Report – Part B

Nature	Report	**Work Package**	WP7
Due Date	29/08/2020	**Submission Date**	21/08/2020
Main authors	Alexandra Birch (UEDIN)		
Co-authors	WP and Task Leaders		
Reviewers	Barry Haddow (UEDIN)		
Keywords	Periodic, technical, deviations, expenditure, effort		
Version Control			
v0.1	**Status**	Draft	15/08/2020
v1.0	**Status**	Final	21/08/2020
v1.1	**Status**	Updated UEDIN expenses	02/09/2020
v1.2	**Status**	Updated person months to M18 figures	22/10/2020

Contents

List of Figures

1 Explanation of the work carried out by the beneficiaries and overview of progress

Europe is currently facing immense challenges. Old certainties are crumbling, and the current world order is changing rapidly, with new political and economic powers emerging. In order to successfully navigate through these choppy waters, it is essential that we look outwards.

Two of Europe's most well-respected international news broadcasters, the BBC and DW, provide a window to this globalised world by broadcasting in many European and non-European languages. They are flagships of European news broadcasting and renowned around the world for their editorial independence and accurate news reporting. The BBC and DW are acutely aware of their responsibility, and want to meet the growing demands of their audiences with innovative technology.

Machine translation (MT) is key technology for achieving a global reach. It allows analysts and journalists to quickly and effectively gather information across very diverse languages, and to understand how these events are being perceived and reported on. It is also a powerful tool for speeding up the dissemination of news reports in multiple languages.

The uptake of MT technology has gradually increased over the last ten years, but recent advances in neural machine translation (NMT) have resulted in significant interest in industry and have lead to very rapid adoption of the new paradigm (eg. Google, Facebook, UN, World International Patent Office). These high performing neural models require a massive amount of training data to deliver these results, many millions of translated sentences.

Even though there are a few language pairs which are well-resourced, there are many more language pairs where little or no translated text is available for training translation models and some under-resourced languages have large populations which are important either for commercial, strategic or humanitarian reasons.

Inspired by human learning, and leveraging machine learning techniques which have been successfully applied in other low-resource machine learning applications, we are making translation more widely applicable, across languages and across low-resource domains. We also invest significant resources into the tools for extracting language resources for new languages and domains, in order to quickly and easily deliver new translation engines to cope with the constantly changing news landscape.

GoURMET is organised around three main use cases:

- **Global content creation** – managing content creation in several languages effciently by providing machinetranslations for correction by humans ;
- **Media monitoring** – for low-resource language pairs — tools to address the challenge of monitoring media instrategically important languages;
- **International business news analysis** – reliably translating and analysing news in the highly specialised financial domain.

1.1 Objectives

The GoURMET project is structured around five objectives. We have made significant progress on each of these during the first period of the project.

Objective 1: Advancing low-resource deep learning for natural language applications

The core objective of GoURMET is the development of methods for low-resource deep learning which is able to optimally learn from small amounts of training data. This objective is the focus of WP1, WP2, WP3 and WP4.

Our progress with regard to this objective has been significant. The strong research component of the project is reflected in our scientific publications (at the time of writing, 26 publications shared in an open-access manner via OpenAIRE, the European Open Science Intiative [1]). We have more work which is under review, or published as pre-prints or theses. We have released 22 different repositories related to research software that accompanies these publications [2]. These represent a significant output for the project and are related to the following milestones:

- **MS4**: Release of initial morphological models (M12).
- **MS5**: Release of initial structural models (M12).
- **MS6**: Release of initial transfer learning approaches (M12).

Objective 2: Development of high-quality machine translation for under-resourced language pairs and domains

This objective concerns collecting and augmenting the data for our low-resource translation tasks and training models which deliver high-quality translations to the media partners. It also covers delivering tools which make it easier to deploy new low-resource MT engines in a short time frame. This objective mainly concerns work done in WP1 and WP5.

We address this objective by pursuing research into improving the data collection pipeline. Some of the publications and software referenced in Objective 1 are related to crawling and cleaning parallel corpora from the web. We have released 10 data sets so far, both parallel data sets and some monolingual data sets.

We further ensure our success at pursuing this objective by having a 9 month-cycle of building, delivering, and evaluating translation models for low-resource languages. We have currently completed two rounds of translation model buidling. In round one we delivered translation models into and out of English for Gujarati, Bulgarian, Turkish and Swahili. In round two we delivered Amharic, Kyrgyz, Serbian and Tamil. In round three we are currently collecting corpora for Hausa, Igbo, Amharic and Macedonian and engaging with the Masekane project and Translators without Borders to create and develop African models. We prioritise languages which are strategically important for the BBC and DW, and then from a shortlist, the research partners select languages which are interesting for their research and have a variety of resources to work with. These models are released to the public.

