Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D6.4 – Final Dissemination and Exploitation Report

Nature	Report	**Work Package**	WP6
Due Date	30/06/2022	**Submission Date**	30/06/2022
Main authors	Kay Macquarrie (DW), Peggy van der Kreeft (DW)		
Co-authors			
Reviewers	Sevi Sariisik Tokalac (BBC)		
Keywords	Dissemination, Exploitation		
Version Control			
v0.1	**Status**	Draft	21/05/2022
v1.0	**Status**	Final	28/06/2022

Contents

List of Figures

List of Tables

1 Abstract

This deliverable D6.4 relates to Work Package 6 “Dissemination and Exploitation” and provides an overview of the dissemination and exploitation activities and achievements for GoURMET for the reporting period of M0-M42 (1.1.2019-30.6.2022).

2 Introduction

This report is divided into two main parts: Section 3 (Dissemination) and Section 4 (Exploitation). These relate to Tasks 6.1 and 6.2 respectively.

Section 3 (Dissemination) shows activities to inform and engage with other researchers and potential users of the GoURMET models and the platform as well as about the project's intentions and results. The project has contributed to and inspired other (low-resource) language translation projects with the goal of building networks and showing that translation technologies can contribute to more efficient and broader news coverage as well as media monitoring activities in under-represented and under-served languages.

Section 4 (Exploitation) outlines the activities aimed at successful exploitation of GoURMET. There are two main ways in which GoURMET is being exploited: Model-Based Exploitation (single applications) and Platform Exploitation (GoURMET as a whole). This section elaborates on the highly promising results in both of these areas.

3 Dissemination Report

This section of the deliverable relates to Task 6.1 (Dissemination), focusing on providing visibility of the project results in the scientific community, the broader community of users and stakeholders and other related research and innovation projects. It provides the final impact report.

3.1 Dissemination Strategy

GoURMET's dissemination and communication strategy is outlined here. It presents the guidelines and procedures for communicating internally and externally and disseminating the results of the project.

Strategy The project's communication strategy (cf. D6.3) has laid down how the communication works internally, with peer researchers and other stakeholders, and with the public at large - and how this communication has been implemented efficiently.

The consortium disseminated information regarding the project goals, research, results and experiences to industrial communities (SME and Industry), to academic and research institutions, as well as to the general audience interested in the project. Project results were promoted and disseminated during the entire project, as an appropriate prerequisite for a successful exploitation, and at the end of the project in order to engage its stakeholders.

Target Groups The consortium targeted the following diverse range of interested groups as part of the dissemination activities (see Table 1).

TARGET GROUP	HOW GOURMET REACHED OUT TO THEM (EXAMPLES)
Peer research groups in academia in the EU and globally	Through participation in academic events (see: 3.4.1 Academic Events), publications (see: 3.5 Publications) and provision of research output publicly available (open-source) on GitHub
Peer research groups in other H2020 projects	Through a dedicated place on the GoURMET website ("related projects"), addressing them on Social Media (Hashtags, Mentions)
Specialist press in research and technology sectors	Through presentation and participation at events such as the QURATOR Conference
International broadcast technology users	Through presentation at EBU events
Key stakeholder groups for all project partners	Through dissemination on the GoURMET project website and on corresponding partner websites
Diverse industry sectors: broadcasting, media monitoring, translation	Through participation at various industry events (see: 3.4.2 Industry Events)
Policymakers and interest groups	Through participation in conferences such as META FORUM

Users of the GoURMET platform, including journalists, editors, monitors, analysts	GoURMET continuously informed potential users via its website and Social Media channels, and organized two workshops (see: 3.3 GoURMET Events)
Educational outreach audiences (the project seeks to encourage young people into scientific careers by showcasing the interesting challenges tackled by GoURMET)	GoURMET encouraged audiences to use project results (cf. https://gourmet-project.eu/data-model-releases/) and as part of the Open Workshop GoURMET ran an Ideation session which also addressed younger people to engage with language technologies

Table 1: GoURMET Target Groups and how they were addressed

3.2 Dissemination: Channels, Materials, Coverage

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 and was used for the various channels, such as online media including the project website and social media channels as well as for offline media including banner and poster sessions.

The development of the Corporate Design has been completed in the initial phase of the project and has been documented in D6.3 Interim Dissemination and Exploitation Report.

3.2.1 Website

The website www.gourmet-project.eu was continuously updated throughout the project and served as the primary online communication channel. It has been designed as a central information site with updates about the latest developments and achievements of the project. In the blog section, we highlighted key events and releases of the website.

More than 6400 visits were registered throughout the course of the project (M0-M42). Connections came from all over Europe, especially Germany and Spain, but also from the United States, Africa, South America and Asia.

The page that has been most frequently visited by users, with over 5200 views, was the "GoURMET Project" page followed by the "Project Output" page including the freely available models and datasets with over 1900 views. These numbers indicate a high interest in GoURMET's language technology resources.

3.2.2 Social Networks

Twitter and LinkedIn have vibrant communities involved in the machine translation and HLT fields. Sustaining a presence on these platforms, with DW providing dissemination has helped inform the wider community and thus sustained awareness of the project developments. Additionally, it provided a feedback channel allowing followers to easily and directly engage with the GoURMET project, either by following, sharing or commenting.

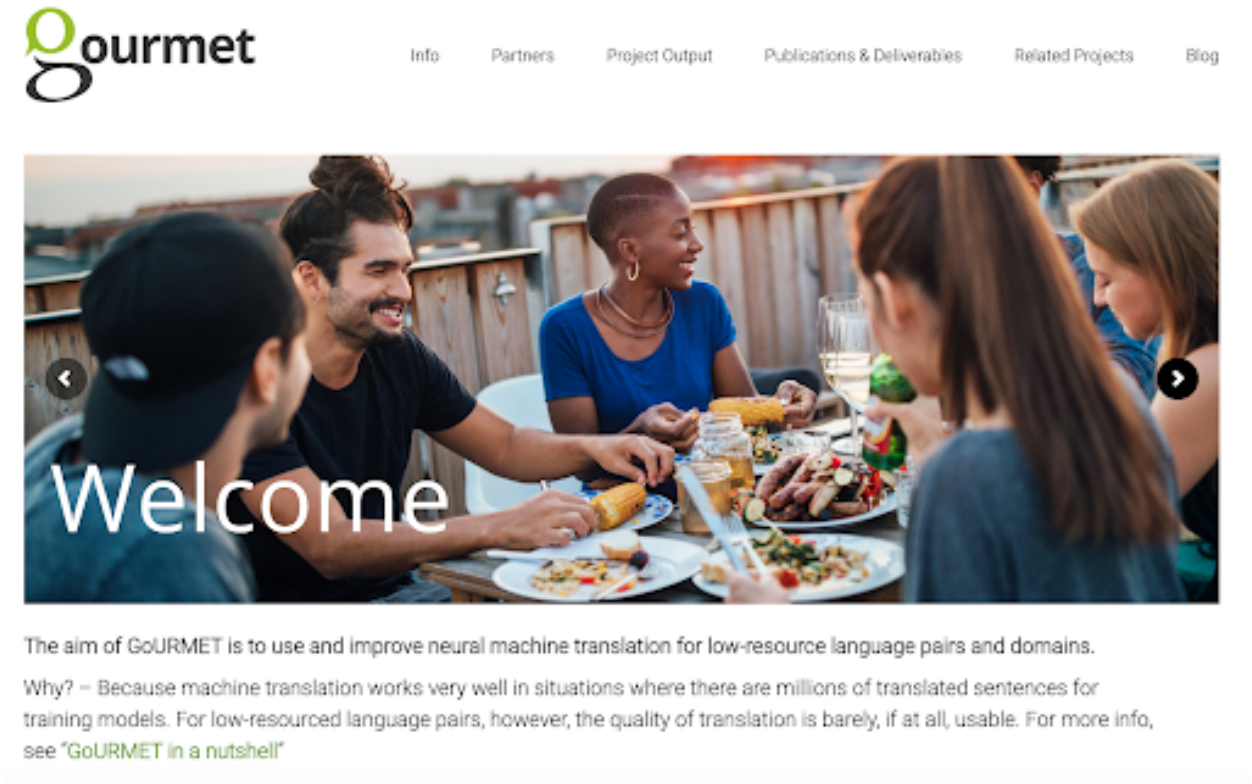

Figure 1: Screenshot showing the global distribution of the website's visits (6445 visits in total)

