

Chapter 9

Language Technology Companies, Research Organisations and Projects

Georg Rehm, Katrin Marheinecke, Rémi Calizzano, and Penny Labropoulou

Abstract The European Language Grid is meant to develop into the primary platform of the European Language Technology community. In addition to LT tools and services (Chapter 7) and Language Resources (Chapter 8), ELG represents the actual members of this community, i. e., the companies and research organisations that develop language technologies and that are engaged in related activities. The goal of becoming the primary platform for LT in Europe implies that ELG should ideally represent *all* European companies and *all* European research organisations with corresponding metadata records in the ELG catalogue, which are interlinked with the respective LT tools and services as well as language resources they offer. This chapter describes the European stakeholders and user groups that are relevant for the ELG initiative, the composition of the community and the locations of the companies and research groups as currently listed in ELG. Furthermore, we describe a number of technical and organisational challenges involved in the preparation of our list of stakeholders, and outline the process of catalogue population.

1 Introduction

The European Language Grid is meant to develop into the primary platform of the European LT community. This is why, in addition to functional LT tools and services and more static Language Resources (LRs), ELG also represents the actual members of this community, i. e., the companies and research organisations that develop LTs and that are engaged in related activities such as the integration of LT into existing systems or support services such as data annotation at scale. This overall goal of eventually establishing ELG as the primary platform for LT in Europe implies that ELG should ideally represent *all* European companies and *all* European research

Georg Rehm · Katrin Marheinecke · Rémi Calizzano

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de, katrin.marheinecke@dfki.de, remi.calizzano@dfki.de

Penny Labropoulou

Institute for Language and Speech Processing, R. C. "Athena", Greece, penny@athenarc.gr

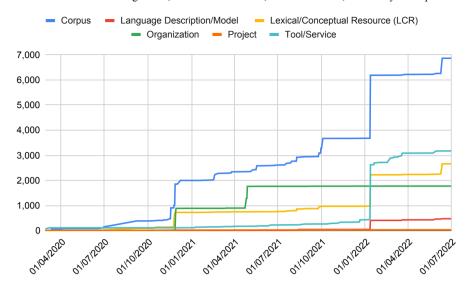


Fig. 1 Evolution of resources in ELG over time broken down by resource type

organisations in the ELG catalogue, which are interlinked with the respective LT tools and services as well as language resources these organisations offer on and through the European Language Grid. In other words, the European Language Grid also functions as the "yellow pages" of the European LT community, ideally listing and promoting *all* relevant members of this community, i. e., small and medium-sized companies as well as large enterprises, research centers, universities and other academic institutions that develop LT but also organisations in the periphery of this core, e. g., integrators and annotation service providers (Rehm et al. 2020, 2021). ¹

In addition to serving as the central directory for members of the European LT community, ELG also includes information about relevant projects in the area.² The reasoning behind this is the way many LTs are typically developed, i. e., through publicly funded project consortia in which academic or commercial organisations participate. Such projects often result in concrete tools and technologies as well as language resources, which can then be made available, among others, through ELG, which allows representing and interlinking these project artefacts (LTs, LRs), the projects that helped create these artefacts and the members of the respective project consortia. Technically, project consortia can provide relevant metadata to create and later edit and update their own project pages in ELG ensuring more visibility as well as an additional dissemination channel for their projects' outputs.

In the second half of the ELG project's runtime, corresponding activities in terms of populating the ELG catalogue with information about companies, academic organisations and projects have been drastically increased so that, towards the end of the project, ELG now includes convincing figures in terms of community members,

¹ https://live.european-language-grid.eu/catalogue/?entity_type term=Organization

² https://live.european-language-grid.eu/catalogue/?entity_type term=Project

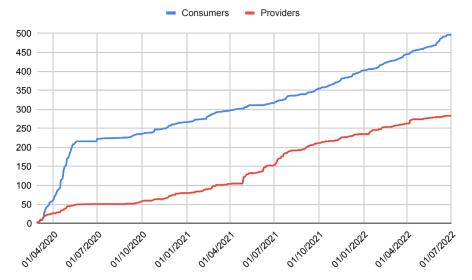


Fig. 2 Number of ELG consumer and provider accounts over time

projects and also active users of the platform. At the time of writing, ELG lists more than 13,000 metadata records on tools and services, resources, organisations and projects. Figure 1 shows the corresponding development of the ELG catalogue and its population over time, differentiated by type of entry.