- **MS1**: First release of translation models (M6).
- **MS7**: Second release of translation models (M18).
- **MS9**: Initial release of project data (M18).

[1] https://explore.openaire.eu/search/project?projectId=corda__h2020::0dac160633eb5c7e8a63df801ba0ec58
[2] We have release software, datasets and trained models here: https://gourmet-project.eu/data-model-releases/

Objective 3: Development of tools for analysts and journalists

This objective relates to the design and implementation of interfaces to translate, edit and evaluate translations. We also work towards combining these interfaces with content creation platforms and media monitoring platforms like plainX (previously the news.bridge platform) to deliver multilingual content for low-resource language pairs and domains. This objective mainly concerns work done in WP5.

Our progress towards this objective has been largely achieved through the development of the GoURMET translation platform. This platform is based on serverless AWS architecture and currently supports seven of the eight languages delivered in the project. It is available via an API, and there is also a demo front end, which is linked to from the GoURMET website [3].

We have reached the following milestones from the first period of the project:

- **MS2**: Initial release of content creation user interface (M8).
- **MS3, MS8**: First and second evaluation of initial translation models (M9, M20).

Objective 4: Sustainable, maintainable platform and services

This objective concerns the development of a plan for sustainable exploitation and use of the platforms, systems and technologies developed in GoURMET. This objective mainly concerns work done in WP6. Progress towards this objective is mainly relevant to the second period of the project. However, the design and planning of these services is described in Deliverable 5.2 Use Case Description and Requirements. The milestones relating to this objective (MS10, MS15, MS20 and MS25) are all timetabled for the second period of the project. However, planning for the surprise language event and the hackathon is already underway.

Objective 5: Dissemination and communication of project results to stakeholders

We engage with organisations from all along the innovation chain, including broadcasters, commercial players, governmental/EU agencies in the single digital market and relevant research communities. This objective mainly concerns work done in WP6.

The milestones relating to this objective (MS10, MS15, and MS20) are all timetabled for the second period of the project. However, measurable progress has been made through project presentations and engagement at a variety of external events, and a large number of project publications. See Deliverable 6.3 Interim Dissemination and Exploitation Report for details.

In the next section we summarise the work.

[3] https://gourmet-project.eu/project-output/gourmet-translate-tool/

1.2 Explanation of the work carried out per WP

1.2.1 WP1: Data Gathering and Augmentation

WP Leader: UA
Participating partners: UEDIN (5.96/12 pm), UA (25.23/40 pm), BBC (0.5/3 pm), DW (7.3/10 pm)

Task T1.1: Identifying, collecting and evaluating monolingual and bilingual language resource This task kicked off by working on deliverable *D1.1 Survey of relevant low-resource languages*, which reports on the identification of monolingual and bilingual resources and corpora for the languages of interest to BBC, DW and the GoURMET project. The first version, delivered on April 2019, covered the 16 languages included in the proposal; last version to date covers 3 additional languages that were not part of the project proposal but for which we developed translation models in the second round of languages. Many of the 19 languages in this deliverable are from the Indian sub-continent and Africa, but there are also languages from Eastern Europe and the Middle and Far East.

Each language is described, including contrasts with English based off WALS[4] and its challenges for corpus-based MT. Available resources are provided. These resources include monolingual and bilingual corpora as well as linguistic resources. This is an extensive and thorough piece of work that we plan to updated as we encounter new languages of interest or new corpora.

Language resources from our partners in the BBC and DW were also leveraged. These data dumps consist of collections of pieces of news that have been published by BBC or DW in their respective websites. They have been processed mainly to obtain test corpora, except for Serbian for which we also obtained data for training.[5] These data dumps can be considered comparable corpora, which made the procedure usually followed with crawled data unsuitable, specially for the BBC data dump; a detailed description of the process followed to process these data dumps is provided in deliverable *D1.2 Initial progress report on data gathering and augmentation*, Section 3. The test sets obtained from these data dumps were subject to human curation, a process not initially envisaged in the proposal, in order to achieve high quality data for evaluation purposes.

