Selected Infrastructure Activities around Language Data at DFKI

Prof. Dr. Georg Rehm

DFKI GmbH and **Humboldt-Universität zu Berlin**
georg.rehm@dfki.de

06 June 2024 – Inria and ECDF Collaboration Event – Berlin, Germany

Outline

- European Language Data Space (LDS)
- European Language Grid (ELG)
- NFDI for Data Science and Artificial Intelligence (NFDI4DS)

European Language Data Space

Relevant Observations

- Large language models are the most disruptive breakthrough in AI in recent history (BERT, GPT-3, ChatGPT, GPT-4 etc.)
- LLMs are trained on vast amounts of training data (language data)
- LLMs use dozens, some even hundreds of terabytes (trillions of tokens) of language and also image, video, audio etc. training data
- Europe's languages are vastly under-resourced, except English
- In 2024, language data is more relevant than ever
- A concerted effort for the collection of (very large amounts of) language data for all European languages is very much needed
- The global NLP/LT/Gen-AI market is in the hundreds of billions of US-$ already and expected to grow to 439.85B US-$ by 2030 – no significant players from Europe
- Already now billions and billions are made but …

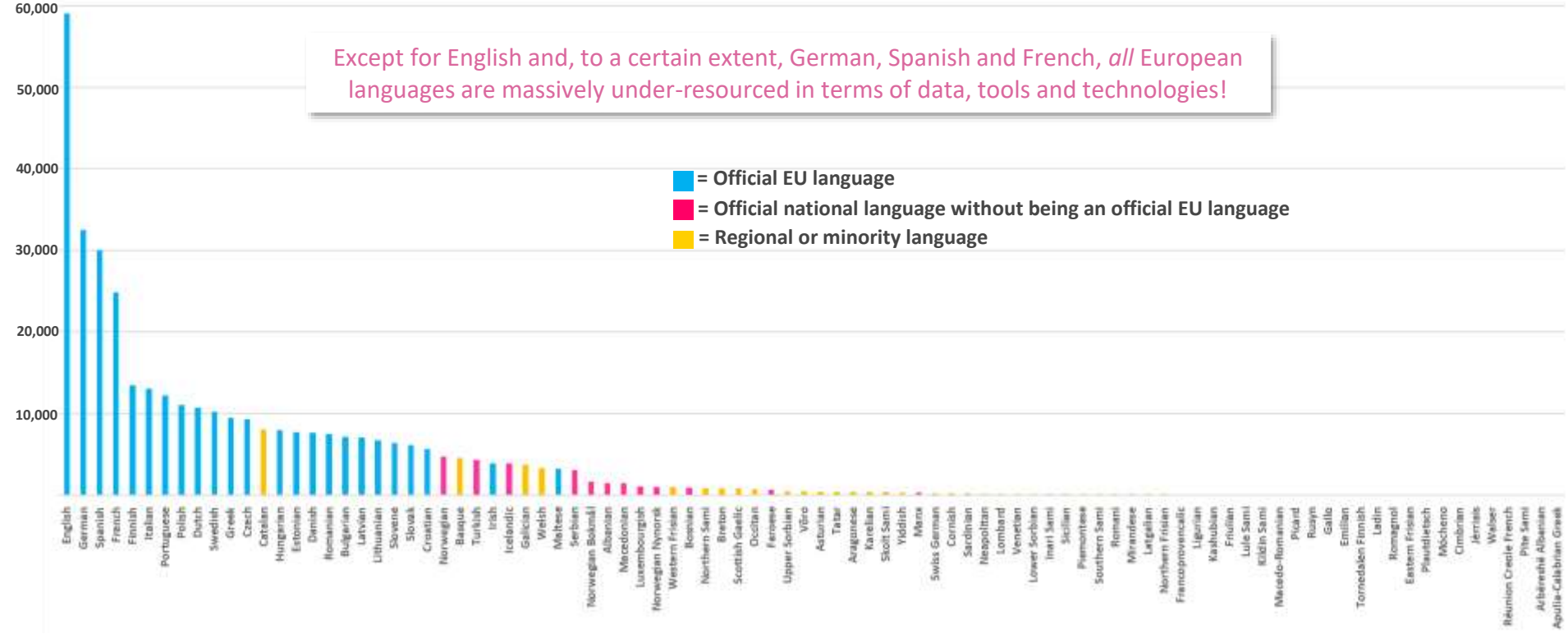

The Washington Post
Democracy Dies in Darkness
Sign in
BUSINESS
ChatGPT Shows Just How Far Europe Lags in Tech
Analysis by Lionel Laurent | Bloomberg
February 21, 2023 at 2:12 a.m. EST
Comment 1
Gift Article
Share
Europe is where ChatGPT gets regulated, not invented. That's something to regret. As unhinged as the initial results of the artificial-intelligence arms race may be, they're also another reminder of how far the European Union lags behind the US and China when it comes to tech.

European Initiatives

- European initiatives for the development of LLMs
  - Large research projects in almost every country, e.g., Spain, Denmark, Italy, Germany etc.
  - Companies in many countries, e.g., Finland (Silo.ai), France (Mistral), Germany (Aleph Alpha)
  - EU and nationally funded projects, e.g., HPLT, TrustLLM
  - New pan-European initiative: ALT-EDIC
- Challenges:
  - HPC facilities
  - Speed of the big tech players in the US and Asia vs. speed of Europe
  - Availability of data for European languages

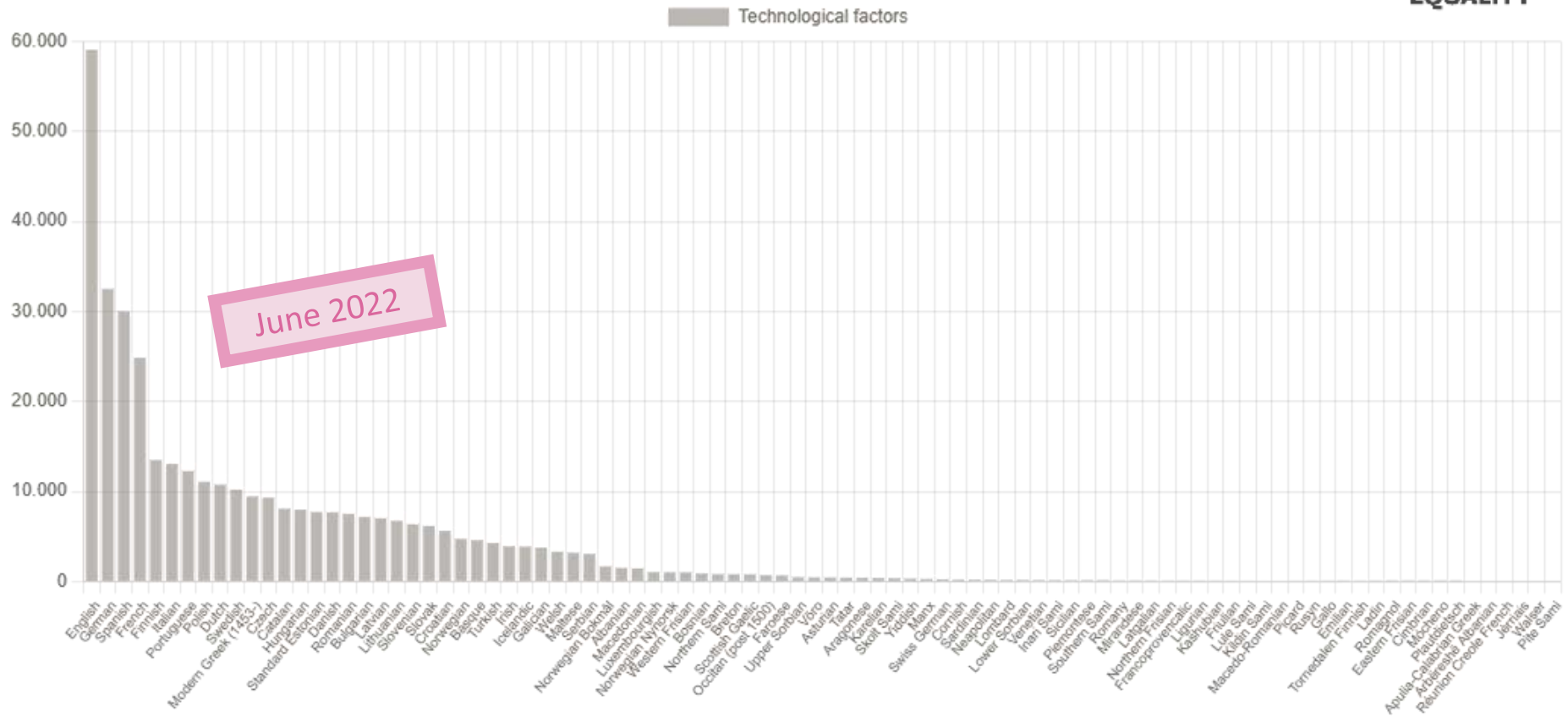
Digital Language Equality Metric: Technological Scores

