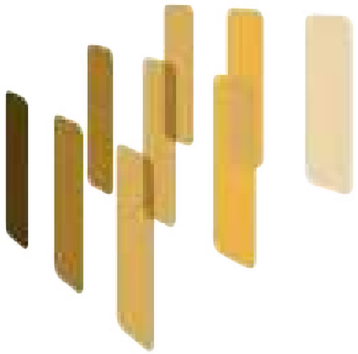
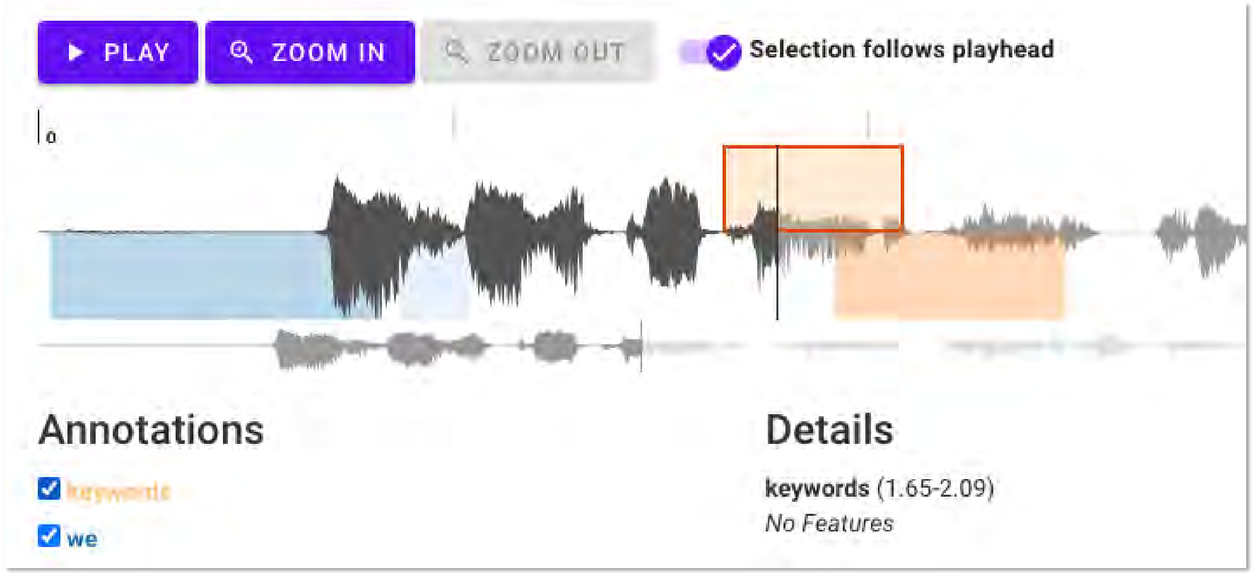


European Language Grid – META-FORUM 2021
Kalina Bontcheva – European Language Grid: LT Services and Resources