Task T1.2: Identifying, collecting and evaluating monolingual and bilingual corpora This task is responsible for the process to acquire parallel and monolingual data from the Internet as a resource to train MT systems. Deliverable *D1.3 Initial release of project data* provides pointers for downloading them, whereas deliverable *D1.2 Initial progress report on data gathering and augmentation* describes the process we have followed for crawling each corpus. This process can be split into three different steps: crawling, document and segment alignment, and cleaning.

For crawling we initially relied on the CommonCrawl's list of language-identified websites. While this was sufficient for most language pairs, for some of them there were very few parallel websites in CommonCrawl, or they were no longer available. To circumvent this problem we developed LinguaCrawl,[6] a tool for exploring top level domains to find website from which to crawl data. This tool has been released under a free/open-source license.

[4] The World Atlas of Languages Structures; https://wals.info/
[5] It is worth noting that to be able to freely distribute our NMT models, BBC data cannot be used for training, since BBC has restrictions on derivative work.
[6] https://github.com/transducens/linguacrawl

For alignment we used Bitextor. This tool relies on bilingual lexicons to identify candidate parallel documents, which were then sentence-aligned using Hunalign or BLEUalign. Some of the languages of interest to GoURMET are agglutinative and morphologically rich and this exacerbates the under-resourced scenario we are addressing. As a result, for some language pairs the amount of bilingual resources to obtain reliable bilingual lexicons is insufficient. To mitigate this problem we extended Bitextor so that the documents are lemmatised or stemmed before their alignment.

For cleaning we have used Bicleaner or LASER, depending on the language pair. For each corpus we first produced a initial version that was cleaned with both tools using low threshold values and then evaluated by the user partners. The results of these evaluations allowed us to decide on the final tool (Bitextor in all cases but one) and threshold to be used for producing the corpora used to train the neural MT systems. Bicleaner relies on probabilistic bilingual dictionaries, which, as stated above, may have low coverage for some language pairs. To improve its performance, we have devised new features, based on word frequencies computed on monolingual data, new noise models for better scoring and used a new classifier. We have also developed a tool to be able to train LASER-like models for those languages not supported by LASER.

Task T1.3: Synthetic data and lexical augmentation This task deals with the generation of synthetic parallel corpora for training neural MT systems. In order to leverage the monolingual corpora available for the languages of interest to GoURMET we have applied the standard back-translation technique; in one case the system used for back translation was a statistical MT system, in the rest of cases it was a neural MT system. In addition, for some language pairs we applied an iterative back-translation algorithm that simultaneously leverages monolingual data in the two languages.

We have also conducted research on a novel method based on the use of variational autoencoders that operate on sentences to generate synthetic text. The aim of this approach is to enhance the diversity of existing datasets and improve the performance of the neural MT systems trained on them. Section 4.1 in deliverable *D1.2 Initial progress report on data gathering and augmentation* describes this research and outlines its main results. This line of research was started during the research stay of Dr. Víctor M. Sánchez-Cartagena (from UA) at the University of Amsterdam and will continue during the second half of the project.

1.2.2 WP2: Modelling Morphological Structure

WP Leader: UEDIN
Participating partners: UEDIN (15.77/36 pm), UA (5.6/14 pm), UVA (0/24 pm))

Task T2.1: Linguistically informed NMT models using morphology The scope of this task has expanded slightly since the initial proposal, with more emphasis now on using morphological models and morphologically guided intuitions applied to neural machine translation (NMT).

The goal of this task is to use morphological information and intuitions to guide the development of NMT models, either by exploiting information form external analysers or using linguistic intuitions to supervisedly guide the design of the NMT architecture. There are two ways in which morphology can be used within this task: (i) to guide subword segmentation in a linguistically informed way and (ii) the use of external morpho-syntactic information to encourage enrich the representation of (sub)words and to provide a better capacity to generalise across wordforms.

Three directions have so far been followed in line with this task. The first studies morphologically guided subword segmentation using the morphological analyser from Apertium (in contrast to the standard statistically guided BPE strategy) for English-Kazakh translation. The second uses linguistically guided intuitions in proposing a hierarchical decoded consisting of a word-level recurrent decoder, but whereby words are decoded character by character. The third moves away from segmentation to the exploitation of morphological information in a study on the use of source-side and target-side morphological information using an approach of interleaving part-of-speech and morphological tags into the source and target sentences. Open-source code is available for these three pieces of work.

Task T2.2: Jointly learning alignments and morphology The work in this task is to be completed in the second half of the project. The schedule has changed slightly with respect to what was initially drawn up, with work having been completed earlier in Task 3, and this task having been pushed back slightly.