Not only the number of resources and organisations listed in ELG is constantly growing. In addition, the number of users is rising continuously. The number of ELG users of the consumer category who have a registered a user account went up significantly at the end of April 2020, after the first official release to the public, and has grown further ever since. The number of ELG users of the provider category, i. e., users with the right to integrate metadata, tools and resources in ELG, is also increasing continuously, albeit more slowly, as can be expected (see Figure 2).

As encouraging as this development is, ELG is still at the beginning. The platform has been designed in such a way that it can be actively used by the community and that it can grow. To achieve this goal of a true one-stop shop for the whole European LT community, it is necessary to steadily expand the consumer and provider base and monitor as well as reflect all changes and new developments in the European LT landscape. Only with this momentum will the desired snowball effect be generated eventually, which ultimately helps ELG to achieve sustainable success from which all stakeholders can benefit.

2 The European Language Technology Landscape

One key characteristic of the European Language Technology landscape is its extreme fragmentation, which has been mentioned repeatedly throughout the years, as, for example, in the META-NET White Paper Series (Rehm and Uszkoreit 2012), in the META-NET Strategic Research Agenda (Rehm and Uszkoreit 2013; Rehm et al. 2016), in the *Final study report on CEF automated translation value proposition in the context of the European LT market/ecosystem* (Vasiljevs et al. 2019) or in the various reports of the European Language Equality project (especially see Aldabe et al. 2022). In fact, this extreme fragmentation is one of the main reasons why the ELG platform has been developed in the first place because the fragmentation is generally perceived as one of the main reasons why the European LT community has been unable to unleash its full potential.

The analysis in the CEF LT Market study (Vasiljevs et al. 2019) shows that European LT vendors are often SMEs with local or regional, often highly specialised solutions. In the study, 473 companies were collected that are active in EU member states in the domain of LT and that fully qualify as LT vendors. According to the research, the total size of the LT industry within the EU member states (plus Iceland and Norway) was estimated at approx. 800M€ in the year 2017. In the study sample investigated, only 14% of the LT vendors had a revenue of more than €10M, whereas almost half of them (48%) had a revenue below €1M. In terms of size, 52% of the companies had between 10 and 99 employees, and 26% had less than 10 employees, both combined representing nearly 80% of the 473 companies studied. Only 44% of the EU companies in this sample received external funding or venture capital.

Consequently, the global LT and NLP market continues to be dominated by large technology enterprises from the United States and Asia which establish "data-driven intellectual monopolies" (Rikap and Lundvall 2020) – in that regard, large companies are the exception in Europe. However, these big non-European LT providers have certain deficiencies regarding under-resourced languages, customisation needs, as well as security and privacy requirements which is a frequently expressed demand from corporate clients and European administrations (Overton 2017).

Despite the fact that the LT market is relatively small when compared to the general IT market at large, it is a market with strong competition, which is one of the reasons why many LT developing companies tend to focus on highly specialised niche markets with less intense competition. This, however, affects profitability, which is, on average, rather low and margins are compressed. On the other hand, LT can also be considered a growing market: today, (potential) customers have more awareness of the benefits of LT, which is also due to marketing activities of large international players. From a local vendors' point of view, the large technology enterprises help create a market awareness that simply did not exist ten years ago. Nevertheless, these companies are also the toughest competition of the European LT community as they tend to offer high-quality LT software free of charge or for very low prices, which European SMEs usually cannot afford to do.