- Current State of Play
- Future Plans: Releases 3

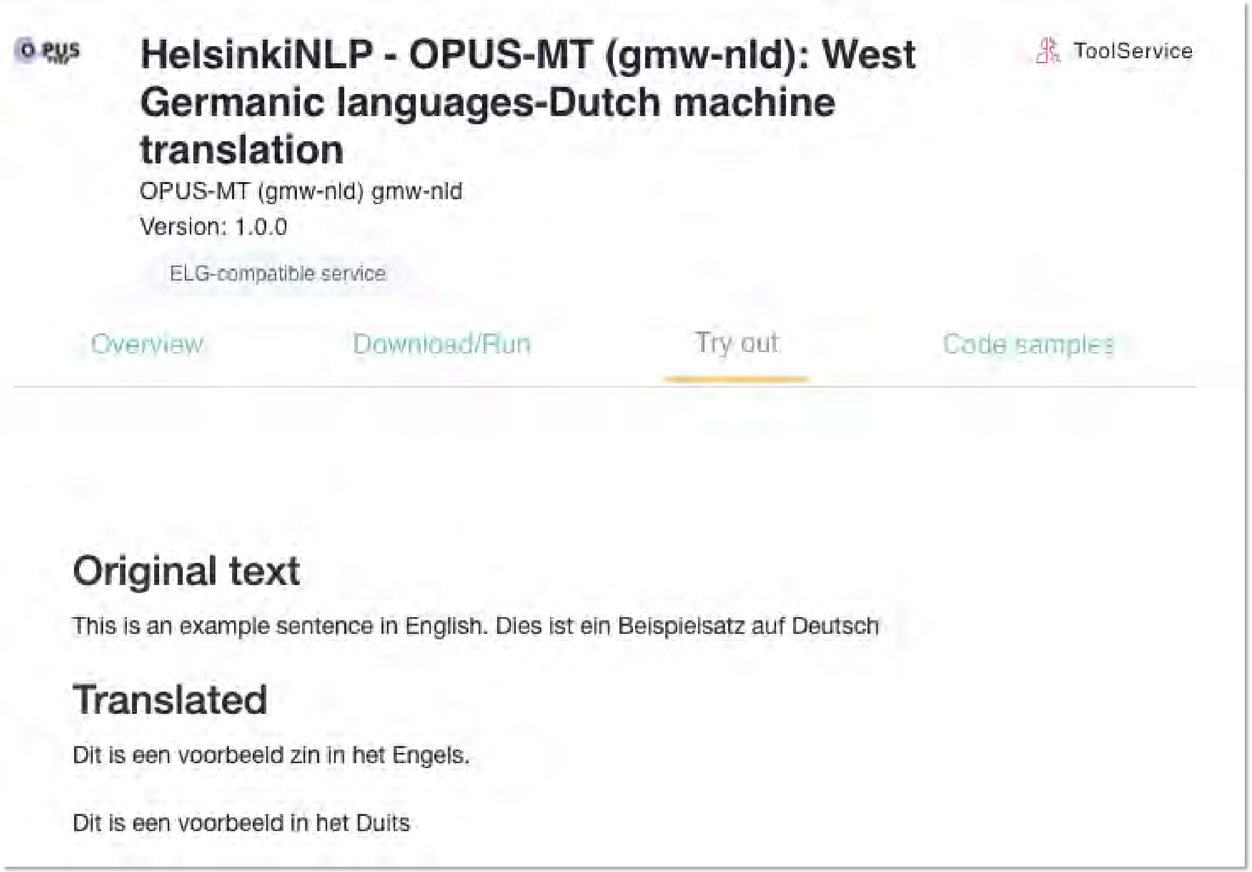
LT Services: From the consortium

- ELG Release 1 in April 2020 finalized APIs for major classes of services (ASR, IE, MT, TTS)
- Concentrated on a subset of EU languages – Czech, English, French, German, Greek, Latvian, Spanish (native languages of the ELG consortium)
  - 9 ASR services, approx. 150 distinct services for IE & Text Analytics, 24 MT, 2 TTS
- Release 2 (early 2021) added support for more EU and related languages
  - 8 further ASR services, approx. 300 IE & Text Analytics, 24 MT, 17 TTS
- … and APIs for additional service types
  - 12 services doing keyword spotting in audio data
- Release 3 (2022) will add more languages beyond the EU (13 ASR, approx. 150 IE, 9 MT), and new service types including OCR



LT Services: From the Pilots

- Many additional services have been contributed by the ELG-funded pilot projects, more to follow for R3
- Services already integrated include:
  - Over 100 MT language pairs from OPUS-MT
    - including non-EN pairs like German-Croatian
    - and multilingual models able to take input in mixed languages
  - Clinical NER in 5 languages from E3C
  - 37 services from Lingsoft (a mix of MT, ASR & IE)
  - Basque language ASR and TTS from Elhuyar
  - Terminology extraction service from Text2TCS
  - Italian classification services from EVALITA
- … more on some of these projects in the next session

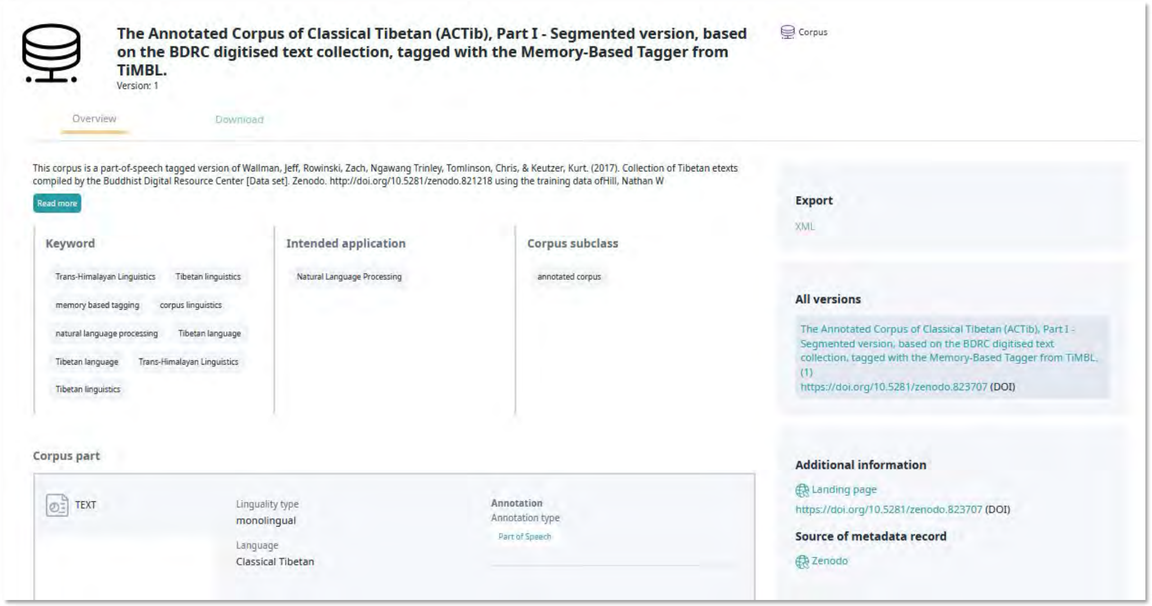


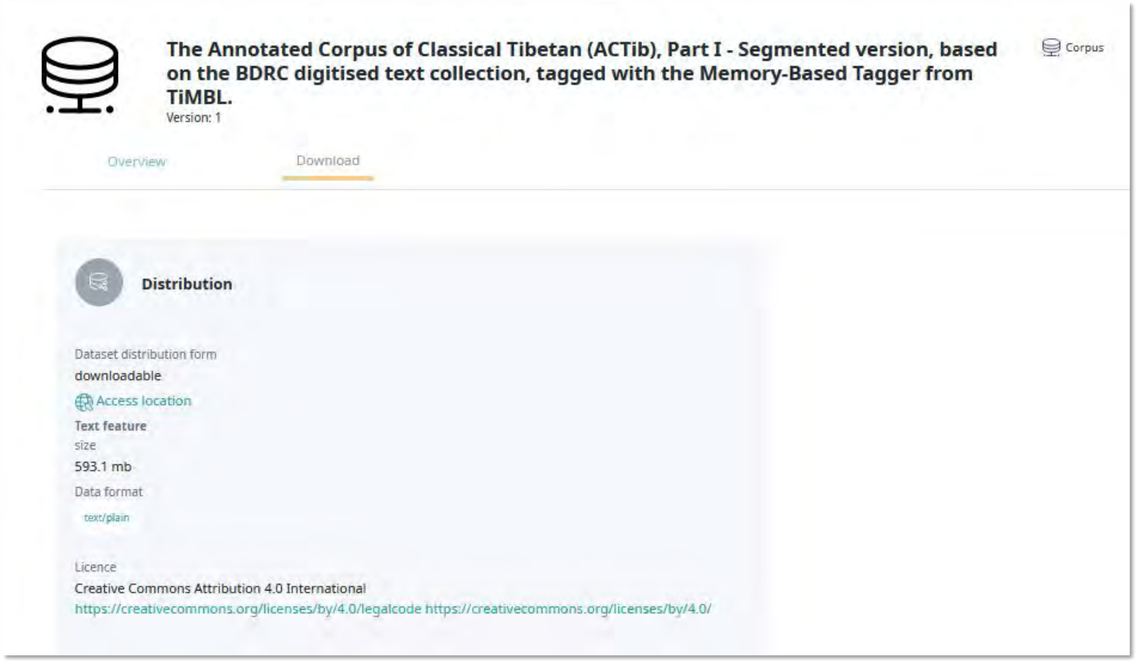
LT Services: from elsewhere

- Other research projects have contributed services including
  - Anonymisation from COMPRISE
  - Sentiment analysis from EMBEDDIA
  - Archaeological NER from ARIADNE
  - Summarisation, document classification and language ID from QURATOR
  - Credibility assessment service from PANQURA
  - Welsh language tools from Bangor University
- Final release will open up the platform to third party *commercial* providers