EUROPEAN LANGUAGE EQUALITY



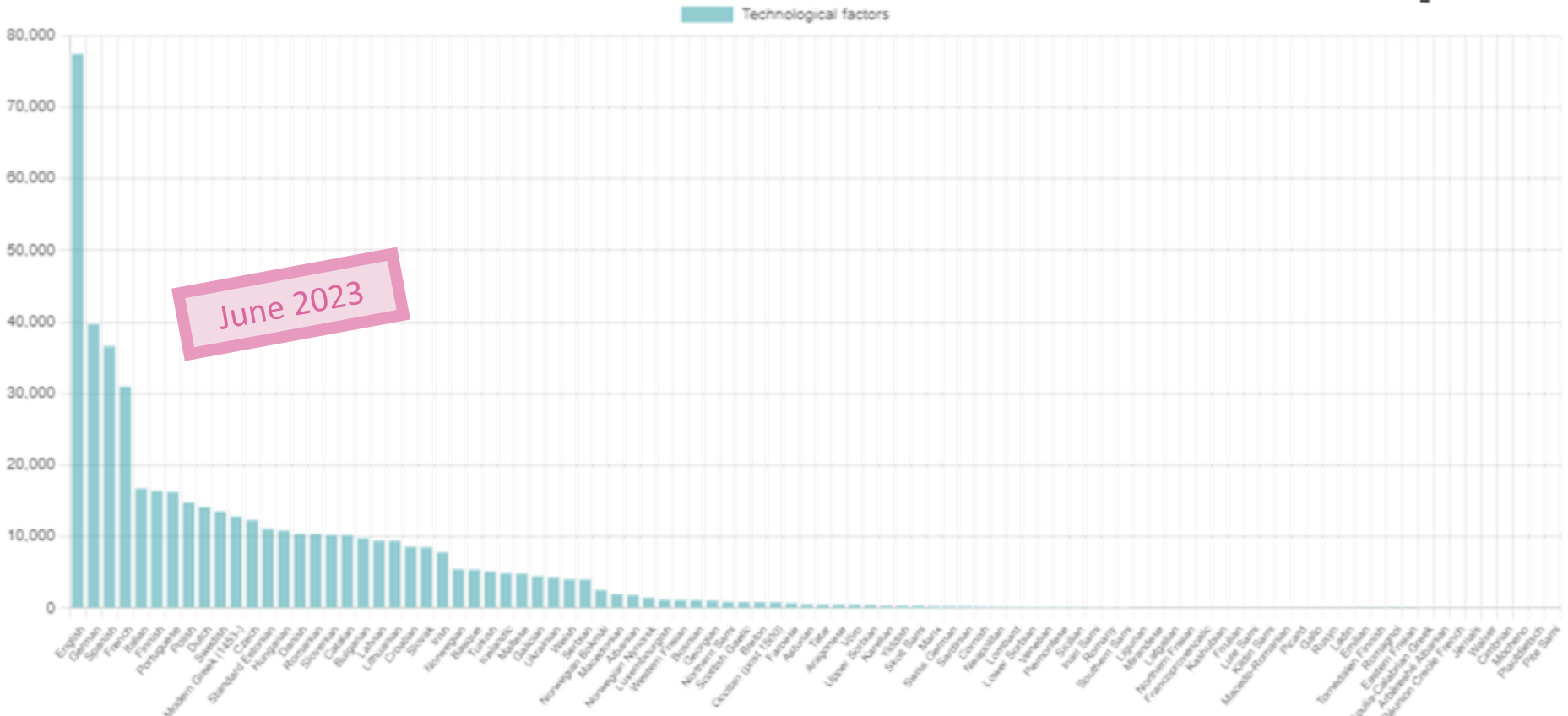
DLE Metric: 2022 vs. 2023 (1/3)