The STOA study Language equality in the digital age – Towards a Human Language Project (STOA 2018), which examines the causes of language barriers in

Europe and formulates recommendations for policies to overcome these barriers, mentions among its 11 key recommendations the need for a pan-European LT Platform of resources and services and ELG has stepped up to solve this problem (also see European Parliament 2018). ELG not only brings together LT resources from all over Europe supporting almost all European languages (although ELG is not limited to European languages) but ELG also has the ambition to unite the European LT community behind these services, tools and resources using one shared umbrella platform to create a common access point and marketplace from which all languages and members of the community will eventually benefit (see Part III of this book).

At the time of writing, ELG contains approx. 1,800 organisations operating in the European LT sphere. One half of these organisations consists of companies, the other half of universities and research groups (Figure 3).³

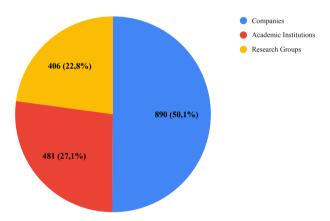


Fig. 3 Distribution of organisations listed in ELG per type

The quantitative distribution of LT developing organisations among the respective countries in Europe already hints at a strongly varying coverage of LT resources for their respective national and regional languages. Whereas countries like the UK, Germany or Spain are well or relatively well equipped with LT developing companies, smaller countries like Malta or Cyprus have only little representation in the European LT community (see Figure 4).⁴ Figure 5 shows the geographical distribution in Europe of organisations listed in ELG.

³ Companies are commercial organisations, academic institutions are universities and research centers, research groups are sub-groups of academic institutions, e. g., faculties or departments.

⁴ In Figure 4, countries are ordered by decreasing number of organisations. The country with the head office of the respective organisation is used as the organisation's country.

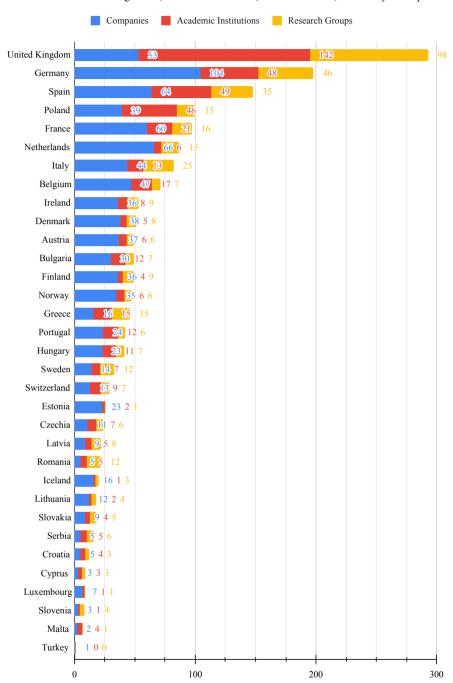


Fig. 4 Distribution of organisations listed in ELG per type and country

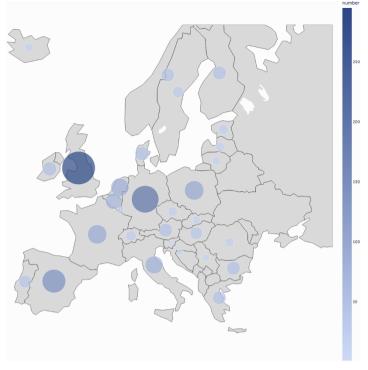


Fig. 5 Organisations listed in ELG per country

3 Organisations in the European Language Grid

To bootstrap the ELG catalogue with as many LT developing European companies and academic organisations as possible, we decided on the following procedure. First, together with the ELG National Competence Centres (see Chapter 11, p. 205 ff.), we collected LT developing organisations semi-automatically and in a decentralised way, i. e., on the national level (Section 3.1). Second, based on the results of this collection, metadata records were prepared that could be automatically ingested into the ELG catalogue (Section 3.2). This resulted in the ELG catalogue being populated with approx. 1,800 metadata records, i. e., pages, each of which describes one LT developing organisation with a basic profile. These organisation profiles can then be claimed by the rightful owners (Section 3.3), i. e., an organisation described in such an ELG page can take over the maintenance of its own page and enrich it with additional information, e.g., upload a logo, associate resources with their organisation etc. (Section 3.4). This bootstrapping procedure enables members of the European LT community to participate actively in ELG with their own organisation within minutes. As a positive side effect, it enabled ELG – including its sister project ELE – to produce a fairly detailed picture of the European LT landscape.