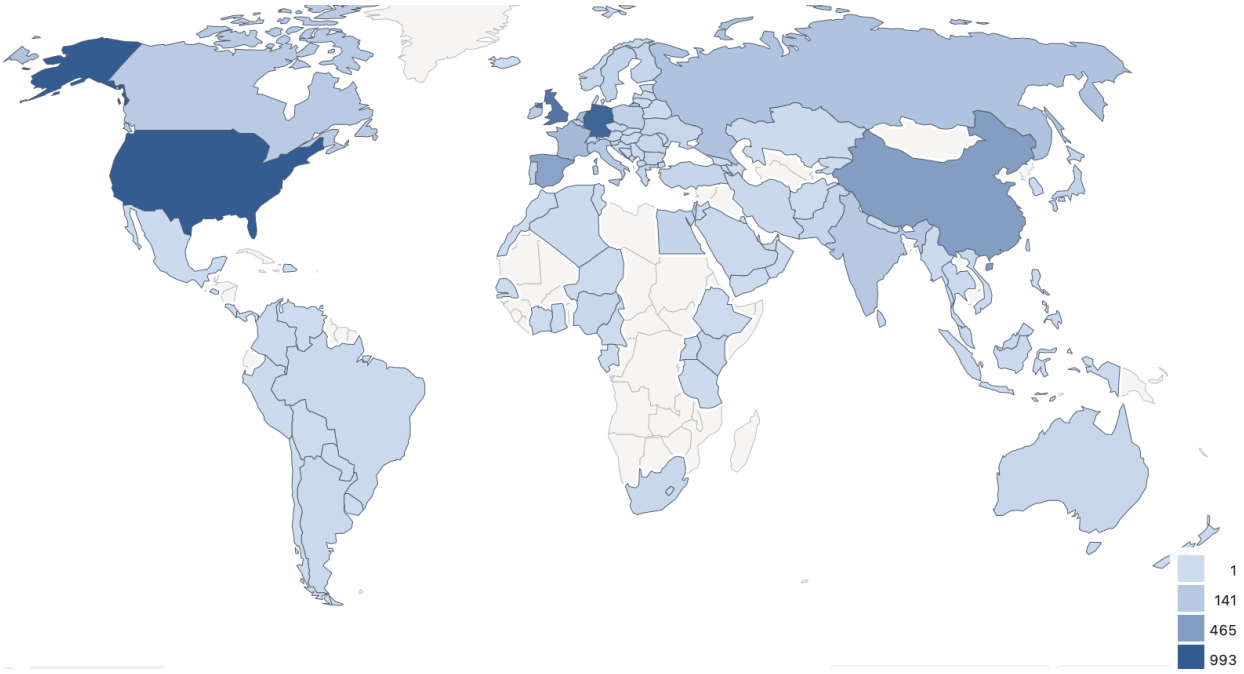

Figure 2: Screenshot showing the global distribution of the website's visits (6445 visits in total)

Twitter

Twitter has been intensively used to inform communities about GoURMET activities and progress. By more than 160 tweets, we created an ongoing awareness of the developments in the project in

specific target groups. During the course of the project, the GoURMET Twitter account doubled its reach with almost 300 followers.

Figure 3: One of the most popular GoURMET tweets with 35 retweets announcing the introduction of the Pashto model to the GoURMET Translate tool

The GoURMET Twitter account has evolved to be a central and successful channel to promote GoURMET's achievements and engage with an interested audience.

LinkedIn

The GoURMET LinkedIn account was used to promote major news and updates on the project specifically in respect to its output formats for further exploitation purposes. In the course of the project the project was able to obtain 60 followers, also more than doubling the numbers in respect to the previous report. The figure below demonstrates how the final user event was promoted, which also listed potential attendees.

3.2.3 GoURMET in the Media

This is a list of GoURMET appearance in the media, mainly as articles in newspapers or posts in partner blogs.

1. GoURMET Kickoff

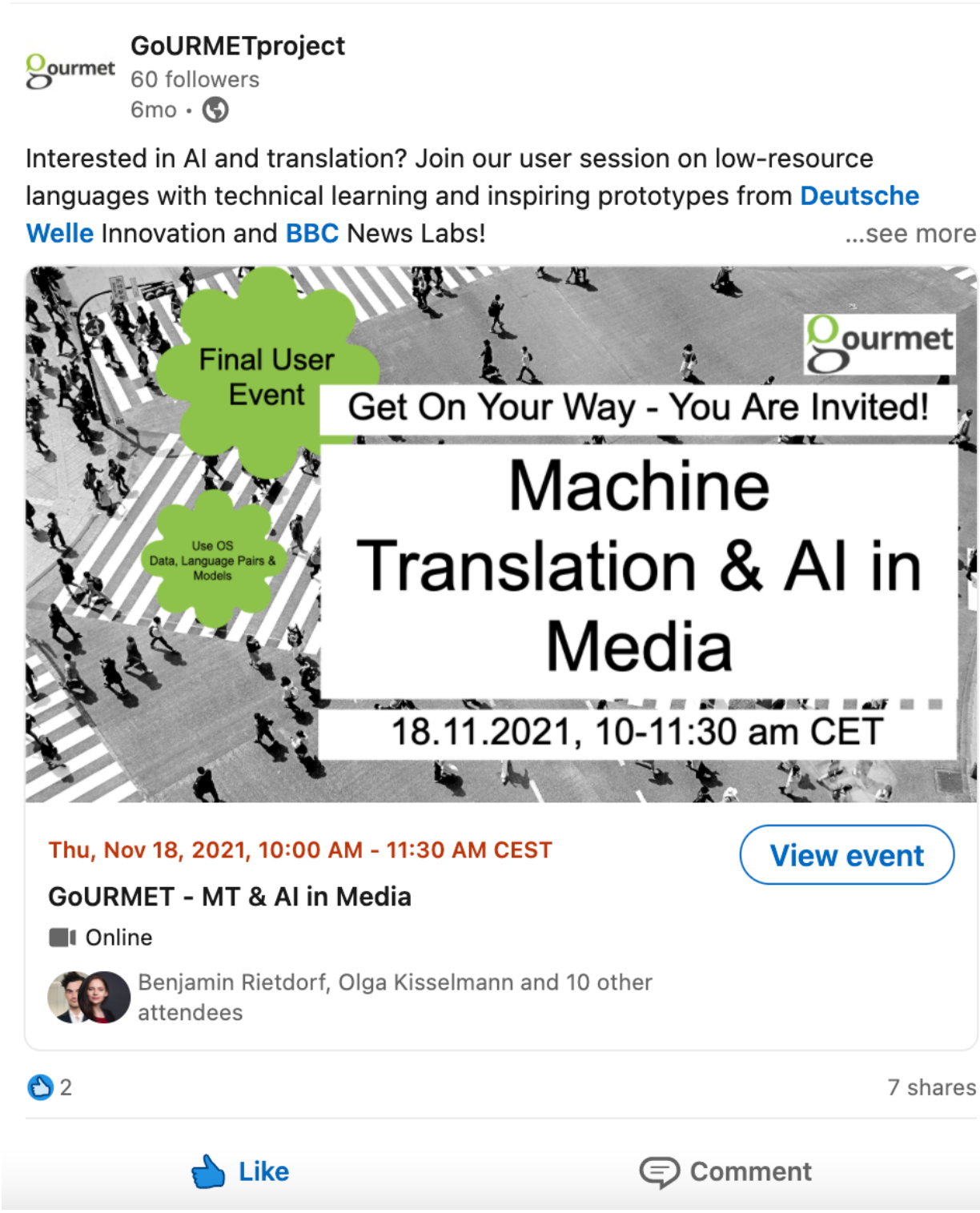

Figure 4: Announcing the GoURMET User Event on LinkedIn

- Coverage in news agencies and digital newspapers: La Vanguardia, Europa Press, Información, 20 Minutos, Intercomarcal, Network of Valencian Universities for Promoting Research & Innovation, NovaCiencia
- Blog post on BBC News Labs, https://bbcnewslabs.co.uk/projects/gourmet/

- Blog post on DW Innovationblog, https://blogs.dw.com/innovation/new-hlt-project-gourmet-to-improve-machine-translation-for-low-resource-languages-and-domains/

2. Provision of a Report to European Broadcasting Union's News Report 2019
   - Report by BBC and DW, November 2019
3. Release of GoURMET Translate, June 2020
   - Blog post on DW Innovations Blog https://innovation.dw.com/gourmet-project-new-free-web-tool-allows-for-translation-into-seven-low-resource-languages
4. Release of GoURMET Data, Models Software for LR Machine Translation, June 2022
   - Blog post on DW Innovations Blog (soon to come) https://innovation.dw.com

3.3 GoURMET Dissemination Events

GoURMET organized two workshops: The GoURMET Open Workshop and the Final User Event. Due to the Covid-19 pandemic both events were reshaped into virtual events using a Zoom platform hosted by University of Edinburgh. Although the limitations to the virtual event came in with certain restrictions it allowed for more geographically diverse contributors and audiences. Deutsche Welle led and coordinated both events with partners contributing with expertise, contacts and leading sections.

3.3.1 Open Workshop

The GoURMET Open Workshop brought together researchers, developers, users and all interested parties engaged in computational processing of multilingual media content - particularly in the machine translation of media to and from low-resourced languages. It took place on three consecutive days between 25-27 May 2021 for three hours in the afternoon each day. As part of the workshop we ran a 90-minute ideation session ("MT for news production and media monitoring") organized and moderated by BBC News Labs (for more information see Appendix). On average, we had more than 50 participants each day listening and interacting across 12 sessions. As guest speakers we had representatives from organizations and projects including: Swedish Radio, The Masakhane open-source project, Translators without Borders, Monition project, SELMA project, Priberam, Fraunhofer IAIS, University Avignon, University of Latvia and EBU Newspilot.