EUROPEAN LANGUAGE EQUALITY



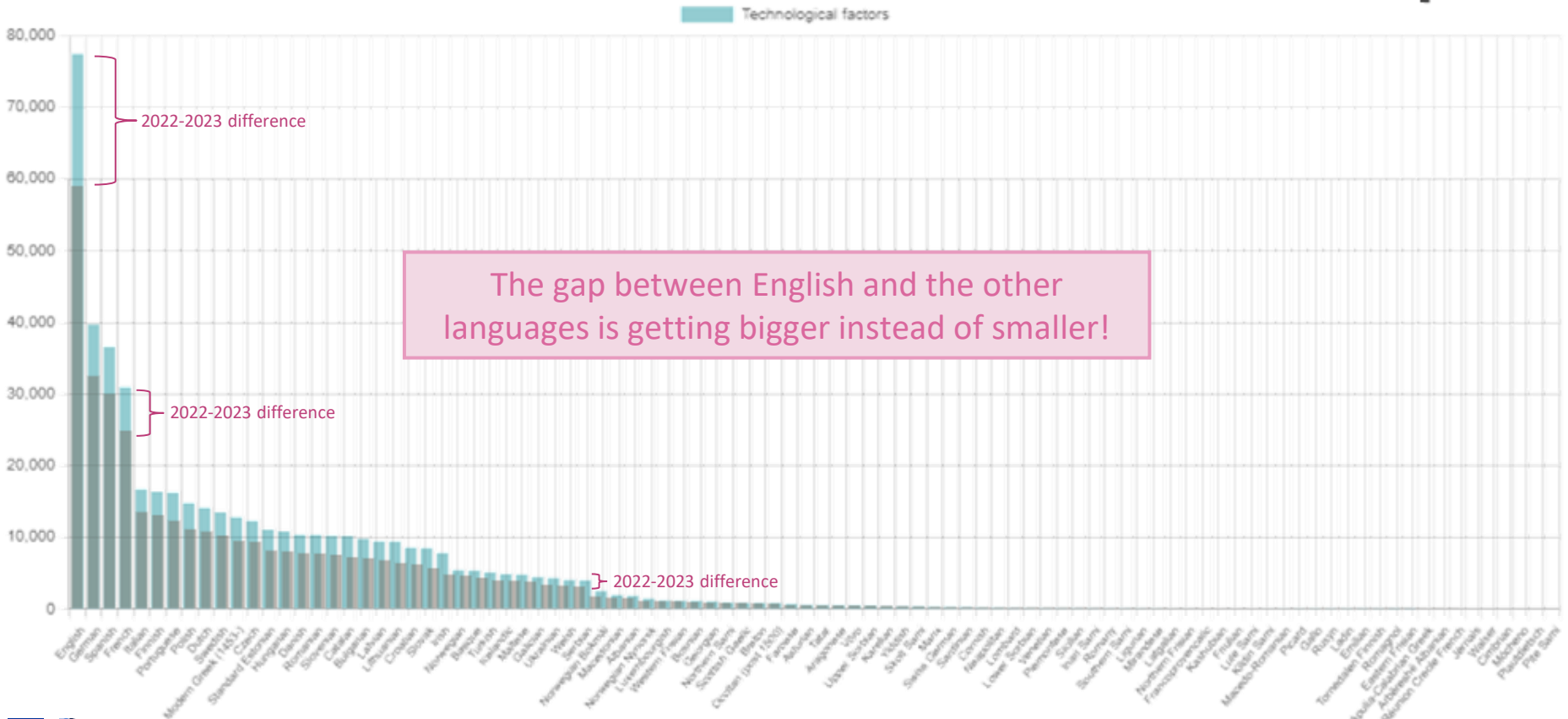
DLE Metric: 2022 vs. 2023 (2/3)