Language Resources: metadata description and harvesting

- ELG has imported metadata from many other repositories:
  - ELRA Catalogue, ELRC-SHARE, ELRA-SHARE-LRs, LINDAT/CLARIAH-CZ, CLARIN PL repository, CLARIN Slovenia repository, META-SHARE-DFKI, META-SHARE-ELDA and META-SHARE-ILSP, Hugging Face, Quantum Stat, Zenodo
  - Some metadata can be imported automatically, some need manual correction





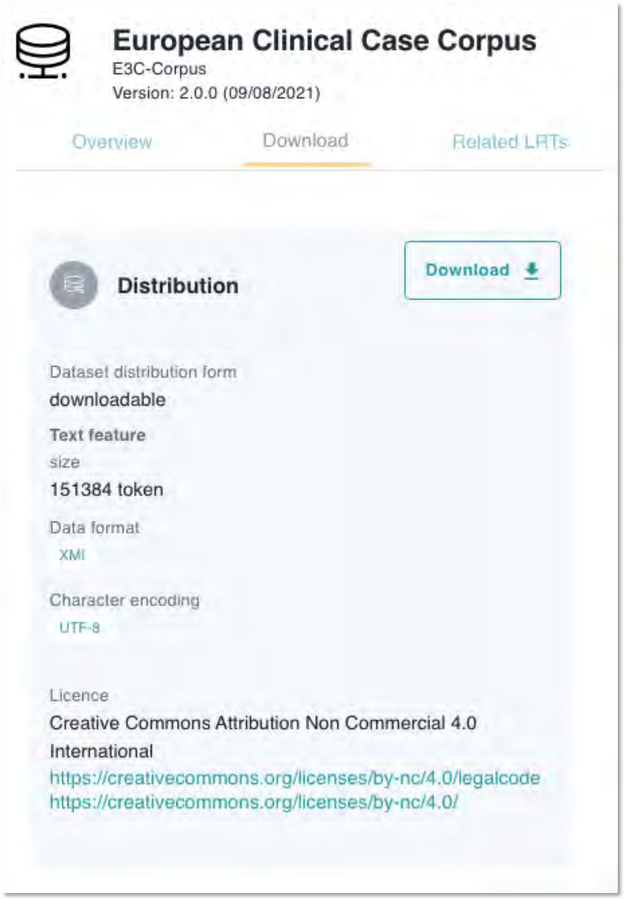
Some statistics on ingested LRs that are publicly visible

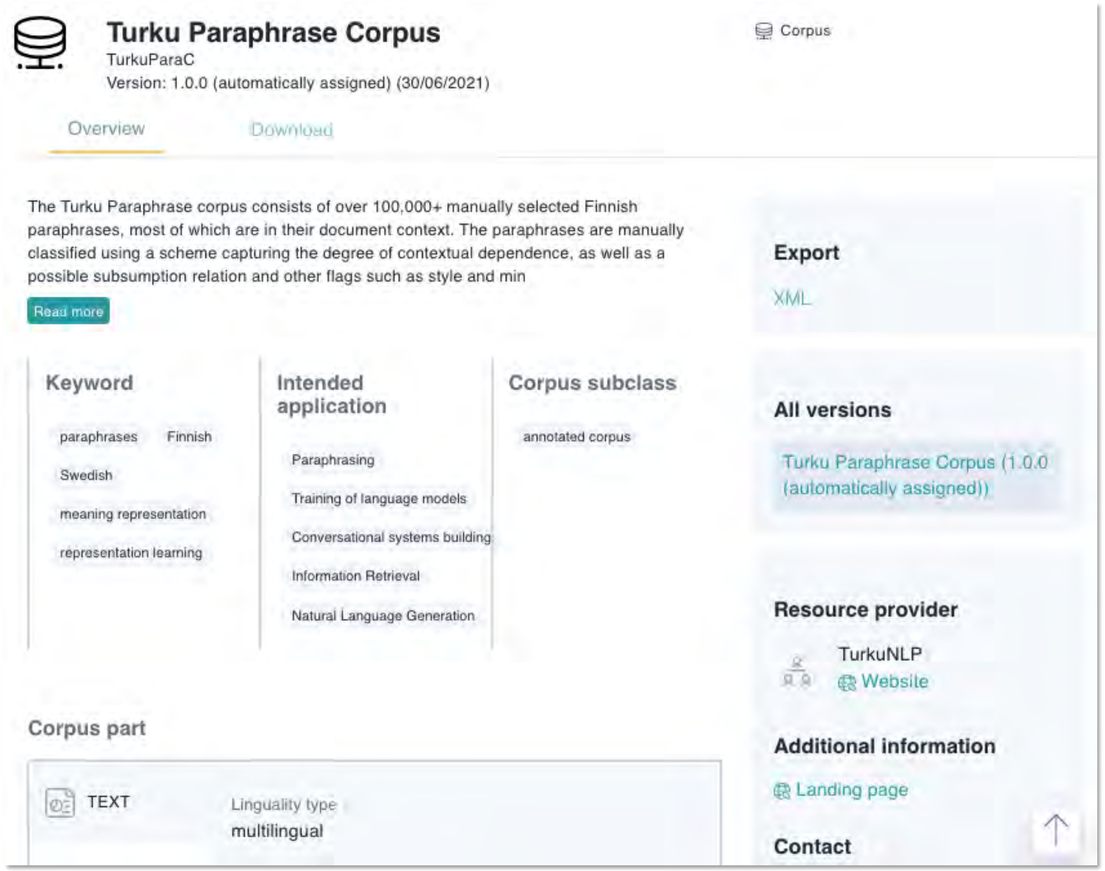
Repository	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	Total
ELRA	635	545	–	**1180**
ELRC-SHARE	1249	50		**1299**
META-SHARE	52	12	7	**71**
ELRA-SHARE-LRs	92	33	1	**126**
LINDAT/CLARIAH-CZ	274	79	–	**353**
CLARIN.SI	140	78	–	**218**
CLARIN.PL	239	15	–	**254**
Quantum Stat	255	5	–	**260**
Zenodo	137	99	24	***260**
Tudatalib	1	–	–	**1**
HuggingFace	385	–	–	**385**
Others	72	28	5	**105**
TOTAL	3531	944	37	**4512**