Further information and presentation slides can be found on the GoURMET website: https://gourmet-project.eu/open-workshop-2021/

3.3.2 Final User Event

The GoURMET User Event took place in November 2021. It focused on prototypes, results and learnings for journalists and the media industry. The event drew insights about machine translation in low-resource language domains from the research partners (University of Edinburgh, Universitat d'Alacant, University of Amsterdam) and showcased our output including live demos from BBC

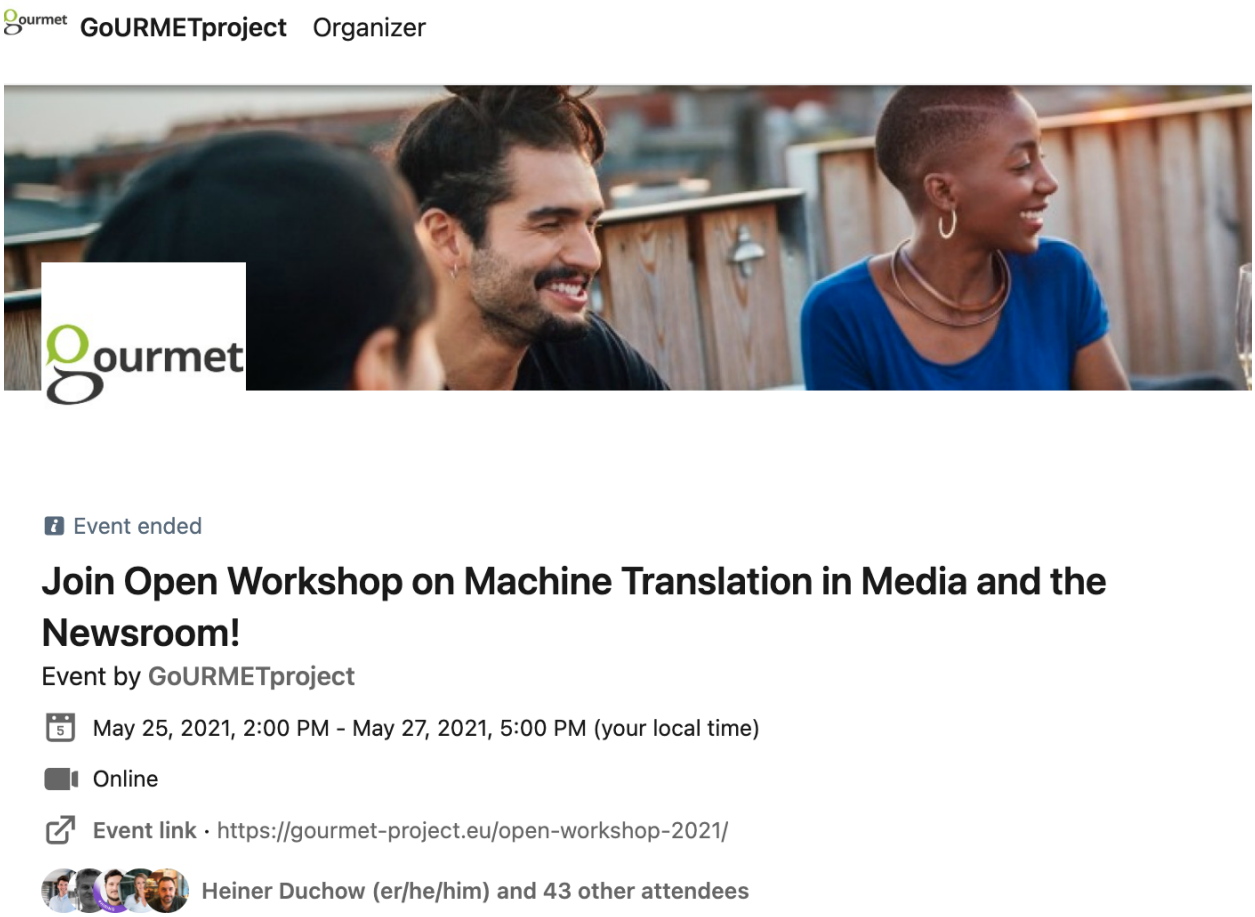

Figure 5: The Open Workshop was successfully promoted on the GoURMET social media channels. The picture above shows the event on LinkedIn indicating 43 attendees

and Deutsche Welle. During the 90-minute user event we reached out to 56 participants. A recording was also made available demonstrating the vivid discussions amongst participants on using AI in the media.

Further information and presentation slides can be found on the GoURMET website: https://gourmet-project.eu/project-output/user-event-2021/

3.4 List of Events

The following two chapters list the events at which the GoURMET partners have represented the project. It should be noted, that the COVID-19 pandemic severely impacted our attendance at events in 2020-2022.

3.4.1 Academic Events M0-M42

GoURMET partners have actively participated in the following academic events (see Table 2 and 3). The number of events visited after the interim report adds up to 16.

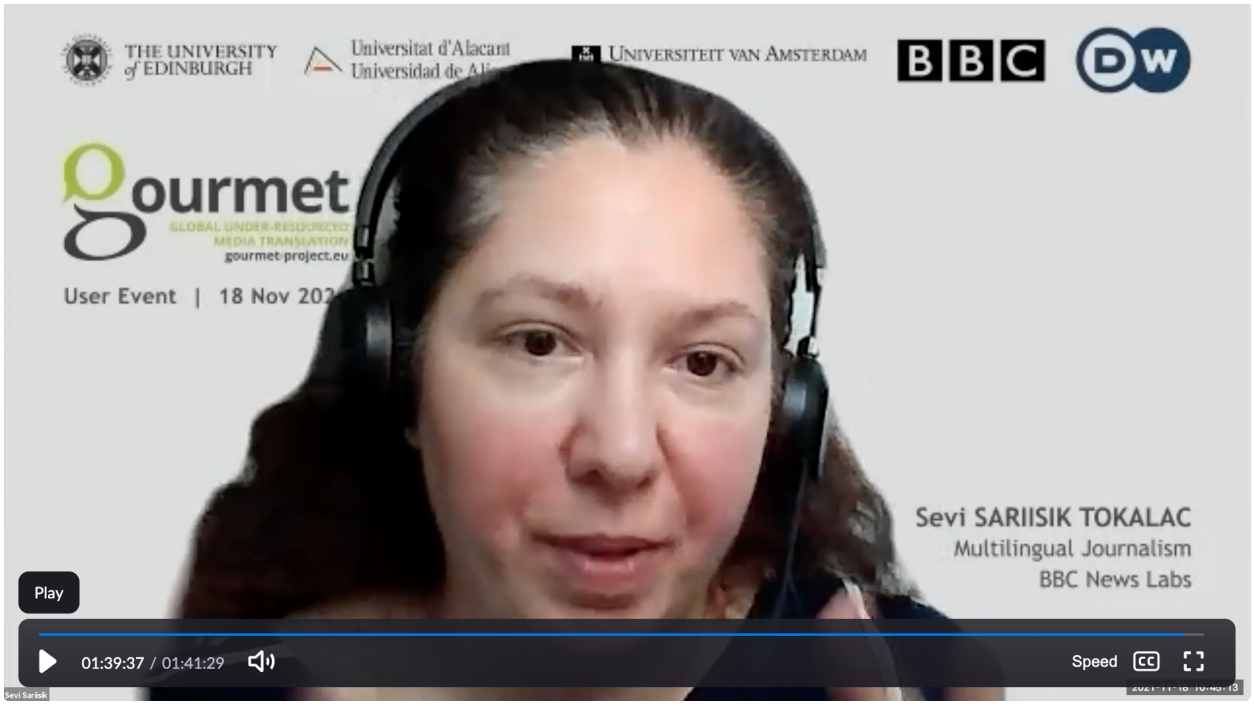

Figure 6: The screenshot of the User Event recording. The backdrop was customized and promotes a link to the GoURMET website

DATE	PLACE	EVENT	DISSEMINATION ACTIVITY	PARTNER
30.07.2019	South Africa, University of KwaZulu-Natal	Machine Learning Workshop	Tutorial One day tutorial on machine learning and data curation mainly for neural machine translation	UEDIN, Alexandra Birch
2019	Online	Lisbon Machine Learning Summer School	Attended	University of Amsterdam, Bryan Eikema
29.07-01.08.2019	Online	ACL 2019	Conference, Co-presented paper	University of Amsterdam, Wilker Aziz, Bryan Eikema
2.08.2019	Florence, Italy	RepL4NLP 2019	Presentation	UA, Bryan Eikema, Wilker Aziz
1-2.08.2019	Florence, Italy	WMT 2019	Presentation (top constrained system for English-to-Gujarati)	UEDIN, Rachel Bawden, Faheem Kirefu, Antonio Valerio, Miceli Barone

1-2.08.2019	Florence, Italy	WMT 2019	Presentation (best performing submission for the English-Kazakh language pair)	University of Alicante, Víctor M. Sánchez-Cartagena
2.11.2019	Hong Kong	WNGT2019	Organiser at Workshop on Neural Generation and Translation	UEDIN, Alexandra Birch
4-6.11.2019	Hong Kong	EMNLP2019	Conference, Co-presented paper	UEDIN, Alexandra Birch
21-23.01.2020	Alcant, Spain	GoURMET tutorial on variational inference	Three day tutorial on variational inference. Tutorial organised by GoURMET and the Universitat d'Alacant's Institute for Computer Research.	UA, Bryan Eikema, Wilker Aziz
26.4-01.05.2020	Online	ICLR 2020	Conference, Co-presented paper	University of Amsterdam, Wilker Aziz