3.1 Collecting the Members of the European LT Community

In order to populate ELG with organisations, we used our own databases, carried out desk research and, most importantly, we involved the 32 National Competence Centres (NCCs) to tap into their detailed knowledge of their respective countries' LT communities. Our general goal was to identify and to record, in a machine-readable format, as many national and regional members of the European LT community as possible so that ELG can eventually provide as complete and up to date a picture as possible. In September 2020, this data collection task was conducted with NCC Leads representing their countries and regions to ideally identify all companies and academic organisations in the European LT community to be listed in ELG.

To streamline the process, based on data gathered in various workshops, conferences and other events over the last ten years, the ELG project team created lists of organisations involved in LT activities in all European countries. Each entry in the list contained, among others, the following information: organisation name, department name, website, address (region, ZIP code, city, country) and LT areas in which they are active. Each NCC Lead received the data records for their country, along with detailed guidelines, and they were asked to check the data included in the list, to correct the data if necessary (e. g., remove duplicates with similar names, correct wrong names of organisations) and to complete them where possible, i. e., to fill in blanks. Furthermore, the NCCs were asked to do their own research and provide new, unlisted organisations. The goal was to find all relevant organisations of each country that develop, market or sell LT in their countries. This way, the ELG consortium wanted to ensure that in addition to well-known organisations also start-ups andyoung research groups are included in ELG.

The feedback received from the NCCs was submitted to a comprehensive internal quality review by the ELG team, which resulted in the final dataset that reflects a fairly complete representation of the relevant stakeholders and providers of Language Technology and language-centric AI in Europe.⁵

3.2 Preparation and Integration of Metadata Records

The efforts of the NCCs and the ELG team for the collection of data regarding LT organisations relevant for ELG resulted in two spreadsheets per country containing companies and research groups respectively. All entries were automatically converted into XML files that are compliant with the ELG metadata schema as described in Chapter 2. Furthermore, for columns corresponding to metadata elements that take values from controlled vocabularies (e. g., LT area), we mapped the input to the values in the controlled vocabulary. This process also served as a sanity check during

⁵ In this procedure, the regulations of the Data Protection Act were adhered to at any time and no personal data have been published without the consent of the data owners.

which errors were identified and resolved. The procedure resulted in 1,740 XML files, 867 for companies and 873 for research groups.

The ELG life-cycle for the publication of individual resources includes a validation process aiming to ensure the quality of the metadata published in ELG (see Chapter 2). For the import of the organisation-related XML files, we applied a different procedure that involved their bulk import with the assignment of the tag "imported by ELG". Metadata records marked as such do not go through a validation process and are immediately published on ELG.

3.3 Claiming and Enriching Organisation Pages

Once the population of ELG with these entries was completed, a campaign was launched inviting (via email) legitimate owners to claim, edit and curate the entries of their own organisations. Since the pages created by the ELG team contained only minimal information, the representatives of the organisations were invited to enrich these pages with reliable and accurate content and also to start providing tools, services and resources. In several email campaigns, we reached out to contact persons identified by the NCCs and we informed them about the existence of their organisations' pages on ELG, also inviting them to take over the pages. To do so, the legitimate owner can "claim" their organisation's page as their own by clicking the "Claim" button on the page (see Figure 6).



Fig. 6 Imported organisation page with a "Claim" button

The claiming process can only be triggered by persons signed in with an ELG account (with provider role). This step serves as a security mechanism ensuring correct and rightful authorisation of eligible persons. Once a request is made, the ELG team checks its validity, which also includes checking the email address used to register the ELG account, making sure that it belongs to the organisation, the page of which is being claimed. Approval of the request entails that the entry is assigned

to the claimant and returns to a status that it can be edited. The claiming person is prompted by email that they can now start editing the metadata entry and ELG page. Once edited, the page needs to be submitted to publication and the usual ELG validation process starts, i. e., the changes made to the resource are reviewed by the ELG team and the entry is made publicly available again.