***approx. 400 ingested resources are still under validation**

Language Resources: hosted datasets

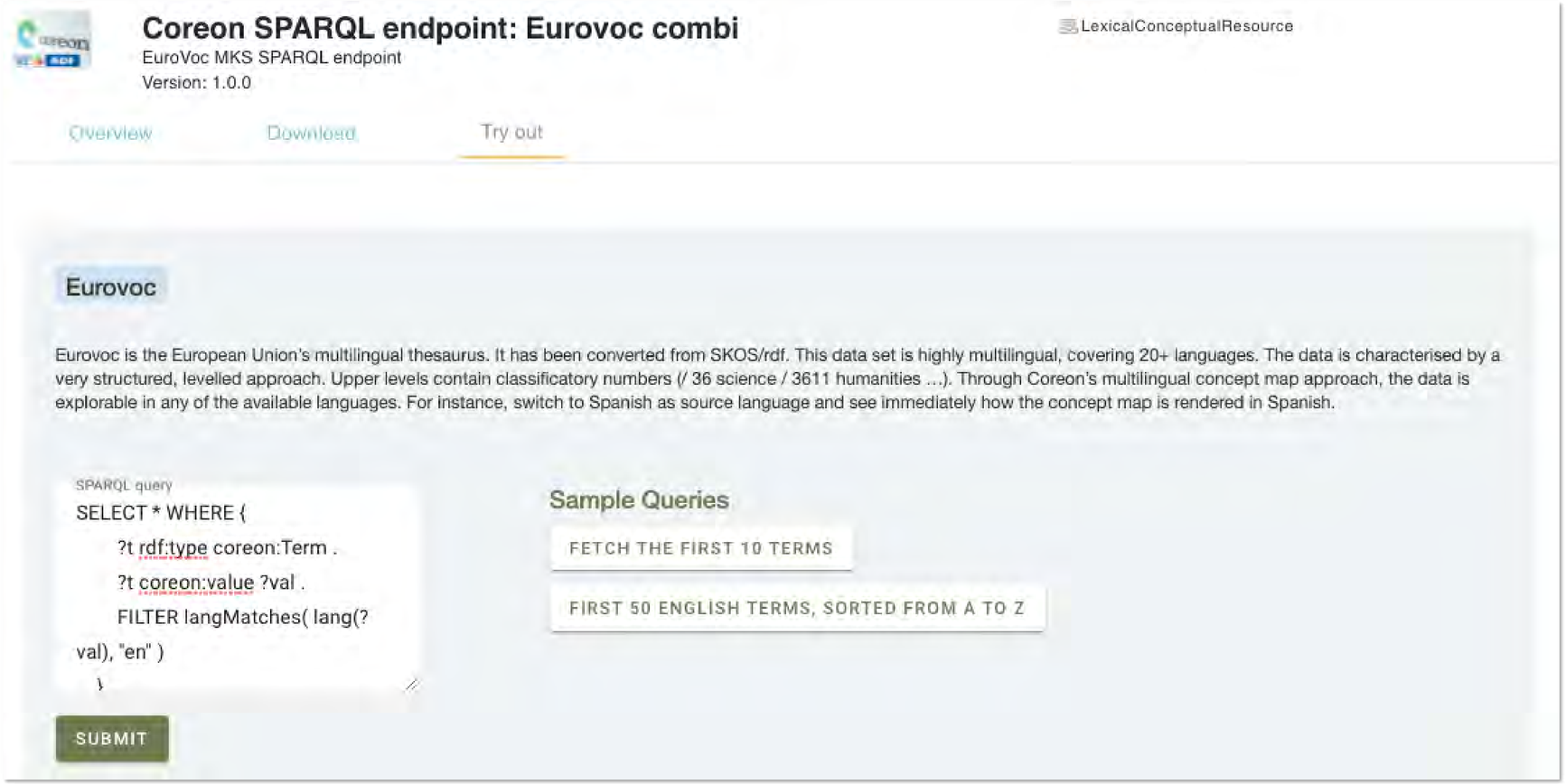
- Data resources can also be hosted at ELG for direct download, e.g., from pilot projects





Language Resources: other LR types

- Besides downloadable LRs, ELG also supports LRs accessed via a query interface
- Coreon pilot project integrated SPARQL endpoints queryable through ELG



Language Resources: current state

- Some large and important repositories (Zenodo, HuggingFace) are particularly complex, with sparse or variable quality metadata
  - Different strategies have been applied to handle this
- Critical issues being addressed include:
  - Duplication – the same resource in multiple repositories
  - Re-use – datasets that claim to be distinct but have data in common, e.g., WMT or RumourEval shared tasks
  - Updates – new datasets added to already imported repositories
- Legal issues
  - ELG is a complex platform with heterogeneous data flows that require monitoring for GDPR compliance
  - ELG services and datasets use over 130 different licences between them
    - Work in progress to characterise and homogenise licence conditions to aid search (e.g., user may only want data sets that are not share-alike)

Come and visit us at the ELG booth in the project expo!

➔ Project expo (today, 12:00-13:30)

Summary and Next Steps

- **Establish ELG as the primary platform and marketplace for Language Technology in Europe.**
- An initiative ***from*** the European LT community ***for*** the European LT community.
- European LT landscape is **highly fragmented**: ELG aims to provide just the right **umbrella platform**.
- Global market size by 2025 is enormous: we want the European LT community to be a **key player**.
- We want to **increase t[illegible]n LT landscape**.

Update slide

- ELG is a long-term initiative: we will establish a **legal entity** for sustainability, which will operate and maintain the technology platform for the whole LT community as a joint marketplace.
- Contribute to **Digital Language Equality** in Europe by giving all our languages one virtual home and umbrella platform that collects **all** services and resources (**ELE**).
- **Next steps:** validation of **ELG products**; attach more **data repositories**; include ELG in **relevant infrastructures** (e.g., NFDI, GAIA-X); **raise awareness** of LT for Europe; organise **META-FORUM 2022**; develop **ELG Release 3** (early 2022); establish the **ELG legal entity** (2022).