Table 2: Academic Events January 2019 - June 2020

DATE	PLACE	EVENT	DISSEMINATION ACTIVITY	PARTNER
06-08.07.2020	Online	ACL 2020	Conference, presented paper	University of Amsterdam, Wilker Aziz
03-05.11.2020	Online	EAMT 2020	Conference, Co-presented paper	University of Alicante, Víctor M. Sánchez-Cartagena
03-05.11.2020	Online	EAMT 2020	Conference, Co-presented paper	University of Alicante, Mikel Forcada
2020	Online	EMNLP 2020	Conference	University of Edinburgh, Alexandra Birch, University of Amsterdam, Bryan Eikema

19-20.11.2020	Online	WMT 2020	Conference, Co-presented paper	University of Alicante, Miquel Esplà-Gomis, Víctor M Sánchez-Cartagena, University of Edinburgh, Barry Haddow, Rachel Bawden
8-13.12.2020	Online	COLING 2020	Conference, Co-presented paper	University of Alicante, Víctor M. Sánchez Carragena, University of Amsterdam, Bryan Eikema, Wilker Aziz
01.2021	Online	ALPS Winter School	Winter school, presentation	University of Amsterdam, Bryan Eikema
03.2021	Online	Principled Statistical Modeling with Stan	Course	University of Amsterdam, Bryan Eikema
16.08.2021	Online	LoResMT 2021	Workshop	University of Edinburgh, Barry Haddow
2021	Online	MTSummit 2021	Conference	University of Edinburgh, Alexandra Birch
7-11.11.2021	Online	EMNLP 2021	Conference	University of Alicante, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez
2021	Online	EMNLP 2021	Conference	University of Edinburgh, Alexandra Birch, University of Amsterdam, Bryan Eikema
10-11.11.2021	Online	WMT 2021	Conference	University of Edinburgh, Barry Haddow
26-27.05.2022	Dublin, Ireland	SPNLP 2022 (colocated with ACL 2022)	Presentation	University of Amsterdam, Wilker Aziz

23-27.05.2022	Dublin, Ireland	ACL 2022	Conference	University of Edinburgh, Alexandra Birch, Antonio Barone, University of Amsterdam, Bryan Eikema (online)
1-3.6.2022	Gent, Belgium	EAMT 2022	PPaper: GoURMET – Machine Translation for Low-Resourced Languages, Peggy van der Kreeft, ed al, European Association for Machine Translation Conference	DW, Peggy van der Kreeft,

Table 3: Academic Events July 2020 - June 2022

3.4.2 Industry Events M0-M42

GoURMET has actively participated in the following industry events (see Table 4 and 5). The number of events attended after the interim report adds up to 17. Two more are scheduled for July and August 2022.

DATE	PLACE	EVENT	DISSEMINATION ACTIVITY	PARTNER
08.02.2019	London, UK	MESA - Media and Entertainment Services Alliance	Conference	DW, Peggy van der Kreeft
18.04.2019	Cologne, Germany	ECIR 2019	Presentation: Human Language Technologies for Information Retrieval	DW, Peggy van der Kreeft, Alexander Plaum
3-5.4.2019	Nairobi, Kenya	Toward a Network of Excellence in Artificial Intelligence for Development (AI4D) in sub-Saharan Africa	Planning for creating data and capacity for natural language processing in African languages	UEDIN
27-28.05.2019	Bonn, Germany	GMF 2019	Presentation: Language Technologies at DW, Deutsche Welle Global Media Forum	DW, Peggy van der Kreeft

10-13.06.2019	Geneva, Switzerland	EBU-MDN Workshop	Presentation: Language Technologies at DW, EBU Metadata Developer Network	DW, Peggy van der Kreeft
11.06.2019	Alicante	Presentation of the project to EUIPO	Presentation and Workshop	UA
11-12.06.2019	London, UK	newsHACK: Tools for Multilingual Newsrooms and Talk	Hackathon Team	BBC, DW, UEDIN, UA
25.06.2019	London, UK	Managing and Translating Audiovisual Content	Invited talk	BBC
19-23.8.2019	Dublin, IR	Machine Translation Summit 2019	Poster Presentation	DW, UEDIN, UA
08.10.2019	London, UK	Ada Lovelace Day 2019	Talk	BBC
08-09.10.2019	Brussels, Belgium	META-FORUM 2019	Presentation: Language technology goals and challenges	DW, Peggy van der Kreeft
09.10.2019	Brussels, BE	META FORUM - ELG and the European LT Industry	Booth and Presentation	BBC, DW
19.11.2019	Online	European Broadcasting Union's News Report 2019	Report	BBC, DW
27-28.11.2019	Salford, UK	EBU Technology in Production and Distribution Workshop	Workshop	BBC
20-21.01.2020	Berlin, Germany	Qurator 2020	Presentation: Curation of Content Use Cases at a World Broadcaster, Curator Conference	DW, Peggy van der Kreeft
10.02.2020	online	EU/LT Workshop	Presentation: AI in Multilingual Media Analysis and Production, EC Expert Consultation Workshop on the deployment of language technologies in Europe	DW, Peggy van der Kreeft
21-22.02.2020	Bern, AT	SwissInfo Hackathon	Presentation	DW, Kay Macquarrie

08.03.2020	London, UK	International Women's Day special	Podcast	BBC
16.05.2020	Marseille, FR	International Workshop on Language Technology Platforms (Event cancelled, but Proceedings are available)	Workshop	UEDIN

Table 4: Industry Events January 2019 - June 2020

DATE	PLACE	EVENT	DISSEMINATION ACTIVITY	PARTNER
24.09.2020	Prague, Czech Republic (virtual)	AI for Journalism Conference 2020	Presentation: Automation in NLP to optimize editorial workflows for international broadcasters, Prague	DW, Peggy van der Kreeft
10.09.2020	online	ARD	Presentation: HLT at DW, ARD Technology Group, Bonn	DW, Peggy van der Kreeft
12.2020	London,UK	Show and Tell	Workshop, Presentation	BBC, Lei He
1-3.12.2020	Online	META FORUM 2020	Workshop, Presentation	DW, Peggy van der Kreeft
08.2021	London,UK	Show and Tell	Workshop, Presentation	BBC
08.2021	London,UK	BBC World Service presentations	Demo	BBC
09.2021	Berlin, Germany (online)	Languages and Media Conference	Poster Presentation	BBC, Sevi Sariisik
10.2021	London,UK	BBC Visual Journalism Team	Demo	BBC
15.10.2021	online	EBU AIDI 2021	Presentation: AI at DW – HLT Focus, EBU AI and Data Initiative, Geneva	DW, Peggy van der Kreeft
15.11.2021	Online	META FORUM 2021	Workshop, Presentation	UEDIN, Alexandra Birch, BBC, Lei He

15.11.2021	online	Goethe Institute	Presentation: Sprachtechnologien (HLT) in der Deutschen Welle - vom Prototypen zur Einführung in den Redaktionellen Alltag, Annual Meeting DW and Goethe Institute	DW, Peggy van der Kreeft
11.2021	London,UK	BBC News Awards	Presentation	BBC
04.2022	London,UK	Data Science Research Project (DSRP)	Presentation	BBC, Lei He and Sevi Sariisik
28.04.2022	online	EU Workshop Common European Language Data Space	Presentation: Common European Language Data Space – Data Management, Opportunities, Challenges - EC, Brussels	DW, Peggy van der Kreeft
1-3.6.2022	Gent, Belgium	EAMT 2022	Poster presentation: GoURMET – Global Under-Resourced Media Translation, European Association for Machine Translation Conference	DW, Peggy van der Kreeft
22.07.2022	London,UK	BBC Digital Growth Team Away day	Presentation	BBC, Sevi Sariisik
08.2022	London,UK	BBC Research and Development Session	Presentation	BBC, Sevi Sariisik

Table 5: Industry Events July 2020 - June 2022

Figure 7: Image of promotion for the final user event on Twitter

3.5 List of Publications

See the openAIRE GoURMET page for further details on these publications.

1. Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality. Carolina Scarton, Mikel L. Forcada, Miquel Esplà-Gomis, and Lucia Specia. IWSLT 2019, link

2. Block neural autoregressive flow. De Cao, Nicola, Ivan Titov, and Wilker Aziz. Arxiv 2019, link

3. Widening the Representation Bottleneck in Neural Machine Translation with Lexical Shortcuts. Emelin, Denis; Titov, Ivan; Sennrich, Rico. WMT 2019, link

4. Findings of the 2019 Conference on Machine Translation. Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post and Marcos Zampieri. WMT 2019, link

5. On the Importance of Word Boundaries in Character-level Neural Machine Translation. Duygu Ataman. WNGT 2019, link
6. Towards a Multi-view Language Representation: A Shared Space of Discrete and Continuous Language Features. Arturo Oncevay, Barry Haddow and Alexandra Birch. TyP-NLP 2019, link
7. Auto-Encoding Variational Neural Machine Translation. Bryan Eikema and Wilker Aziz. RepL4NLP 2019, link
8. Interpretable Neural Predictions with Differentiable Binary Variables. Joost Bastings, Wilker Aziz, and Ivan Titov. ACL 2019, link
9. The Universitat d'Alacant submissions to the English-to-Kazakh to the WMT19 News Translation Task. Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez. WMT 2019, link
10. The University of Edinburgh's Submissions to the WMT19 News Translation Task. Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, Alexandra Birch. WMT 2019, link
11. Effective Estimation of Deep Generative Language Models. Tom Pelsmaeker and Wilker Aziz. ACL 2020, link
12. A multi-source approach for Breton-French hybrid machine translation. Víctor M. Sánchez-Cartagena, Mikel L. Forcada, Felipe Sánchez-Martínez. EAMT 2020, link
13. An English-Swahili parallel corpus and its use for neural machine translation in the news domain. Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L. Forcada, Miquel Esplà-Gomis, Andrew Secker, Susie Coleman, Julie Wall. EAMT 2020, link
14. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. Biao Zhang, Philip Williams, Ivan Titov, Rico Sennrich. ACL 2020, link
15. MultiWord Expression Aware Neural Machine Translation. Andrea Zaninello, Alexandra Birch LREC 2020, link
16. Toward Making the Most of Context in Neural Machine Translation. Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen and Alexandra Birch. IJCAI 2020, link
17. A Latent Morphology Model for Open-Vocabulary Neural Machine Translation. Duygu Ataman, Wilker Aziz, Alexandra Birch. ICLR 2020, link
18. Document Sub-structure in Neural Machine Translation. Radina Dobreva, Jie Zhou and Rachel Bawden. LREC 2020, link
19. Document-level Neural MT: A Systematic Comparison. António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang and André T. Martins. EAMT 2020, link
20. Architecture of a Scalable, Secure and Resilient Translation Platform for Multilingual News Media. Susie Coleman, Andrew Secker, Rachel Bawden, Barry Haddow and Alexandra Birch. IWLTP 2020, link

2nd reporting period

1. Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation. Bryan Eikema and Wilker Aziz. Coling 2020, link
2. A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing. Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar and Matt Post. EMNLP Findings 2020, link
3. Bridging linguistic typology and multilingual machine translation with multi-view language representations. Arturo Oncevay, Barry Hadddow, Alexandra Birch. EMNLP 2020, link
4. Language Model Prior for Low-Resource Neural Machine Translation. Christos Baziotis, Barry Haddow, Alexandra Birch. EMNLP 2020, link
5. Understanding the effect of morphological tags in under-resourced neural machine translation. Víctor Manuel Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez. COLING 2020, link
6. Bicleaner at WMT 2020: Universitat d'Alacant-Prompsit's submission to the parallel corpus filtering shared task. Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, Felipe Sánchez-Martínez. WMT 2020, link
7. Efficient Marginalization of Discrete and Structured Latent Variables via Sparsity. Gonçalo M. Correia, Vlad Niculae, Wilker Aziz, and André F. T. Martins. NeurIPS 2020, link
8. Exploring Unsupervised Pretraining Objectives for Machine Translation. Christos Baziotis, Ivan Titov, Alexandra Birch, Barry Haddow, ACL (Findings) 2021 link
9. Few-shot learning through contextual data augmentation. F Arthaud, R Bawden, A Birch. EACL 2021, link
10. Rethinking data augmentation for low-resource neural machine translation: a multi-task learning approach. Víctor M.Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez. EMNLP 2021, link
11. The University of Edinburgh's English-German and English-Hausa Submissions to the WMT21 News Translation Task. Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, Kenneth Heafield. WMT 2021 link
12. Findings of the 2021 Conference on Machine Translation. Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, Marcos Zampieri. WMT 2021 link

13. Surprise Language Challenge: Developing a Neural Machine Translation System between Pashto and English in Two Months. Alexandra Birch, Barry Haddow, Antonio Valerio Miceli-Barone, Jindřich Helcl, Jonas Waldendorf, Felipe Sánchez-Martínez, Mikel L Forcada, Víctor Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Wilker Aziz, Lina Murady, Sevi Sariisik, Peggy van der Kreeft, Kay Macquarrie. MTSummit 2021, link

14. Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking. N Moghe, M Steedman, A Birch. EMNLP 2021, link

15. On Sparsifying Encoder Outputs in Sequence-to-Sequence Models. Biao Zhang, Ivan Titov, Rico Sennrich. ACL 2021 link

16. Analyzing the Source and Target Contributions to Predictions in Neural Machine Translation. Elena Voita, Rico Sennrich, Ivan Titov. ACL 2021 link

17. Language Modeling, Lexical Translation, Reordering: The Training Process of NMT through the Lens of Classical SMT. Elena Voita, Rico Sennrich, Ivan Titov. EMNLP 2021 link

18. Distributionally Robust Recurrent Decoders with Random Network Distillation. Antonio Valerio Miceli-Barone, Alexandra Birch, Rico Sennrich. RepNLP 2022 link

19. Revisiting End-to-End Speech-to-Text Translation From Scratch. Biao Zhang, Barry Haddow, Rico Sennrich. ICML 2022 link

20. Survey of Low-Resource Machine Translation. Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, Alexandra Birch. Computational Linguistics 2022 link

21. GoURMET – Machine Translation for Low-Resourced Languages. Peggy van der Kreeft, Sevi Sariisik, Wilker Aziz, Alexandra Birch, Felipe Sánchez-Martínez. EAMT 2022 link

3.6 Awards

The paper "**Towards a Multi-view Language Representation: A Shared Space of Discrete and Continuous Language Features**" (Link) from Arturo Oncevay, Barry Haddow and Alexandra Birch won as Best Paper at TyP-NLP 2019.

Description: The First Typology for Polyglot NLP was held at ACL 2019, aiming to promote the research development in linguistic typology for multilingual NLP tasks, such as machine translation. In our awarded submission, we combine language-level representations from typological knowledge bases, and task-driven learned embeddings of languages from a multilingual machine translation model. Our approach demonstrated that it is possible to fuse both kinds of representations with minimal information loss.

The paper "**Efficient Marginalization of Discrete and Structured Latent Variables via Sparsity**" (Link) from Gonçalo M. Correia, Vlad Niculae, Wilker Aziz, and André F. T. Martins was a Spotlight Paper at NeurIPS 2020.

Description: Training neural network models with discrete (categorical or structured) latent variables can be computationally challenging, due to the need for marginalization over large or combinatorial sets. To circumvent this issue, one typically resorts to sampling-based approximations

of the true marginal, requiring noisy gradient estimators (e.g., score function estimator) or continuous relaxations with lower-variance reparameterized gradients (e.g., Gumbel-Softmax). In this paper, we propose a new training strategy which replaces these estimators by an exact yet efficient marginalization.

The paper "**Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation**" (Link) from Bryan Eikema and Wilker Aziz won as Best Paper at Coling 2020.

Description: Recent studies have revealed a number of pathologies of neural machine translation (NMT) systems. Hypotheses explaining these mostly suggest there is something fundamentally wrong with NMT as a model or its training algorithm, maximum likelihood estimation (MLE). Most of this evidence was gathered using maximum a posteriori (MAP) decoding, a decision rule aimed at identifying the highest-scoring translation, i.e. the mode. We argue that the evidence corroborates the inadequacy of MAP decoding more than casts doubt on the model and its training algorithm. In this work, we show that translation distributions do reproduce various statistics of the data well, but that beam search strays from such statistics.

4 Exploitation Report

This section of the deliverable relates to Task 6.2 (Exploitation), focusing on capturing the information required and exploring options to ensure the outputs of the GoURMET project can be exploited by the partners themselves and others. As the exploitation road map, this document sets out the activities required for successful exploitation of GoURMET and reports on progress so far.

- Exploitation Committee - The role of the Exploitation Committee is to set and execute the exploitation strategy of the consortium as well as to ensure that IPR Management is being carried out appropriately.
- IPR Management - The Exploitation Committee has mechanisms in place to manage the record of IPR being contributed to the project by the consortium members.

Building on this foundation there are two main ways in which GoURMET can be exploited:

- Model-Based Exploitation (for other applications) The models that GoURMET has produced are available as open-source to be included in other platforms and are offered to power other services.
- Platform Exploitation (GoURMET as a whole) During the project the partners will investigate the options further into the revenue potential of the best options.

4.1 Exploitation Committee

The project established an Exploitation Committee to coordinate the management of IPR and to set and execute the exploitation strategy of the consortium. The Exploitation Committee consisted of one representative of each project partner:

- Alexandra Birch, University Edinburgh
- Kay Macquarrie, Deutsche Welle
- Wilker Aziz, University Amsterdam
- Mikel L. Forcada, University Alicante
- Sevi Sariisik Tokalac (preceded by Andrew Secker) BBC

4.2 IPR Management

As part of GoURMET's efforts to increase resources and tools available for low-resource machine translation, the project has released the corpora, models and software created during the project under the Open source license. These corpora are also available at OPUS (http://opus.nlpl.eu/GoURMET.php) and Translation Models (including a "how to use them") as well as Software developed within the scope of the GoURMET project are available on GitHub (https://github.com/EdinburghNLP/gourmet-models).