3.4 Organisation Pages in the European Language Grid

Organisation pages can include different tabs. The "Overview" tab includes a description of the organisation as well as an info box on the right with data such as postal address and contact email as well as a link to the organisation's own website. This tab can also include keywords that describe the general domain and LT areas an organisation addresses. ELG pages can also be exported in XML format. The tab "Related LRTs & projects" lists all resources and technologies the respective organisation has made available on ELG and the projects they are involved in. This helps companies to promote their tools and resources and to show connections between companies or research organisations and their research projects and corresponding results. The "Related organisations" tab is especially important for academic institutions and universities to reflect their relationship to other departments, faculties or the umbrella organisation (usually the university). Figure 7 provides an example for a page of an academic organisation. Figure 8 (p. 182) shows a company page.

4 Projects in the European Language Grid

ELG is also able to represent research projects, especially for the purpose of acknowledging the funding that made the development of a technology or resource possible and also to interlink projects with organisations and resources. ELG project pages are structured in a similar way, but they are especially adapted to the characteristics and metadata of a typical research project. In addition to information regarding the start and end of the project, the info box also contains details on the funding agency, the funding country, the type of project and the amount of funding provided. Besides the project description and keywords, the "Overview" tab contains the list of consortium partners, that are linked to their respective ELG pages if they exist. Again, the tab "Related LRTs" lists all technologies and resources associated with or resulting from the project. Two examples are shown in Figures 9 (p. 183) and 10 (p. 184).

⁶ At the time of writing, we are preparing a list with more than 500 projects that will be imported into the ELG catalogue in the second half of 2022; this list was put together in a similar manner as the list of organisations described in Section 3.1.

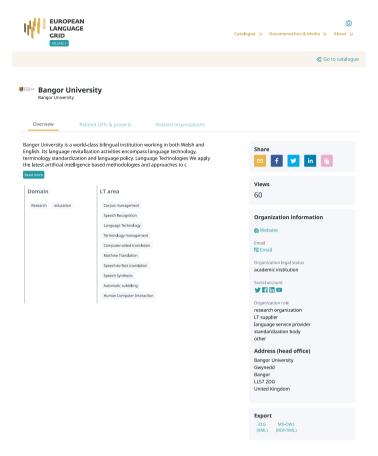


Fig. 7 Example ELG organisation page: Bangor University

5 Conclusions

The European Language Grid is meant to develop into the primary platform of the European LT community. In addition to the technical resources, ELG also represents the actual members of this community: companies and research organisations that develop LTs and related organisations. Our ambition is for ELG eventually to represent *all* companies and *all* research organisations active in the European LT community. In order for ELG to function as a marketplace for European LT, it also needs to provide core information about the European LT community (i. e., "yellow pages" functionalities).

ELG currently contains approx. 1,800 organisations active in the European LT community. Like every similar repository or digital catalogue with certain artefacts, one of the key challenges is the maintenance of the records and metadata entries, i. e., keeping the entries up to date and also making sure that the community is fully