European Language Grid: An Overview

Georg Rehm[1], Maria Berger[1], Ela Elsholz[1], Stefanie Hegele[1], Florian Kintzel[1], Katrin Marheinecke[1], Stelios Piperidis[2], Miltos Deligiannis[2], Dimitrios Galanis[2], Katerina Gkirtzou[2], Penny Labropoulou[2], Kalina Bontcheva[3], David Jones[3], Ian Roberts[3], Jan Hajič[4], Jana Hamrlová[4], Lukáš Kačena[4], Khalid Choukri[5], Victoria Arranz[5], Andrejs Vasiļjevs[6], Orians Anvari[6], Andis Lagzdiņš[6], Jūlija Meļņika[6], Gerhard Backfried[7], Erinç Dikici[7], Miroslav Janosik[7], Katja Prinz[7], Christoph Prinz[7], Severin Stampler[7], Dorothea Thomas-Aniola[7], José Manuel Gómez Pérez[8], Andres Garcia Silva[8], Cristian Berrio[8], Ulrich Germann[9], Steve Renals[9], Ondrej Klejch[9]

[1] DFKI GmbH, Germany • [2] ILSP/Athena RC, Greece • [3] University of Sheffield, UK • [4] Charles University, Czech Republic • [5] Evaluations and Language Resources Distribution Agency (ELDA), France • [6] Tilde, Latvia • [7] SAIL LABS Technology GmbH, Austria • [8] Expert System Iberia SL, Spain • [9] University of Edinburgh, UK

Corresponding author: Georg Rehm georg.rehm@dfki.de

Abstract

Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid

Penny Labropoulou[1], Katerina Gkirtzou[1], Maria Gavriilidou[1], Miltos Deligiannis[1], Dimitrios Galanis[1], Stelios Piperidis[1], Georg Rehm[2], Maria Berger[2], Valérie Mapelli[3], Mickaël Rigault[3], Victoria Arranz[3], Khalid Choukri[3], Gerhard Backfried[4], José Manuel Gómez Pérez[5], Andres Garcia Silva[5]

[1]ILSP/Athena RC, Greece, [2]DFKI GmbH, Germany, [3]Evaluations and Language Resources Distribution Agency (ELDA), France, [4]SAIL LABS Technology GmbH, Austria, [5]Expert System Iberia SL, Spain

{penny, gkirtzou, maria, mdel, galanisd, spip}@athenarc.gr, {georg.rehm, maria.berger}@dfki.de, {mapelli, mickael, arranz, choukri}@elda.org, Gerhard.Backfried@sail-labs.com, {agarcia, jmgomez}@expertsystem.com

Abstract

The current scientific and technological landscape is characterised by the increasing availability of data resources and processing tools and services. In this setting, metadata have emerged as a key factor facilitating management, sharing and usage of such digital assets. In this paper we present ELG-SHARE, a rich metadata schema catering for the description of Language Resources and Technologies (processing and generation services and tools, models, corpora, term lists, etc.), as well as related entities (e.g., organizations, projects, supporting documents, etc.). The schema powers the European Language Grid platform that aims to be the primary hub and marketplace for industry-relevant Language Technology in Europe. ELG-SHARE has been based on various metadata schemas, vocabularies, and ontologies, as well as related recommendations and guidelines.

Keywords: metadata, language technology, language technology services, language resources

1. Introduction

The rise of data-driven approaches that use Machine Learning (ML), and especially the breakthroughs in the Deep Learning field, has put data into a central place in all scientific and technological areas, Natural Language Processing (NLP) being no exception. Datasets and NLP tools and services are made available through various repositories (institutional, disciplinary, general purpose, etc.), which makes it hard to find the appropriate resources for one's purposes. Even if they are brought together in one catalogue, such as the European Open Science Cloud[1] or the Google dataset search service[2], the difficulty of spotting the right resources and services among thousands still remains. Metadata plays an instrumental role in solving this puzzle, as it becomes the intermediary between consumers (humans and machines) and digital resources.

In addition, in the European Union, with the 24 official and many additional languages, multilingualism, cross-lingual and cross-cultural communication in Europe as well as an inclusive Digital Single Market[3] can only be enabled and firmly established through Language Technologies (LT). The boosting of the LT domain is thus of utmost importance. To this end, the European LT industry needs to be strengthened, promote its products and services, integrate them into applications, and collaborate with academia into advancing research and innovation, and bringing research outcomes to a mature level of entering the market. The European Language Grid (ELG) project[4] aims to drive forward the European LT sector by creating a platform and establishing it as the primary hub and marketplace for the LT community. The ELG is developed to be a scalable cloud platform, providing in an easy-to-integrate way, access to hundreds of commercial and non-commercial LTs for all European languages, including running tools and services as well as data resources. Discovery of and access to these resources can only be achieved through an appropriate metadata schema. We present here the ELG-SHARE schema, which is used for the description of LT-related resources shared through the ELG platform and its contribution to the project goals.

2. Objectives

The ELG project (Rehm et al., 2020a) aims to foster European LT by addressing the fragmentation that hinders its development; see indicatively (Rehm and Hegele, 2018; Rehm et al., 2016). To this end, it builds a platform dedicated to the *distribution and deployment of Language Resources and Technologies (LRT)*, aspiring to establish it as the primary platform and marketplace for industry-relevant LT in Europe. The *promotion of LT stakeholders and activities* and growth of their visibility and outreach is also one of its goals. Together with complementary material in the portal (e.g., training material, information on events, job offerings, etc.), ELG offers a comprehensive picture of the European LT sector.

The ELG platform[5] will offer access to hundreds of *commercial* and *non-commercial LTs* and ancillary *data LRs* for all European languages and more; these include processing and generation services, tools, applications for written and spoken language, corpora, lexicons, ontologies, term lists, models, etc. All resources are accessed through their descriptions in the ELG catalogue. LRT providers can describe, upload, and integrate their assets in ELG, and LRT

[1]https://www.eosc-portal.eu
[2]https://toolbox.google.com/datasetsearch
[3]https://ec.europa.eu/digital-single-market/en
[4]https://www.european-language-grid.eu

[5]The ELG platform has just been launched (alpha release) and will continue to be updated with new resources and functionalities (official release dates are on April of 2020, 2021 and 2022.