EUROPEAN LANGUAGE EQUALITY



DLE Metric: 2022 vs. 2023 (3/3)

EUROPEAN LANGUAGE EQUALITY



EU Data Strategy & Data Spaces

- Data Spaces are an inherent part of the EU Data Strategy
- Data Spaces will help to establish a data economy in Europe
- Various data economy and data infrastructure initiatives in Europe with slightly different goals and individual positioning but conceptual, technical, legal and operational overlap:
  - Data Spaces Business Alliance (DSBA): Gaia-X, IDSA, FIWARE, BDVA
  - EU: DSSC (incl. DSBA), Simpl, approx. 20 data spaces eventually
- The Common European Language Data Space is one of the 14 official EU data space projects with a strong focus on industry

Common European Language Data Space

- Type of action: procurement (CNECT/LUX/2022/OP/0026)
- Budget: 6M€ (+ 2M€ if renewed)
- Runtime: 36 months (+ 12 months if renewed)
- Objective: Develop and deploy a European platform and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data
- Salient features: governance framework, technical architecture and infrastructure, openness, promotion
- Stakeholders: industry, research, public administration, cultural associations, NGOs and citizens

Consortium and Subcontractors

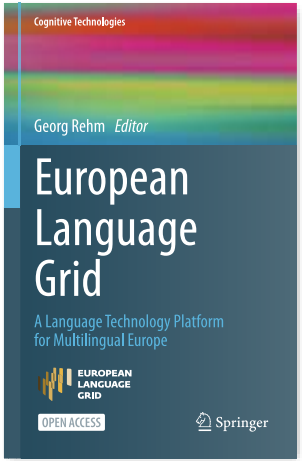
Coordinator		
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH	DFKI	DE
Partners and Operation Leads		
R.C. “Athena”, Institute for Language and Speech Processing	ILSP	GR
Evaluations and Language Resources Distribution Agency	ELDA	FR
TILDE	TILDE	LV
Main Subcontractors		
3pc GmbH Neue Kommunikation	3pc	DE
CLARIN ERIC	CLARIN	NL
Big Data Value Association (Data, AI and Robotics) AISBL	BDVA	BE

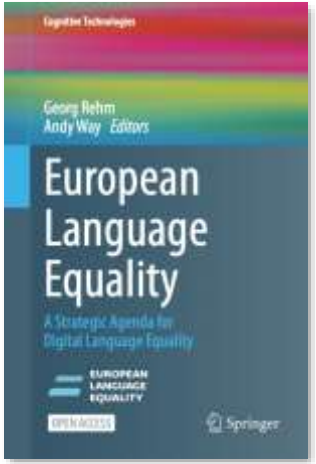
Plus legal experts (Delcade, France) and approx. 30 organisations for the logistics of multiple country workshops

Previous Projects and Initiatives

- The four core partners – DFKI, ILSP, ELDA, TILDE – have been involved in many projects, including:
- **META-NET** (FP7, 2010-2013)
  - META-SHARE
- **ELRC** (CEF, 2014-2023)
  - ELRC-SHARE
- **ELG** (H2020, 2019-2022)
  - ELG Cloud Platform
- **ELE** (PP/PA, 2021-2023)

The **technical development work in LDS** will be informed by ELG, ELRC-SHARE, META-SHARE.