The aim of Task 3 is to learn models of subword segmentation that are guided by the word (or subword) alignment induced between sentences in parallel texts (where sentences in one language are aligned to their translations in another language). Inspired by work in Work Package 3 Task 1 on modelling latent alignments, the proposal is to jointly model subword segmentation and subword-level alignment between sentences. The idea is that the subword segmentation informs the alignment model and the alignment model in turn informs subword segmentation. Especially for language pairs in which one language is more morphologically rich than the other, the aim is to find a segmentation that makes the alignment between subwords more straightforward (e.g. a one-to-one mapping between (sub)words in one language and sub(words) in the other).

Task T2.3: Factors encoding latent features of morphology Task 3 is dedicated to models in which morphological attributes are induced as latent variables while training towards a downstream learning signal such as word generation or translation. This lies in contrast to the models in Task 1, for which the morphological information is provided more explicitly to the models.

Three pieces of research correspond to this task. The first involves designing a generative model of inflected wordforms ina semi-supervised way, applied to a single language. This work is then revisited and extended in the second piece of research to apply it to MT (rather than in a monolingual

setting), where the decoder must inflect every single word in the target sequence. In this second work, latent morphology is induced in an entirely unsupervised way, rather than being partly supervised as in the first work. Finally, the third piece of work develops a variational NMT model with morphological priors, which introduces latent modelling of morphology into NMT as in the second work, but is also designed to benefit from some degree of supervision for the morphological features, rather than them being entirely unsupervised. Code for the second piece of work is freely available online.

1.2.3 WP3: Structure Induction at Sentence Level

WP Leader: UVA
Participating partners: UEDIN (14.88/36 pm), UVA (24.2/36 pm))

Task T3.1: Modeling latent alignments We investigate an unsupervised neural model for alignment that builds upon the simple factorisation of a classic statistical model, namely, the IBM model 1 (Brown et al., 1993). This model can be thought of as a very simple NMT model which generates the target sen- tence one word at a time, each time translating an independently selected subset of source tokens. Because the space of subsets grows exponentially large with source sequence length, we employ variational inference learning via approximate marginalisation with Monte Carlo (MC) methods. We investigate both discrete and approximately discrete alignments and find that by employing some variance reduction techniques discrete alignments are viable and in fact outperform approx- imately discrete ones. Whereas the model shows to be a good alignment model, initial attempts at combining it with a complete NMT model did not lead to improved translation quality. We are currently developing alternative uses of this model for other work packages, for example, as a model of sentence alignment (in WP1) and for unsupervised discovery of sub-word units (in WP2).

Task T3.2: Structured sentence models We made progress on three fronts, one focused on learning with unobserved variables, this aims at advancing technology that will enable latent structure in NMT, another focused on inducing sentence-level structure within NMT, and the last one focused on making use of context beyond the sentence level. On the first front, we (i) improve variational inference for text generation problems by addressing a common failure mode of deep latent variable models known as posterior collapse, (ii) develop sparse relaxations to binary random variables that admit unbiased reparameterised gradients, (iii) improve variational inference for language models with latent syntactic structure, and (iv) improve variational inference for discrete combinatorial structure. On the second front, we (i) develop a joint generative model that induces latent representations of sentence pairs, and (ii) build discrete latent factors into this model. On the last front, we (i) unify sentence-level and document-level translation, (ii) create evaluation resources and present a systematic evaluation of alternative architectures for representing global context, and (iii) investigate the effect of sub-document information.

Task T3.3: Probabilistic neural machine translation In the proposal, Task T3.3 corresponded to "Multilingual learning for sentence structure", aiming to derive structurally informed models across multiple language pairs. However, Task T3.3 turned out too closely-related to T3.2 as well as to the overall goals of work package 4. Learning sentence-level structure is the main objective of T3.2, and T3.3 does not seem to require new methodology as far as modelling with latent variables goes. Thus it was decided to change the task to "Probabilistic neural machine translation" as it would help meet the objectives of the project better. No alterations are needed to the structure of the milestones of the deliverables due to this change.

The probabilistic point of view has a major advantage, namely, uncertainty management. Uncertainty is sometimes seen as a problem, but we argue this is mostly so because of the ways in which predictions are traditionally formed, namely, via mode-seeking search algorithms such as beam search. Harnessed well, it can be used to make the most out of little data. We proposed to

exploit the implications of probabilistic training of NMT models. Moreover, we aim to advance NTM's data efficiency by improved probabilistic regularisation. With resources dedicated in the last quarter, we have already shown a great deal of evidence, in particular, in low- resource settings, for improvements due to a better probabilistic account to NMT.