4.3 Model-Based Exploitation

As part of GoURMET's efforts to increase resources and tools available for low-resource machine translation, several sets of corpora and software created during the project have been released. All public resources are available on the project website (https://gourmet-project.eu/data-model-releases/). We list here for your convenience all the translation models shown in Table 6 "Translation Models" (also maintained here https://github.com/EdinburghNLP/gourmet-models/), all the data sets in Table 7 (further details can be found in Deliverables D1.3 and D1.4.), Table 8 "Evaluation Tools" and Table 9 "Software created or improved during GoURMET". And in Table 10 we list "BBC and DW Prototypes and Products".

Translation Models

NAME	VER.	REPOSITORY
Bulgarian to English	0.1	http://data.statmt.org/gourmet/models/docker/bg-en.20190801.tgz
English to Bulgarian	0.2	http://data.statmt.org/gourmet/models/docker/en-bg.v0.2.tgz
Gujarati to English	0.1	http://data.statmt.org/gourmet/models/docker/gu-en.20190628.tgz
English to Gujarati	0.2	http://data.statmt.org/gourmet/models/docker/en-gu.v0.2.tgz
Swahili to English	0.5	https://data.statmt.org/gourmet/models/docker/translation-sw-en-0-5-0.docker.gz
English to Swahili	0.5	https://data.statmt.org/gourmet/models/docker/translation-en-sw-0-5-0.docker.gz
Tamil to English	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-ta-en.tar.gz
English to Tamil	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-en-ta.tar.gz
Serbian to English	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-sr-en.tar.gz
English to Serbian (Cyrilic)	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-en-sr.cyr.tar.gz
English to Serbian (Latin)	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-en-sr.lat.tar.gz
Hausa to English	0.2	http://data.statmt.org/gourmet/models/docker/mt-engine-ha-en.v0.2.tar.gz
English to Hausa	0.2	http://data.statmt.org/gourmet/models/docker/mt-engine-en-ha.v0.2.tar.gz
Igbo to English	0.2	http://data.statmt.org/gourmet/models/docker/mt-engine-ig-en.v0.2.tar.gz
English to Igbo	0.2	http://data.statmt.org/gourmet/models/docker/mt-engine-en-ig.v0.2.tar.gz
Tigrinya to English	0.1	https://data.statmt.org/gourmet/models/docker/mt-engine-ti-en.tgz

English to Tigrinya	0.1	https://data.statmt.org/gourmet/models/docker/mt-engine-en-ti.tgz
Pashto to English	0.4.1	https://data.statmt.org/gourmet/models/docker/translation-ps-en-0-4-1.docker.gz
English to Pashto	0.4.1	https://data.statmt.org/gourmet/models/docker/translation-en-ps-0-4-1.docker.gz
Turkish to English	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-tr-en.tgz
English to Turkish	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-en-tr.tgz
Turkish to English	0.2	http://data.statmt.org/gourmet/models/docker/mt-engine-v2-tr-en.tgz
English to Turkish	0.2	http://data.statmt.org/gourmet/models/docker/mt-engine-v2-en-tr.tgz
Amharic to English	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-am-en.tgz
English to Amharic	0.1	http://data.statmt.org/gourmet/models/docker/mt-engine-en-am.tgz
Kyrgyz to English	0.1.1	https://data.statmt.org/gourmet/models/docker/translation-ky-en-0-1-1.docker.gz
English to Kyrgyz	0.1.1	https://data.statmt.org/gourmet/models/docker/translation-en-ky-0-1-1.docker.gz
Macedonian to English	0.1.1	https://data.statmt.org/gourmet/models/docker/translation-mk-en-0-1-1.docker.gz
English to Macedonian	0.1.2	https://data.statmt.org/gourmet/models/docker/translation-en-mk-0-1-2.docker.gz
Urdu to English	0.2	http://data.statmt.org/gourmet/models/docker/mt-engine-ur-en.v0.2.tar.gz
English to Urdu	0.2	http://data.statmt.org/gourmet/models/docker/mt-engine-en-ur.v0.2.tar.gz
Myanmar to English	0.1 Slower	http://data.statmt.org/gourmet/models/docker/translation-my-en-slower-0-1-0.docker.gz
English to Myanmar	0.1 Slower	http://data.statmt.org/gourmet/models/docker/translation-en-my-slower-0-1-0.docker.gz
Myanmar to English	0.1 Faster	http://data.statmt.org/gourmet/models/docker/translation-my-en-faster-0-1-0.docker.gz
English to Myanmar	0.2.3 Faster	https://data.statmt.org/gourmet/models/docker/translation-en-my-faster-0-2-3.docker.gz
Yoruba to English	0.1	https://data.statmt.org/gourmet/models/docker/mt-engine-yo-en.tgz
English to Yoruba	0.1	https://data.statmt.org/gourmet/models/docker/mt-engine-en-yo.tgz

Table 6: Translation models

Data Sets

DETAILS	REPOSITORY
Parallel and monolingual corpora of languages of India (PM India)	http://data.statmt.org/pmindia/
Swahili–English parallel, and Swahili monolingual	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-sw.zip
Serbian–English parallel data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.sr-en.zip
English–(Serbo or Croatian or Bosnian) parallel data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.hbs-en.zip
Hausa–English parallel corpus and Hausa monolingual corpus	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-ha.zip
Training data for WMT 2021 crawled from the President of Iran's website	https://data.statmt.org/wmt21/translation-task/ha-en/khamenei.v1.ha-en.tsv
Igbo–English parallel data and Igbo monolingual data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-ig.zip
Tigrinya–English monolingual data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.ti.zip
Pashto–English parallel data and Pastho monolingual data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-ps.zip
Turkish–English parallel data, and Turkish monolingual	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-tr.zip
Amharic–English parallel data, and Amharic monolingual data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-am.zip
Kyrgyz–English parallel data, and Kyrgyz monolingual data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-ky.zip
Kyrgyz–Russian parallel data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.ky-ru.zip
English–Macedonian parallel data and Macedonian monolingual data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-mk.zip
English–Burmese/Mayanmar parallel corpus and Burmese monolingual corpus	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-my.zip
English–Yoruba parallel data and Yoruba monolingual data	http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-yo.zip
Monolingual News Crawl in 42 languages	http://data.statmt.org/news-crawl

Development data for WMT 2021	https://data.statmt.org/wmt21/translation-task/dev.tgz
Test data for WMT 2021	https://data.statmt.org/wmt21/translation-task/test.tgz

Table 7: Data Sets

Evaluation Tools

NAME	REPOSITORY	DESCRIPTION
Direct Assessment – Sentence Pairs Evaluation Tool	https://github.com/bbc/gourmet-sentence-pairs-evaluation	The goal of Direct Assessment is to evaluate a translation model by asking a human to compare the quality of a machine translated sentence to a human translated sentence where the human translation is assumed to be the gold standard.
Gap Fill Evaluation Tool	https://github.com/bbc/gourmet-gap-fill-evaluation	The goal of Gap Filling Task is to evaluate a translation model by asking a human to fill in the gaps in a sentence that has been translated by a human using the machine translation of the same sentence as a guide to what words should go in that sentence.
GoURMET Translation API	https://github.com/bbc/gourmet-translation-api-server	The GoURMET Translation API aims to provide a single platform where all translation models produced as part of the GoURMET project can be hosted, run and accessed rather than forcing the user of the translation technology to need to integrate the translation models into the project directly. This allows a single point for maintenance and upgrades to the translation models

Table 8: Evaluation Tools (also see D5.3)

Software created or improved during GoURMET

WP	SOFTWARE	URL
WP 1	LinguaCrawl - crawl a number of top-level domains	https://github.com/transducens/linguacrawl
	LASER train - reproduces Artetxe etal. (2018, 2019) to train sentence embeddings	https://github.com/transducens/LASERtrain

	MTL-DA Training scripts using different data augmentation techniques	https://github.com/vitaka/mtl-da
	Diversity in back-translation	github.com/laurieburchell/exploring-diversity-bt
	Bitextor - most widely-used tool to automatically harvest bilingual corpora	https://github.com/bitextor/bitextor
	Bicleaner - detection of noisy sentence pairs in parallel corpora	https://github.com/bitextor/bicleaner
WP 2	Morphological segmentation using Apertium (Ataman et al., 2019)	https://github.com/transducens/smart-segmentation
	Code accomanying (Ataman et al., 2020)	https://github.com/d-ataman/Char-NMT
	Morphological Tagging and Lemmatization in Context	https://github.com/d-ataman/lmm
	Morphological Tagging and Lemmatization in Context	https://github.com/IvoOVerhoeven/morph_tag_lemmatize
WP 3	Alignment models	https://github.com/Roxot/m-to-n-alignments
	Deep latent language models	https://github.com/tom-pelsmaeker/deep-generative-lm
	Sparse approximations to binary variables	https://github.com/bastings/interpretable_predictions
	Language models with latent syntax	https://github.com/daandouwe/thesis
	Deep latent translation models	https://github.com/Roxot/AEVNMT.pt
	Contrastive test sets for document-level machine translation	https://github.com/rbawden/Large-contrastive-pronoun-testset-EN-FR
	Training data for document-level machine translation	https://github.com/radidd/Doc-substructure-NMT
	Bayesian data analysis of NMT models	https://github.com/probabll/bda-nmt
	Constrained optimisation for torch	https://github.com/EelcovdW/pytorch-constrained-opt.git
	Probabilistic modules for torch	https://github.com/probabll/dgm.pt
	Probability distributions for torch	https://github.com/probabll/dists.pt
	Alignment models	https://github.com/Roxot/m-to-n-alignments