The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe

Georg Rehm[1], Katrin Marheinecke[1], Stefanie Hegele[1], Stelios Piperidis[2], Kalina Bontcheva[3], Jan Hajič[4], Khalid Choukri[5], Andrejs Vasiļjevs[6], Gerhard Backfried[7], Christoph Prinz[7], José Manuel Gómez Pérez[8], Luc Meertens[9], Paul Lukowicz[1], Josef van Genabith[1], Andrea Lösch[1], Philipp Slusallek[1], Morten Irgens[10], Patrick Gatellier[11], Joachim Köhler[12], Laure Le Bars[13], Dimitra Anastasiou[14], Albina Auksoriūtė[15], Núria Bel[16], António Branco[17], Gerhard Budin[18], Walter Daelemans[19], Koenraad De Smedt[20], Radovan Garabík[21], Maria Gavrilidou[2], Dagmar Gromann[18], Svetla Koeva[22], Simon Krek[23], Cvetana Krstev[24], Krister Lindén[25], Bernardo Magnini[26], Jan Odijk[27], Maciej Ogrodniczuk[28], Eiríkur Rögnvaldsson[29], Mike Rosner[30], Bolette Sandford Pedersen[31], Inguna Skadiņa[32], Marko Tadić[33], Dan Tufiș[34], Tamás Váradi[35], Kadri Vider[36], Andy Way[37], François Yvon[38]

[1] DFKI GmbH, Germany • [2] ILSP/Athena RC, Greece • [3] University of Sheffield, UK • [4] Charles University, Czech Republic • [5] ELDA, France • [6] Tilde, Latvia • [7] SAIL LABS Technology GmbH, Austria • [8] Expert System Iberia SL, Spain • [9] CrossLang, Belgium • [10] Oslo Metropolitan University, Norway • [11] Thales Group, France •

European Language Grid: A Joint Platform for the European Language Technology Community

Georg Rehm[1], Stelios Piperidis[2], Kalina Bontcheva[3], Jan Hajič[4], Victoria Arranz[5], Andrejs Vasiļjevs[6], Gerhard Backfried[7], José Manuel Gómez-Pérez[8], Ulrich Germann[9], Rémi Calizzano[1], Nils Feldhus[1], Stefanie Hegele[1], Florian Kintzel[1], Katrin Marheinecke[1], Julián Moreno-Schneider[1], Dimitrios Galanis[2], Penny Labropoulou[2], Miltos Deligiannis[2], Katerina Gkirtzou[2], Athanasia Kolovou[2], Dimitris Gkoumas[2], Leon Voukoutis[2], Ian Roberts[3], Jana Hamrlová[4], Dušan Variš[4], Lukáš Kačena[4], Khalid Choukri[5], Valérie Mapelli[5], Mickaël Rigault[5], Jūlija Meļņika[6], Miroslav Jánošík[7], Katja Prinz[7], Andrés García-Silva[8], Cristian Berrío[8], Ondrej Klejch[9], Steve Renals[9]

[1]DFKI GmbH, Berlin, Germany; [2]R. C. Athena, ILSP, Greece; [3]University of Sheffield, UK; [4]Charles University, Czechia; [5]ELDA, France; [6]Tilde, Latvia; [7]HENSOLDT Analytics, Austria; [8]Expert System, Spain; [9]University of Edinburgh, UK

Corresponding author: georg.rehm@dfki.de

Abstract

Europe is a multilingual society, in which dozens of languages are spoken. The only option to enable and to benefit from multilingualism is through Language Technologies (LT), i. e., Natural Language Processing and Speech Technologies. We describe the European Language Grid (ELG), which is targeted to evolve into the primary platform and marketplace for LT in Europe by providing one umbrella platform for the European LT landscape, including research and industry, enabling all stakeholders to upload, share and distribute their services, products and resources. At the end of our EU project, which will establish a legal entity in 2022, the ELG will provide access to approx. 1300 services for all European languages as well as thousands of data sets.

1 Introduction

Europe is a multilingual society with 24 EU Member State languages and dozens of additional languages including regional and minority languages and languages spoken by immigrants, trade partners and tourists. The only option to enable and to benefit from multilingualism is through Language Technologies (LT) including Natural Language Processing (NLP) and Speech Technologies (Rehm, 2017). While the European LT landscape is world class, it is also massively fragmented (Vasiļjevs et al., 2019; Rehm et al., 2020d).

We describe Release 2 of the European Language Grid (ELG) cloud platform.[1] This scalable system is targeted to evolve into the primary platform for LT in Europe. It will provide one umbrella platform for all LTs developed by the European LT landscape, including research and industry, addressing a gap that has been repeatedly raised by the European LT community for many years (Rehm and Uszkoreit, 2013; Rehm et al., 2016b; STOA, 2017; Rehm, 2017; Rehm and Hegele, 2018; European Parliament, 2018). ELG is meant to be a virtual home and marketplace for all products, services and organisations active in the LT space in Europe (Rehm et al., 2020a). The platform can be used by all stakeholders [illegible] case, share and distribute their products, tools and resources. At the end of the EU ELG (2019-2022), which will establish a [illegible] tity in early 2022, the platform will provi[illegible] to approx. 1300 commercial and non-com[illegible] tools and services for all European langu[illegible] well as thousands of language resource[illegible] ELG will enable the European LT comm[illegible] deposit and upload their technologies and [illegible] and to deploy them through the grid. T[illegible] is also meant to support *digital languag[illegible] ity* in Europe (STOA, 2017; European Pa[illegible] 2018), i. e., to create a situation in whic[illegible] guages are supported through technologie[illegible] well. The current imbalance is characteri[illegible] stark predominance of LTs for English, [illegible] most all other languages are only marginally supported and, thus, in danger of digital language extinction (Rehm and Uszkoreit, 2012; Kornai, 2013; Rehm et al., 2014, 2016a; ELRC, 2019).

Section 2 gives an overview of the ELG platform and related activities. Section 3 touches upon related work. Section 4 concludes the paper.

[1]https://www.european-language-grid.eu. We provide a screencast demo video at https://youtu.be/LD6QadkkZiM.

Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability

Georg Rehm[1], Dimitrios Galanis[2], Penny Labropoulou[2], Stelios Piperidis[2], Martin Welß[3], Ricardo Usbeck[3], Joachim Köhler[3], Miltos Deligiannis[2], Katerina Gkirtzou[2], Johannes Fischer[4], Christian Chiarcos[5], Nils Feldhus[1], Julián Moreno-Schneider[1], Florian Kintzel[1], Elena Montiel[6], Victor Rodriguez Doncel[6], John P. McCrae[7], David Laqua[3], Irina Patricia Theile[3], Christian Dittmar[4], Kalina Bontcheva[8], Ian Roberts[8], Andrejs Vasiļjevs[9], Andis Lagzdiņš[9]

[1] DFKI GmbH, Germany • [2] ILSP/Athena RC, Greece • [3] Fraunhofer IAIS, Germany • [4] Fraunhofer IIS, Germany • [5] Goethe University Frankfurt, Germany • [6] Universidad Politécnica de Madrid, Spain • [7] National University of Ireland Galway, Ireland • [8] University of Sheffield, UK • [9] Tilde, Latvia

Corresponding author: Georg Rehm georg.rehm@dfki.de

Abstract

LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

IWLTP 2020 – 1st International Workshop on Language Technology Platforms

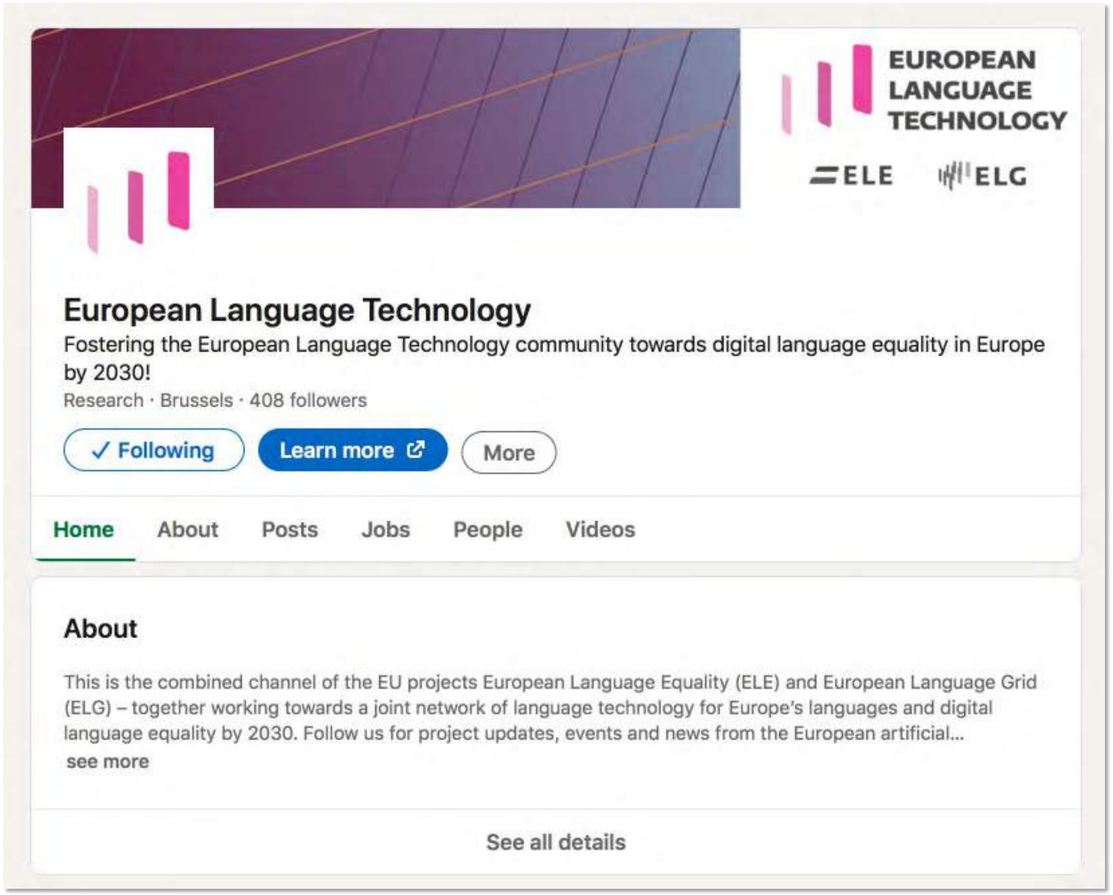
PROCEEDINGS

Edited by:
Georg Rehm, Kalina Bontcheva, Khalid Choukri,
Jan Hajič, Stelios Piperidis, Andrejs Vasiļjevs

Coming up:

- European Language Grid book (2022)
- European Language Equality book (2022)

Plus: ELG Online Documentation



EUROPEAN LANGUAGE TECHNOLOGY

EUROPEAN LANGUAGE GRID

EUROPEAN LANGUAGE EQUALITY

More than 14k visitors on Twitter already!
More than 65k impressions already!