This workpackage has so far resulted in six conference publications and various pre-prints and theses. There have also been 11 different software repositories released for research completed in this workpackage.

1.2.4 WP4: Transfer Learning

WP Leader: UEDIN
Participating partners: UEDIN (12.78/24 pm), UA (2.32/14 pm), UVA (7.3/9pm)

Task T4.1: Learning from Multilingual Data In this task we take advantage of the similarity between languages to improve the quality of machine translation between English and low-resource languages by leveraging resources for other languages. For most low-resource languages it is possible to identify one or more related medium-resource or high-resource languages that have more and better parallel data with English, which we can exploit using model pretraining and data augmentation approaches.

We performed an extensive analysis of the similarity of languages for the purposes of machine translation, which was recognized by the academic community with a best paper award, we collected and released a corpus of parallel sentences between English and 99 languages, we carried out research on techniques to to best exploit multilingual data which we applied to our English-Gujarati and English-Tamil models that we delivered to the project user partners. We reported our progress in deliverable D4.1 section 2.

Task T4.2: Learning from Monolingual Corpora Monolingual text is widely more available than parallel text, especially for the news domain. Exploiting this vast data can greatly improve the quality of machine translation systems, especially for low-resource languages. In this task we focus on researching and applying these techniques.

We investigated cross-lingual word embedding induction, probabilisitc translation models and semi-supervised machine translation. We successfully applied these techniques to our English-Gujarati, English-Tamil and English-Turkish models that we delivered to the project user partners. We reported our progress in deliverable D4.1 section 3.

Task T4.3: Learning from Lexical Resources In this task we seek to exploit the bilingual dictionaries that are available between English and most languages, even low-resource ones. While these dictionary may be small, they are curated by linguists or domain experts and they are therefore high quality and can provide the appropriate terminology for a specific domain which might not be well reflected in general parallel training corpora.

We conducted a study on different techniques to leverage dictionaries for the French-Breton language pair. We plan to expand this line of research for English-Macedonian translation. We reported our progress in deliverable D4.1 section 4.

1.2.5 WP5: Integration and Evaluation

WP Leader: BBC
Participating partners: UEDIN (1.7/7 pm), UVA (0/2pm), BBC (24.5/54pm), DW (16.2/35pm)

Task T5.1: Requirements gathering The full description of the first two use cases is described in D5.2 – Use Case Description and Requirements. This document describes the use case definition and user requirements for the GoURMET project. GoURMET envisages three different use cases: global content creation, and media monitoring are the primary use cases and currently under development. In the third year, a third use case on financial journalism will be added. The use case definition as described in this deliverable forms the basis use cases, meant to improve currently existing procedures at the two participating user partners, i.e., Deutsche Welle and BBC. It encompasses a description of the as-is situation, as well as the targeted situation, personae descriptions, a common use case model and detailed user scenarios, serving to develop the use case and prototype.

The requirements derive from the use cases. D5.2 describes the current as-is workflow (for both the editorial and monitoring workflow at Deutsche Welle and BBC World Service and Monitoring) and also the to-be workflow. The as-is situation highlights requirements coverage and describes current personae involvement. The to-be situation, on the other hand, is meant to understand the new (targeted) workflow. It reviews prioritised requirements from a workflow point of view and identifies resource requirements.

This deliverable also describes detailed scenarios, with a scenario model which shows the interaction between the different personae, the envisaged activities and the GoURMET platform. Finally, we included the list of user requirements as derived from the personae, work flow descriptions and detailed scenarios.

The development of the Financial Use Case is scheduled for the second half of the project.

Task T5.2: Creation of shared interfaces The full description of the creation of shared interfaces is outlined in D5.3 - Initial Integration Report. This report describes how the translation models, supplied by research partners, have been integrated and are made available via a cloud-based translation service. It also covers how this approach makes the translation models available for use in prototypes without the need to integrate the translation models directly and how this centralised approach to integration makes maintenence and updates to models easier.

Task T5.3: Platform integration and deployment Based on requirements gathered in T5.1 the BBC and DW have worked to integrate the research technology into existing platforms as well as developing new prototypes to serve needs identified that cannot be fulfilled using preexisting platforms.

First steps have been taken towards integration of GoURMET for real-time translation of BBC News live pages between languages, and for use in BBC's Dashboard style tool monitoring prototype.