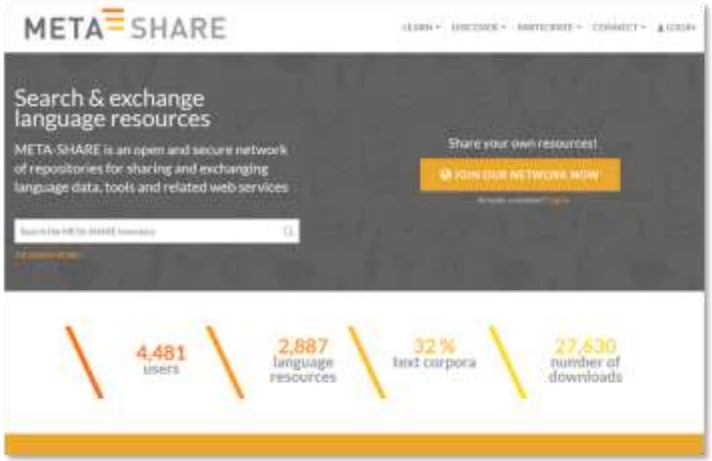
Classes of Data

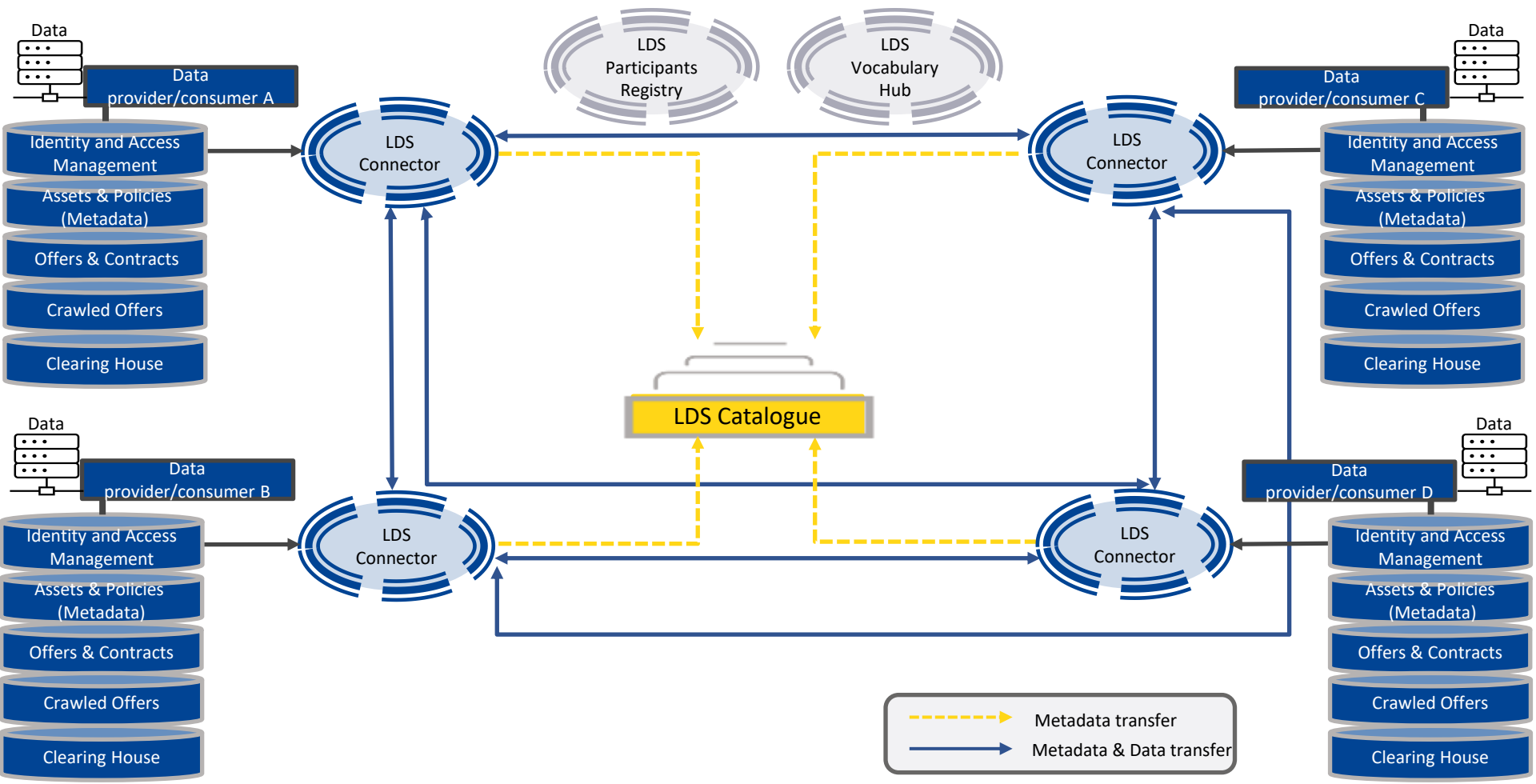
Class of Data	Typical Size	Providers	Integration into LDS	Relevance for LLMs
Regular Corpora and Language Resources	Small (MB, GB)	Primarily NLP/LT research: ELG, META-SHARE, CLARIN, ELRA, ELDA etc.	Can be easily integrated by connecting the repositories to LDS	Usually very high quality data and thus relevant for LLMs but not as base data
Web Crawls	Very big (TB, PB)	Common Crawl (and OSCAR-processed CC dumps), Internet Archive dumps etc.	Challenge due to their size (hard to transfer, hard to preprocess, hard to store; must be close to the HPC)	Indispensable due to their size and coverage – but: high level of noise, massive need for pre-processing
New, fresh data from industry and other organisations	Arbitrary size, ideally as large as possible	Publishing houses, media companies, libraries, call centres, broadcasters etc.; also: Media Data Space	Can be easily integrated by connecting these organisations to LDS	Especially high quality data or domain-specific data or data covering specific languages and thus highly relevant for LLMs