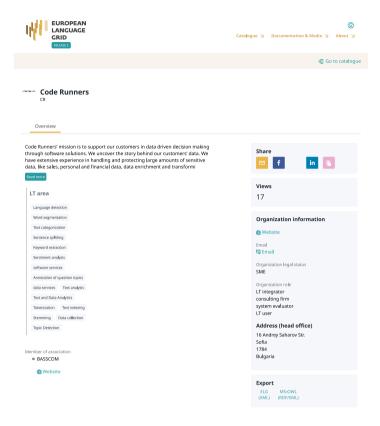


Fig. 8 Example ELG organisation page: Code Runners

represented. Our long-term vision for ELG is to become the primary platform of the European LT community, which entails that all members of the European LT community, both commercial and academic, immediately recognise the value, importance and relevance of ELG and, thus, actively want to participate in ELG, keepting their pages up to date, sharing technologies and resources, benefiting from this European marketplace. Until this intended snowball effect is fully in place, i. e., all stakeholders recognise the benefit ELG brings about and participate actively, we will perform, even if time-consuming and logistically challenging, manual updates of the ELG catalogue, we will continue to convert as many members of the community as possible into active users and also active providers of ELG and we will increase our our outreach activities, encouraging more organisations to claim their ELG pages. As soon as the snowball effect is in place and ELG is accepted as the primary platform of the European LT community, all participating organisations will have a sufficient amount of intrinsic motivation to maintain their ELG pages and to keep their information, technologies and resources up to date. At this time, ELG strives to be an established player, which is known throughout the community so

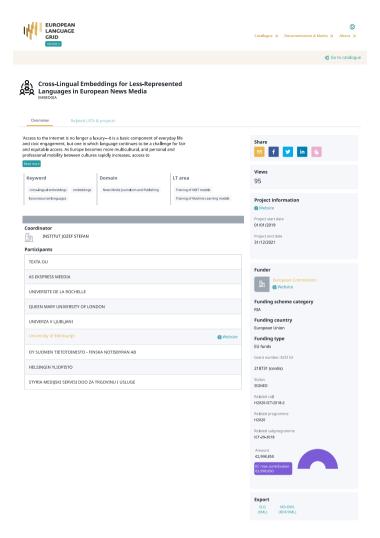


Fig. 9 Example ELG project page: EMBEDDIA (Overview)

that also new companies are attracted by and to ELG. In addition to simplifying the claim process, the attractiveness of ELG will be further enhanced through increased community-related promotions, new features and improved offerings.

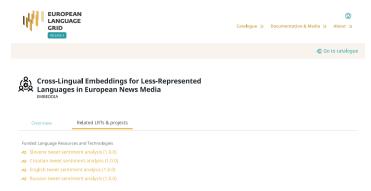


Fig. 10 Example ELG project page: EMBEDDIA (Related LRTs)

References

Aldabe, Itziar, Georg Rehm, German Rigau, and Andy Way (2022). *Deliverable D3.1 Report on existing strategic documents and projects in LT/AI (second revision)*. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE. URL: https://european-language-equality.eu/wp-content/uploads/2022/06/ELE___Deliverable D3 1 second revision 2.pdf.

European Parliament (2018). Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI). URL: http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Overton, David (2017). Next Generation Internet Initiative – Consultation. URL: https://ec.europa.eu/futurium/en/system/files/ged/ec_ngi_final_report_1.pdf.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiljevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. URL: https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiljevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings*

of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021). Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Rehm, Georg and Hans Uszkoreit, eds. (2012). META-NET White Paper Series: Europe's Languages in the Digital Age. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Rehm, Georg and Hans Uszkoreit, eds. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020.* Heidelberg, New York, Dordrecht, London: Springer. URL: http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf.

Rehm, Georg, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen García-Mateo, Josef van Genabith, Jan Hajič, Inma Hernáez, John Judge, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Joseph Mariani, John McNaught, Maite Melero, Monica Monachini, Asunción Moreno, Jan Odjik, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Sandford Pedersen, Inguna Skadina, Koenraad De Smedt, Marko Tadić, Paul Thompson, Dan Tufiş, Tamás Váradi, Andrejs Vasiljevs, Kadri Vider, and Jolanta Zabarskaite (2016). "The Strategic Impact of META-NET on the Regional, National and International Level". In: Language Resources and Evaluation 50.2, pp. 351–374. DOI: 10.1007/s10579-015-9333-4.

Rikap, Cecilia and Bengt-Åke Lundvall (2020). "Big Tech, Knowledge Predation and the Implications for Development". In: *Innovation and Development*, pp. 1–28. DOI: 10.1080/2157930 X.2020.1855825.

STOA (2018). Language equality in the digital age – Towards a Human Language Project. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. URL: https://data.europa.eu/doi/10.2861/136527.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). Final study report on CEF Automated Translation value proposition in the context of the European LT market/e-cosystem. DOI: 10.2759/142151. URL: https://op.europa.eu/de/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