	Deep latent language models	https://github.com/tom-pelsmaeker/deep-generative-lm
	Sparse approximations to binary variables Language models with latent syntax	https://github.com/bastings/interpretable_predictions https://github.com/daandouwe/thesis
	Deep latent translation models	https://github.com/Roxot/AEVNMT.pt
	Contrastive test sets for document-level machine translation	https://github.com/rbawden/Large-contrastive-pronoun-testset-EN-FR
	Training data for document-level machine translation	https://github.com/radidd/Doc-substructure-NMT
	Bayesian data analysis of NMT models	https://github.com/probabll/bda-nmt
	Constrained optimisation for torch	https://github.com/EelcovdW/pytorch-constrained-opt.git
	Probabilistic modules for torch	https://github.com/probabll/dgm.pt
	Probability distributions for torch	https://github.com/probabll/dists.pt
	Minimum Bayes risk decoding for NMT	https://github.com/Roxot/mbr-nmt
WP 4	WMT19 Gujarati system models and scripts	http://data.statmt.org/wmt19systems
	Tool for fusing, extending and using representations	github.com/aoncevay/multiview-langrep
	Code for the improving massively multilingual NMT work	https://github.com/bzhangGo/zero
	Code for the language model prior work	github.com/cbaziotis/lm-prior-for-nmt
	Code for the auto-encoding variational NMT work	github.com/Roxot/AEVNMT
	Exploitation of large pre-trained models	https://github.com/transducens/tune-n-distill
	Cross-Lingual Intermediate Fine-Tuning for Dialogue State Tracking	https://github.com/nikitacs16/xliftdst
WP 5	GoURMET translation platform	https://github.com/bbc/gourmet-translation-api-server
WP 6	Direct Assessment – Sentence Pairs Evaluation Tool	https://github.com/bbc/gourmet-sentence-pairs-evaluation
	Gap Fill Evaluation Tool	https://github.com/bbc/gourmet-gap-fill-evaluation

Table 9: Software created or improved during GoURMET

4.4 Sustainable Platform Exploitation

There are several scenarios for GoURMET technology as being exploited and used within an integrated product (cf. Table 9).

Number	PROTOTYPES / PRODUCTS	PARTNER	DESCRIPTION
1	Frank	BBC	Journalist gets an overview of content available from a machine translated text and can create adaptions. *Global content creation*
2	Live Page Translation	BBC	A dashboard provides an editorial overview of "live" stories produced within BBC. *Multilingual real time monitoring*
3	Graphical Storytelling Tool	BBC	A novel approach (experimental stadium) using machine translation as facilitator to enable under-resourced languages to benefit from machine learning solutions. *Global content creation*
4	plain X	DW	GoURMET models were integrated to an AI-supported editorial system creating subtitles, translations and voice-overs. *Global content creation*
5	SELMA open source NLP Tool	DW	GoURMET models were added to the SELMA open-source tool. SELMA is an ongoing broad language technology project, targeting low-resource as well as high-resource languages. *Global content creation*
6	Benchmarking prototype	DW	The prototype evaluates various NLP processes and models and compares the output. *Serves both content creation and media monitoring*

Table 10: BBC and DW Prototypes and Products

4.4.1 BBC Integrated Tools

During months M19-42 of the project, BBC's focus has been on identifying areas of opportunity for MT-assisted solutions, designing and developing interfaces that would expose journalists to MT-assisted workflows, gathering feedback on the models developed and building stakeholder relationships to facilitate the exploitation of the outcomes of the research both within and outside the BBC.

BBC has enabled exploitation of the models for all three key use cases as detailed in the deliverable report 5.5 sections 1.2 and 9.3. The Live Page Translation tool and Frank both serve to achieve editorial compliance and oversight goals by transposing all content to a level playing field, viewable by all. The benefits of real-time monitoring and content creation offered by Live Pages Translation tool were clearly proven during the Russia-Ukraine War, with core English-language offers making use of the Russian and Ukrainian output.

In contrast to the "breaking/developing" nature of the content for LPT, the second prototype has focused on the "slow news" genre, as per the feedback from WS journalists. This was identified

as the right space to allow for post-edits and checks required for content creation from a machine translated text.

Away from the GoURMET project, development work is underway (to be completed by the end of August 2022) to link the Frank prototype and the models deployed with the content enrichment pipeline which derives content from a wider range of BBC content management tools which are coming online, to future-proof the MT workflows generated.

Meanwhile, Graphical Storytelling Tool (GST) has experimented with a novel approach where ”translation” was not the intended output but rather the facilitator to enable under-resourced languages to benefit from the kind of machine learning solutions available only for the English language. The above-mentioned ”Suite of Multilingual Prototypes” News Labs has developed under the GoURMET project and described in further detail in D5.5 has been shortlisted for the BBC News Awards 2021 in Outstanding Digital Innovation category and has provided internal dissemination opportunities.

BBC News Labs has been working with the European Broadcasting Union (EBU) on EuroVox project (https://tech.ebu.ch/eurovox), a shared initiative of several European broadcasters, including project partner DW. Discussions with EBU stakeholders to ensure exploitation of some European languages of interest (i.e. Bulgarian, Macedonian, Serbian, Turkish) are ongoing. Making GoURMET models accessible to EBU members representing 112 organisations in 56 countries would be a significant outcome.

4.4.2 DW Integrated Tools

The results of the GoURMET achievements are being implemented and will continue to be used in Deutsche Welle in different ways.

First of all, it is deployed in an AI-supported editorial system (called plain X - https://www.plain-x.com) which enables journalists to translate, subtitle (for social media and for accessibility) and add voice-over in virtually any language broadcast at DW. This means that the GoURMET models are used to produce text translations, video subtitles, and voice-over output. This system is currently being rolled in Deutsche Welle for the entire editorial workforce, as well as other departments and expected to be completed by the end of 2022.

Publications, whether from text, audio or video, can now easily be reproduced and published in other languages using a variety of translation models, including GoURMET. The platform is also being made available to other (primarily media) clients including public broadcasters e.g. within the ARD group as well private media companies.

Secondly, we have implemented the models in our SELMA prototype (https://selma-project.eu/), which uses natural language processing (NLP) for multilingual media monitoring in over 30 languages. All GoURMET models (in both directions) have been integrated into the OSS (Open-Source System) of the SELMA platform in particular, using the dockerised engines. This ensures the GoURMET models are available to anyone interested in using the platform. Continued dissemination is done within the scope of SELMA project.

Thirdly, the GoURMET models are part of DW’s benchmarking prototype, a system that Deutsche Welle has set up to evaluate different NLP processes and models and compare the output. This includes automated speech recognition (ASR), machine translation (MT), and synthetic voice. We have set up a sustainable internal benchmarking system, which compares the output of different

engines and technologies, looks at various aspects of ASR, MT and synthetic voice. It is and will be further automated as much as possible, so that updated or new engines can efficiently be re-evaluated with minimum effort.

Benchmarking is an ongoing process, otherwise the results soon become obsolete. As NLP processes are interrelated in Deutsche Welle, as explained above for the plain X application, we stack ASR, MT and voice-over, thus qualitative outcome of the different steps - and evaluation thereof - is essential. A combination of evaluation methods is used:

automated evaluation (BLEU for MT and WER for ASR), in addition to human evaluation with findings collected through a user questionnaire. The process is automated in such a way that it is quite easy to rerun the evaluation in case of updated models or new engines to be explored.

4.5 Accumulated Knowledge (per project partner)

4.5.1 University of Edinburgh

As a result of GoURMET, UEDIN has further developed its expertise and know-how in data creation and low-resource machine translation, and the development of new models, algorithms, and evaluation related to the GoURMET use cases. We have better understood the advantages of multi-lingual pre-training and fine-tuning paradigm that has led to large advances in the field of low-resource MT.

4.5.2 University Alicante (UA)

Universitat d'Alacant (UA) has been researching on low-resource machine translation for more than two decades. We started with rule-base machine translation and then moved to statistical machine translation, and finally to neural machine translation. Thanks to the GoURMET project we have deepened our knowledge of neural networks, how they behave in low-resource conditions, how to integrate existing bilingual resources and how to make the most of the little parallel corpora available. We have also learned to better crawl and filter parallel corpora for low-resource languages and how to exploit existing pre-trained models. In the near future, we plan to use this knowledge to develop neural machine translation systems between Spanish and the low-resource languages used by migrants and asylum seekers in Spain. We also plan to improve some of the methods we have developed during the project, in particular the crawling of corpora, the generation of synthetic samples and the exploitation of existing pre-trained models.