Alliance for Language Technologies EDIC (ALT-EDIC)

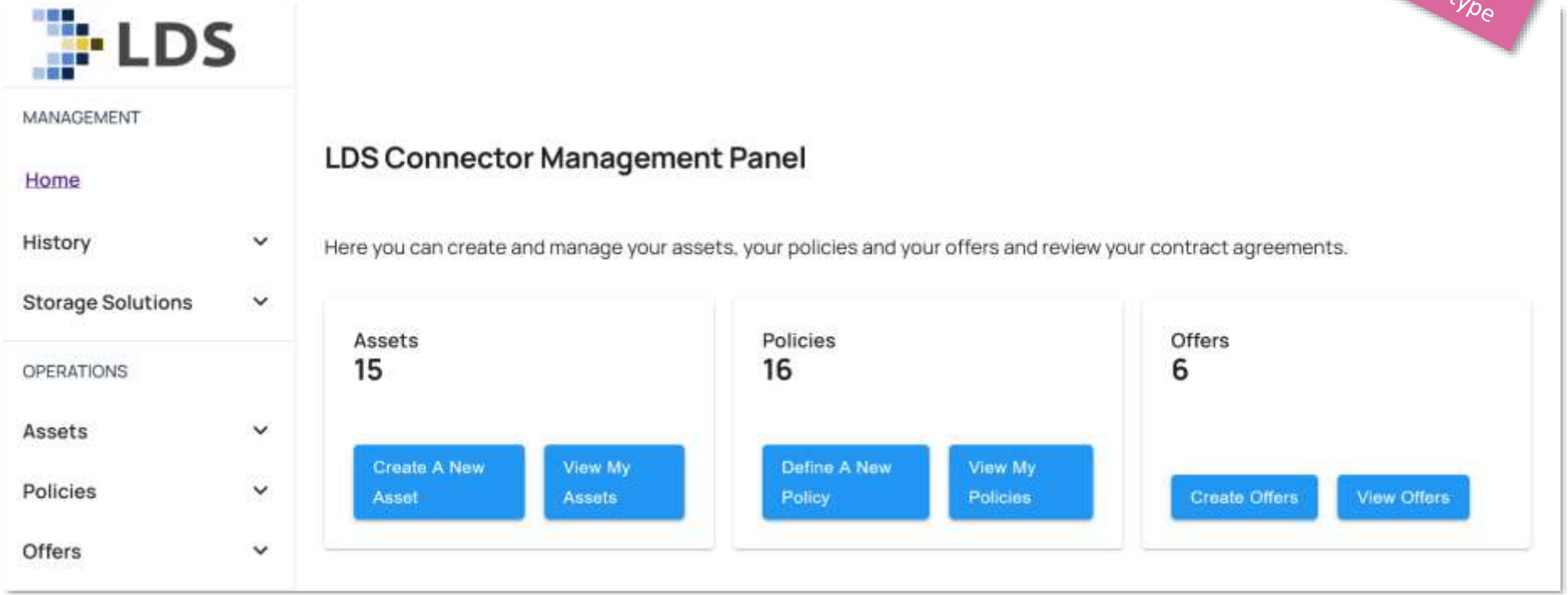
- European Digital Infrastructure Consortium (EDIC): a new legal entity type in the EU
- Two EDICs have recently been established including the ALT-EDIC (not-for-profit association in France)
- Coordinated by the French Ministry of Culture
- Collaboration between: ALT-EDIC Working Group, EC, LDS
- ALT-EDIC action plan will concentrate on:
  - 1. Data
  - 2. Existing language models
  - 3. New language models
  - 4. Evaluation, certification, normalization
  - 5. Ecosystem
  - 6. EDIC implementation
- We expect many synergies between LDS, ALT-EDIC, DSSC, Simpl, other data spaces and other projects!