Two GoURMET translation modules (Serbian and Bulgarian) have been integrated and deployed into DW's multilingual audiovisual production platform called plain X, allowing automated translation, subtitling and voice-over - with subsequent human post-editing for quality assurance - as part of the editorial process at Deutsche Welle.

The available GoURMET translation modules have also been integrated into DW's widescale benchmarking operation.

In the second half of the project both user partner organisations will look to continue this work and further integrate the research technology.

Task T5.4: Media monitoring: user evaluation D5.4 - Initial Progress Report on Evaluation describes the strategy, the plan and the methodology for evaluation, and interim results for both automatic and human evaluations. The evaluation of the research technologies is a key aspect of the GoURMET project and the initial Evaluation Plan D5.1 outlined the approach that will be taken. The media monitoring evaluation focuses on MT from the low-resourced language into English.

Evaluation of the languages of the first and second batches for monitoring purposes has been performed, using the gap filling method, with automated evaluation as well as human assessment by editors from both BBC and DW. A comparative result analysis shows which of the languages perform best.

Task T5.5: Global content creation: user evaluation Similar to T5.4, D5.5 Initial Progress Report on Evaluation describes the strategy, the plan and the methodology for evaluation, and interim results for both automatic and human evaluations. Also here, the GoURMET project and the initial Evaluation Plan D5.1 outlined the approach taken. This evaluation focuses on translation from English into the low-resourced language.

For the purpose of multilingual production, we have evaluated the languages of the first and second batches (with the second batch still ongoing) using Direct Assessment, in which editors from both user partners (BBC and DW) compare MT output with human translation output (reference) into the low-resourced language(s). Combined with automated evaluation, this produces a comparative result analysis with an indication of which of the languages perform best.

In addition to the gap filling and Direct Assessment evaluation, the GoURMET modules are evaluated as part of Deutsche Welle's widescale benchmarking activities, with evaluation at word, sentence and document level, using automated as well as human assessment by editors (native speakers of the low-resourced languages).

1.2.6 WP6: Dissemination and Exploitation

WP Leader: DW
Participating partners: UEDIN (0.4/7 pm), UA (1.35/6pm), UVA (0/2pm), BBC (1.4/4pm), DW (5.3/15pm)

Task T6.1: Dissemination and Communication This task is about establishing the dissemination strategy and doing continuous communication and awareness activities through the website, social media and events to inform and engage with its diverse target groups. Major achievements of the task include the setting up of the project fact sheet (D6.1 M1), the creation of the website (D6.2 M4), the early set up the dissemination strategy which eventually led to the interim dissemination and exploitation report (D6.4 M18). Next to these activities every project partner contributed in keeping the website and the social media channels populated with project related news and relevant blog posts. A module-based "dissemination kit" has been developed, adapted to changing requirements of events, target groups and communication channels. This kit contains a set of key visuals of the GoURMET project such as flyers, poster and banner and has been used by partners at 19 academic and industry dissemination events. In the first half of the project 20 publications have been published including an award-winning paper for multilingual NLP tasks and over 45 data, model and software components have been publicly released as Open Source on the website.

Task T6.2: Exploitation This task explores options to ensure the outputs of the GoURMET project can be exploited by the partners themselves and others. An exploitation committee has been established to set up and execute the exploitation strategy of the consortium as well as to ensure that IPR Management is being carried out appropriately. There are two main ways in which GoURMET can be exploited: Component-Based Exploitation (for improving existing applications) and Platform Exploitation (GoURMET as a whole). Both user partners BBC and Deutsche Welle have created infrastructure to ensure exploitability at the two partner sites. The integration of demonstrators have already been initiated for content production and media monitoring at BBC and Deutsche Welle. Additionally, Deutsche Welle has started to make GoURMET part of DW benchmarking activities. Major work of the exploitation task will be undertaken in the second half of the project with investigations into the revenue potential of the best options.

1.2.7 WP7: Management

WP Leader: UEDIN
Participating partners: UEDIN (7.81/24 pm), UA (0.49/5pm), UVA (0/2pm), BBC (2.5/6pm), DW (1.7/5pm)

Task T7.1: Project management Project management consists of guiding and monitoring key areas of collaboration within the project.

Project structure: There were no changes to the project structure in the first half of the project. The project management approach has worked well, and no changes have been necessary.

Reporting: All required reporting has been carried out.