4.5.3 University Amsterdam

UVA has been contributing to computational linguistics and natural language processing since the early 90s. Prior to neural machine translation, the UVA contributed to statistical machine translation, in particular, developing latent variable models of reordering and translation, establishing a tradition in combining symbolic and statistical approaches. Through GoURMET we have expanded our research portfolio to include neural machine translation and deep learning for NLP more generally. This also had an impact on our teaching and supervision portfolio, for example, through new or redesigned courses at both BSc and MSc levels. GoURMET also brought us closer to European partners (in industry and academia) leading to new funded collaborations. In the near future, we will keep developing our agenda of latent variable approaches to natural language

processing and translation in particular, always seeking to facilitate the use of assumptions about unobserved aspects of data generating processes as a means to advance NLP for low resource languages and domains.

4.5.4 British Broadcasting Company (BBC)

Broadcasting in more than 40 languages, facilitating content exchanges between its language services through technology is a vital aspiration for the BBC, in particular its World Service arm comprising multi-platform language services and BBC Monitoring. News Labs, BBC's News Innovation Hub has been involved in multinational and multilingual projects to this effect since its inception 10 years ago. However, GoURMET represented the first comprehensive attempt focusing on custom translation models for low resourced languages. The immediate team has gained comprehensive insights about the MT landscape in general and challenges in deployment (i.e. reconciling speed, scalability and quality requirements) in particular, as well as a much deeper understanding about editorial needs and resistance points. The project has also given most of the World Service teams their first exposure of machine translation-assisted solutions, and enabled compilation of extensive feedback. We also learned about the various evaluation and benchmarking tools and approaches; growing into advocates for the establishment of a systemic benchmarking hub for the Corporation. Working on the health domain task meant gaining first hand experience about what to look out for building terminology lists. One key learning from the project was the need for a content enrichment hub that could modulate between several different input and output interfaces. A pipeline built for this purpose is coming online imminently and will offer the basis of future translation tooling and experiments. The team will continue its work to turn its prototypes into production scale tools to allow better editorial oversight and exchanges at BBC and champion efforts for more strategic investment into this field.

4.5.5 Deutsche Welle (DW)

Deutsche Welle (DW) as an international broadcaster in 32 languages - most of them being broadcast in the low-resource domain - has a high interest of using and deepening its knowledge in machine translation. The direct need to take advantage of technological progress in the field of automatic language translation has already led to several successful engagements in European projects in the past decade. The GoURMET project has contributed in many ways and especially in the domain of low-resourced languages. We learned how to use low-resource machine learning models and how models are integrated into existing HLT (Human Language Technology) environments at DW. We learned how to validate language pairs using current assessment methodologies (including direct assessment and gap filling) and we learned how to set up a "surprise language challenge" including providing dataset and validating it in a short time period. We learned how low-resource models can be developed. We will continue to add more language models and pairs and improve the services.

5 Conclusion

The GoURMET project was successful in developing and promoting machine translation technology for media and multilingual newsrooms during the course of the project (1.1.2019 - 30.6.2022). The project has developed a significant amount of code - in form of data, models and software releases - which is mostly made accessible Open Source and thus can be publicly used (via the dedicated GoURMET webpage under https://gourmet-project.eu/data-model-releases/ or for more immediate access via the corresponding GitHub account: https://github.com/EdinburghNLP/gourmet-models).

More than 40 research publications were produced and disseminated in more than 60 academic and industry events. Two user events organized by the GoURMET project and led by DW provided platforms for valuable interactions and exchanges with key players in the language and MT field.

Based on the dissemination strategy, GoURMET has exploited various communication means and continuously publicized news, milestones, activities and achievements. The platforms included online media, primarily the project website; social media platforms like Twitter and LinkedIn, as well as direct interactions involving face-to-face and virtual meetings and conferences. The latter were well-attended by both the research and media partners. The exploitation goals were mainly driven by BBC and DW use cases, while also enabling and encouraging utilisation of the GoURMET technology for the wider community. Talks were underway with EBU to this end at the time of the project's completion.

GoURMET technology and models have been integrated into human language tools and / or prototypes at both public media organisations: For BBC this is valid for three prototypes (Frank, LPT and GST) available to both BBC News and World Service teams; for Deutsche Welle this is plain X (https://plain-x.com/), the NLP benchmarking system and the Open Source NLP SELMA platform (https://selma-project.github.io/).

Along with the GitHub account where the code is available for further refinement by the developer community, the GoURMET project offers the GoURMET Translate tool publicly for anyone interested in translations in and out of the low-resourced languages paired with English (https://gourmet-project.eu/project-output/gourmet-translate-tool/). After the project, access to the translation models will be made available through the SELMA open-source platform with all GoURMET dockerized engines being integrated.

6 Appendix

6.1 Ideation session during the Open Workshop

As part of the Open Workshop in May 2021, BBC News Labs organised a 90-minute ideation session. This has been an insightful format with an outcome which also directs into potentially interesting future work. We provide the set-up and the outcome in detail below:

Unlike a traditional full-on 5-day design sprint, participants from all walks of media, academia and wider multilingual industries only had two hours to ideate and come up with a fleshed out idea. Bearing in mind the time constraints, we have facilitated the session with the following approach:

First, we have sent out an exercise upfront for participants to think about the challenges in their own areas, to draw questions about the problem that wanted to address in a "How might we..." phrase. The purpose of the "How Might We (HMW)?" exercise is to reach a broad understanding of the challenges that media organisations are facing and to explore the opportunities of machine translation in tackling those challenges. Below are some of the samples of the How Might We hypothesis proposed by the participants.

Based on the long list, BBC team have shortlisted 6 HMWs for the participants to work on the day.

- How might we use machine translation to help audiences get different perspectives?
- How might we present audience content that is not created in their language but is of interest to them?
- How might we ensure machine translation handles local context / relevance / names / organisational style?
- How might we best harness post edits to help improve future translations
- How might we use machine translation to identify stories / themes emerging in languages on issues of interest?
- How might we ensure visibility of content irrespective of language it is in?

These were then voted, grouped and assigned to teams drawn from a list of participants who pre-registered to attend the event with a view to have a balanced mix of skills aiming i.e ensuring no two people from the same organisation or skills set were in the same group.

They typically included:

- An academic researcher from the Consortium
- A product manager
- A software engineer
- A journalist
- A UX designer
- A linguist

Group No.	Challenge	Outcomes
1	How might we use machine translation to help audiences get different perspectives?	Idea: Comparitise DESCRIPTION: Automatically align stories on same topic from different languages to show similarities and differences. This is related to the ability to search, display and contrast multiple viewpoints on a story from different language communities.
2	How might we present audience content that is not created in their language but is of interest to them?	Idea: TransLearn Buddy An MT solution integrated into the newsroom CMS with learning features: Journalist feeds back to translation server on accuracy; MT learns and improves
3	How might we ensure machine translation handles local context / relevance / names / organisational style?	Idea: Oracle Description: Build a knowledge base in multiple languages and cultures which evolves by learning from comparable local media and on the basis of post-edits and gets integrated into NMT systems, Includes named entities, Phrases, Expressions, Images (Vladimir Putin, Kremlin), Localised events i.e Cultural events (which can be comparable)
4	How might we best harness post edits to help improve future translations	Idea: NameTagger Description: Allow editors to tag named entities, as these are often mistranslated. Run Named Entity Recognition (NER) system before the translation. Names in new breaking stories are often mistranslated, for example Wuhan. In the legacy MT systems names and other hard-to-translate words were copied unaltered to the translation.
5	How might we use machine translation to identify stories / themes emerging in languages on issues of interest?	Idea: ViewPoints Description Find articles from around the world on a certain topic and try to cluster them into different stances or focusses and try to quantify how many articles repeat certain positions or facts vs other positions or facts and how this changes from country to country or language to language.
6	How might we ensure visibility of content irrespective of language it is in?	Idea: TERG: Temporal Entity Relation Graph Description: We convert the content's story into a graph with entities (people, locations, time). A visually appealing dashboard allows people to easily explore the content grouped around the entities and see the strength of the links between them. The representations of entities and relationships are automatically adapted to various languages or writing systems; iconic images are used as support. The temporal evolution of the story can be controlled via a slider. While moving the slider, entities appear and disappear in the graph, while their size reflect their relevance at that specific moment in time.

Table 11: Six Challenges the participants worked on during the ideation session

There were six groups to work on each of the challenges (see table).

As a result of the discussions, each team came up with a presentation at the end of the session. The idea from Group 5 narrowly won over the vote over Group 1 and the team members were presented with book tokens.

Main learnings from this exercise were:

- It was useful to innovate with the community and with people from different backgrounds. Thanks to the GoURMET project, this was made possible.
- The open workshop bridges the silos of research and industry, allowing researchers to learn about the real-world problems in the newsroom where their research can be applied. It also allows people from the industry to know the cutting-edge research and the potential of MT.
- Further prototyping and idea validation could have been very useful to make the best use of the ideas from the workshop
- Forming a community of MT research experts and industry partners to meet regularly to discuss common challenges would be very useful.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D6.4 Final Dissemination and Exploitation Report