https://www.linkedin.com/company/european-language-technology

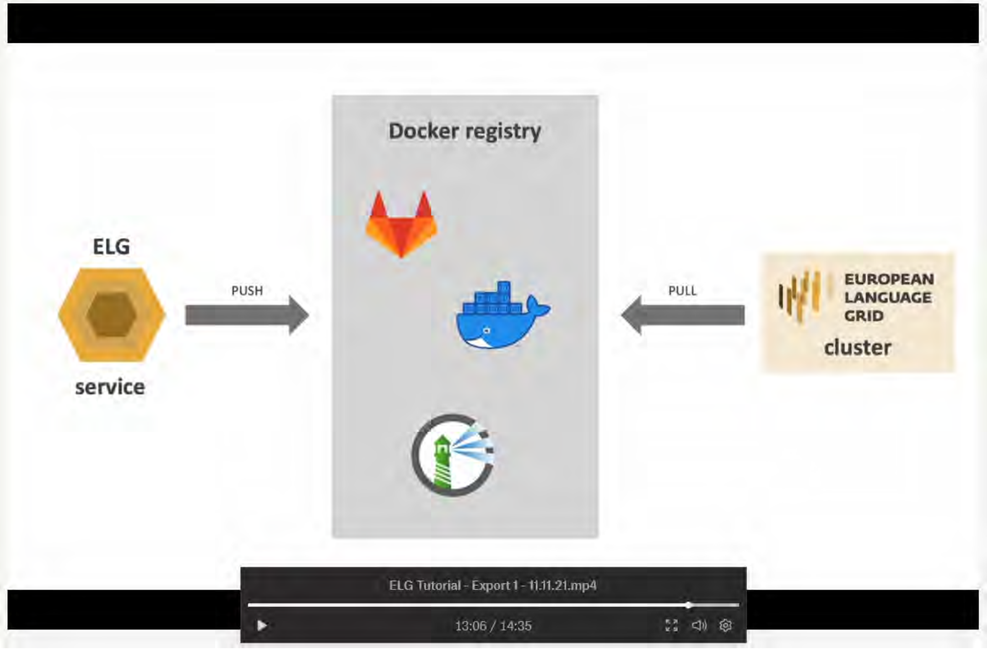
https://twitter.com/EuroLangTech

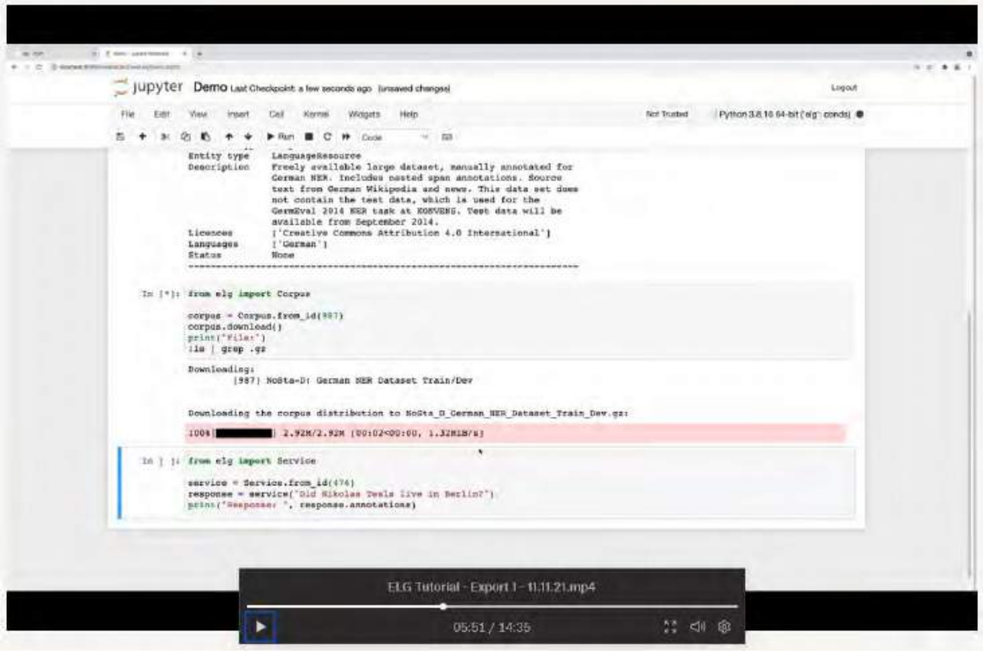
https://www.european-language-technology.eu

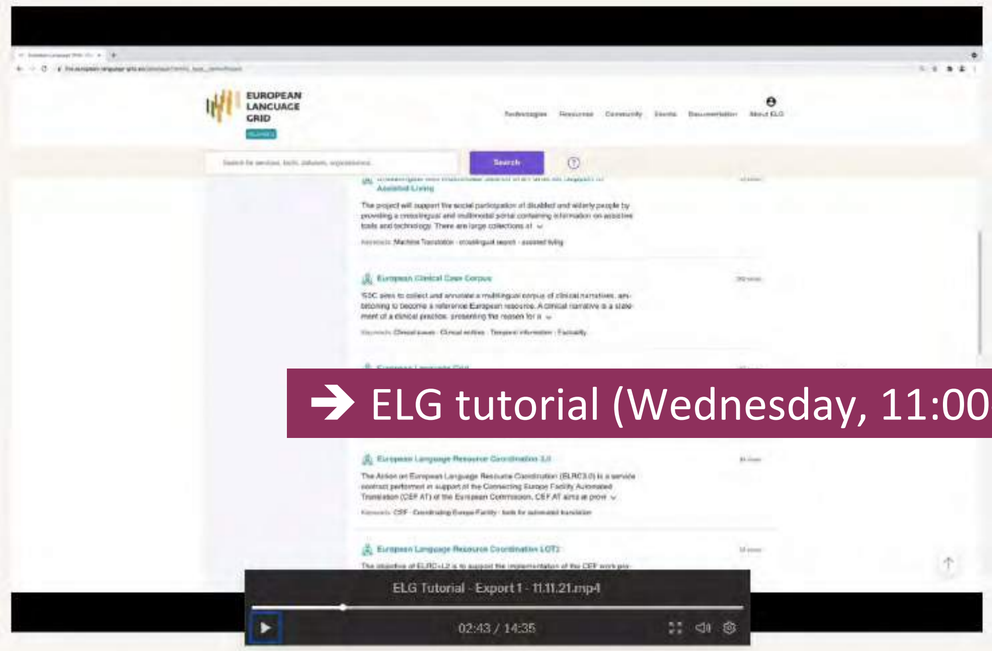
Subscribe to our newsletter

More than 2200 subscribers already!

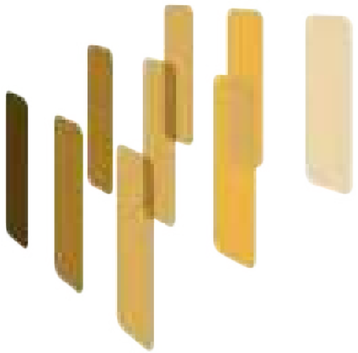
Coming soon: ELG tutorial video that explains how to make resources and tools available.







➔ ELG tutorial (Wednesday, 11:00-12:30)



European Language Grid

Thank you!

The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825627 (ELG).

Georg Rehm (DFKI), Stelios Piperidis (ILSP/R.C. Athena), Kalina Bontcheva (University of Sheffield),
georg.rehm@dfki.de, spip@athenarc.gr, k.bontcheva@sheffield.ac.uk

15/16/17-11-2021 META-FORUM 2021 – Using the European Language Grid (virtual conference)
http://www.european-language-grid.eu