Internal communication: Communication between the project partners has been facilitated by using Slack, a number of monthly video conferencing meetings each with a different focus (general project management, research, innovation, occasional WP meetings), an email list, and an internal project wiki (confluence). There have been two in person meetings, with a further two held online due to COVID-19.

Quality assurance and risk management: The project put quality assurance proceedures in place, for example ensuring that there is a named reviewer for each deliverable. We reviewed project progress against the milestones in an ongoing fashion, and project risks were reviewed, and there were no risks with high likelihood and high severity.

Extension The project has been granted a no-cost 3 month extension to manage the effects of COVID on delayed in-person events, and reduced working capacity due to extra caring responsibilities. Most milestones and all deliverables have been moved to 3 months later. Please see D7.3 for the full list of changed deliverable and milestone dates.

Task T7.2: Innovation management This task is concerned with the coordination of innovation activities such as prototyping, testing, demonstrating. As such it has involved coordinating activities largely undertaken in WP5, and WP6 and in the relevant deliverables (D5.2, D6.3). In March we set up an Exploitation Committe which meet once a month. This is comprised of the user partners and the scientific co-ordinator. We will be responsible for more focussed planning of events such as the surprise language task, the hackathon and the sustainability plan.

Task T7.3: Ethics management The GoURMET ethics process is based around an Ethics Committee, comprised of a representative from each partner. The ethics process has grouped the ethical issues which arise into two broad categories: Protection of personal data; and Broader ethical concerns including dual-use (military and defence) implications of GoURMET technologies, and the social impact of the tools and technologies developed by GoURMET. These issues, and the projects response to them, are discussed in D7.1.

Task T7.4: Data Management The data management plan (D7.2) provides an analysis of the main elements of the data management policy that have been used by the GoURMET consortium with regard to all the datasets collected for or generated by the project. It addresses issues such as collection of data, data set identifiers and descriptions, standards and metadata used in the

project, data sharing, property rights and privacy protection, and long-term preservation and reuse, complying with national and EU legislation.

The data management plan also presents the project's privacy strategy, including the protective measures developed to address explicit consent for processing personal data, and the storage, transmission, processing, and analysis of personal data.

1.3 Impact

The information in section 2.1 of the DoA is still relevant, and no update is needed.

2 Deviations from Annex1 and Annex2

2.1 Tasks

All tasks were fully implemented and all critical objectives were fully achieved to schedule.

2.2 Use of resources

Figures 1 and 2 show the use of resources in terms of costs and effort during the first period of the project. The expenditure numbers are the final amounts to the end of the 18 month period. The numbers for person month activity reflect figures to the end of month 17 (except for Alicante who report these numbers to month 18). About 38% of project resources have been used in the first period. The work is proceeding according to plan with about 41% of the total predicted effort expended.

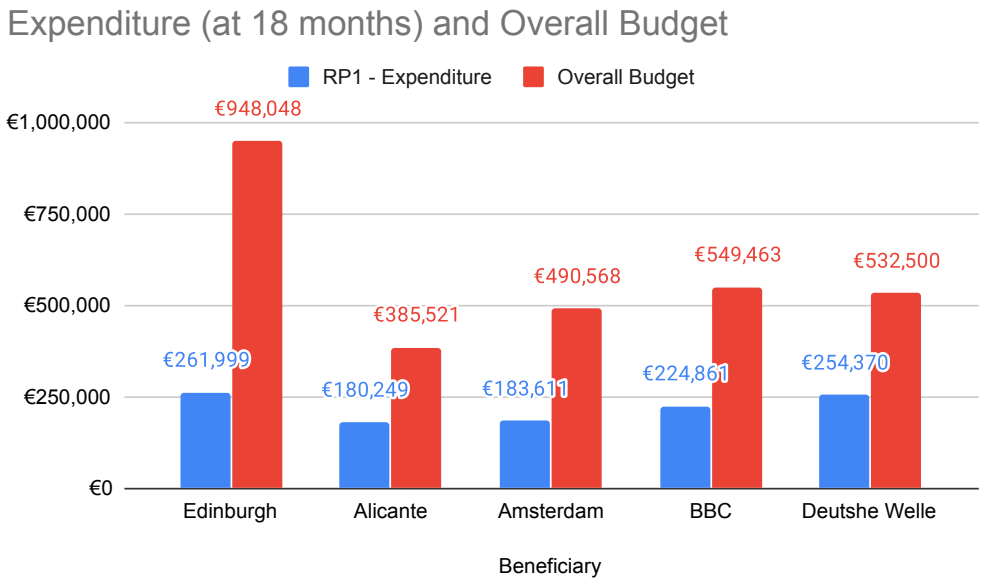

Figure 1: Summary of period 1 costs by partners, showing expenses incurred vs overall budget. 38% of the total budget has been spent (€ 1,105,090 of € 2,906,099)

Below we outline the use of resources for each partner and each WP

UEDIN

WP1 (5.96/12 p-months): according to plan
WP2 (15.77/36 p-months): according to plan
WP3 (14.88/36 p-months): according to plan
WP4 (12.78/24 p-months): according to plan
WP5 (1.7/7 p-months): increased effort in integration and evaluation expected during M19–39
WP6 (0.4/7 p-months): increased effort in dissemination and exploitation was planned for M19–39;
WP7 (7.81/24 p-months): increased effort in reporting and exploitation was planned for M19–39.

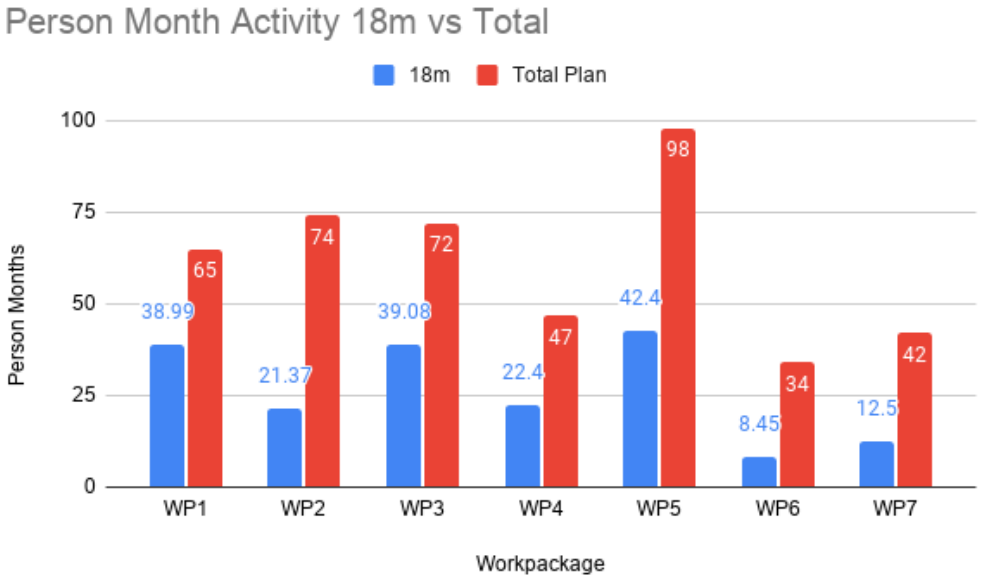

Figure 2: Summary of period 1 effort by WP, showing effort applied (to M18) vs total amount of effort planned. 43% of the person months have been expended (185.2 of 432)

UA

WP1 (25.23/40 p-months): according to plan
WP2 (5.6/14 p-months): according to plan
WP3 (0/0 p-months): n/a
WP4 (2.32/14 p-months): according to plan
WP5 (0/0 p-months): n/a
WP6 (1.35/6 p-months): according to plan.
WP7 (0.49/5 p-months): according to plan.

UVA

WP1 (0/0 p-months): n/a
WP2 (0/24 p-months): according to plan; increased effort in final project phase M18-M36
WP3 (24.2/36 p-months): according to plan
WP4 (7.3/9 p-months): according to plan
WP5 (24.5/54 p-months): according to plan;
WP6 (1.4/4 p-months): according to plan; increased effort in initial project phase M1-M18
WP7 (0/5 p-months): according to plan.

BBC

WP1 (0.5/3 p-months): according to plan;
WP2 (0/0 p-months): n/a
WP3 (0/0 p-months): n/a
WP4 (0/0 p-months): n/a
WP5 (24.5/54 p-months): according to plan;
WP6 (1.4/4 p-months): according to plan;

WP7 (2.5/6 p-months): according to plan.

DW

WP1 (7.3/10 p-months): according to plan;
WP2 (0/0 p-months): n/a
WP3 (0/0 p-months): n/a
WP4 (0/0 p-months): n/a
WP5 (16.2/35 p-months): according to plan;
WP6 (5.3/15 p-months): according to plan;
WP7 (1.7/5 p-months): according to plan.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

First Periodic Technical Report Part B