Long History of Language Data Sharing in the European NLP/CL Community

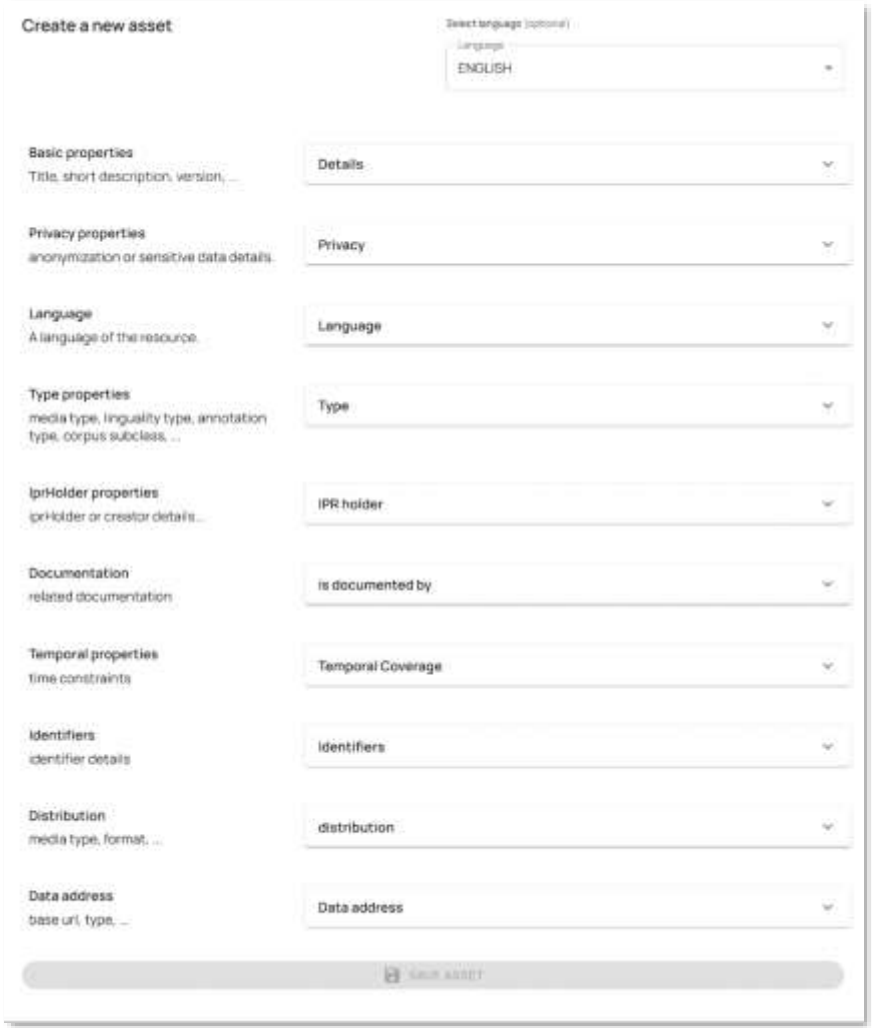



Data
Data provider/consumer A
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
LDS Participants Registry
LDS Vocabulary Hub
LDS Connector
LDS Connector
Data
Data provider/consumer C
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
LDS Catalogue
Data
Data provider/consumer B
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
LDS Connector
LDS Connector
Data
Data provider/consumer D
Identity and Access Management
Assets & Policies (Metadata)
Offers & Contracts
Crawled Offers
Clearing House
Metadata transfer
Metadata & Data transfer

Current LDS Prototype

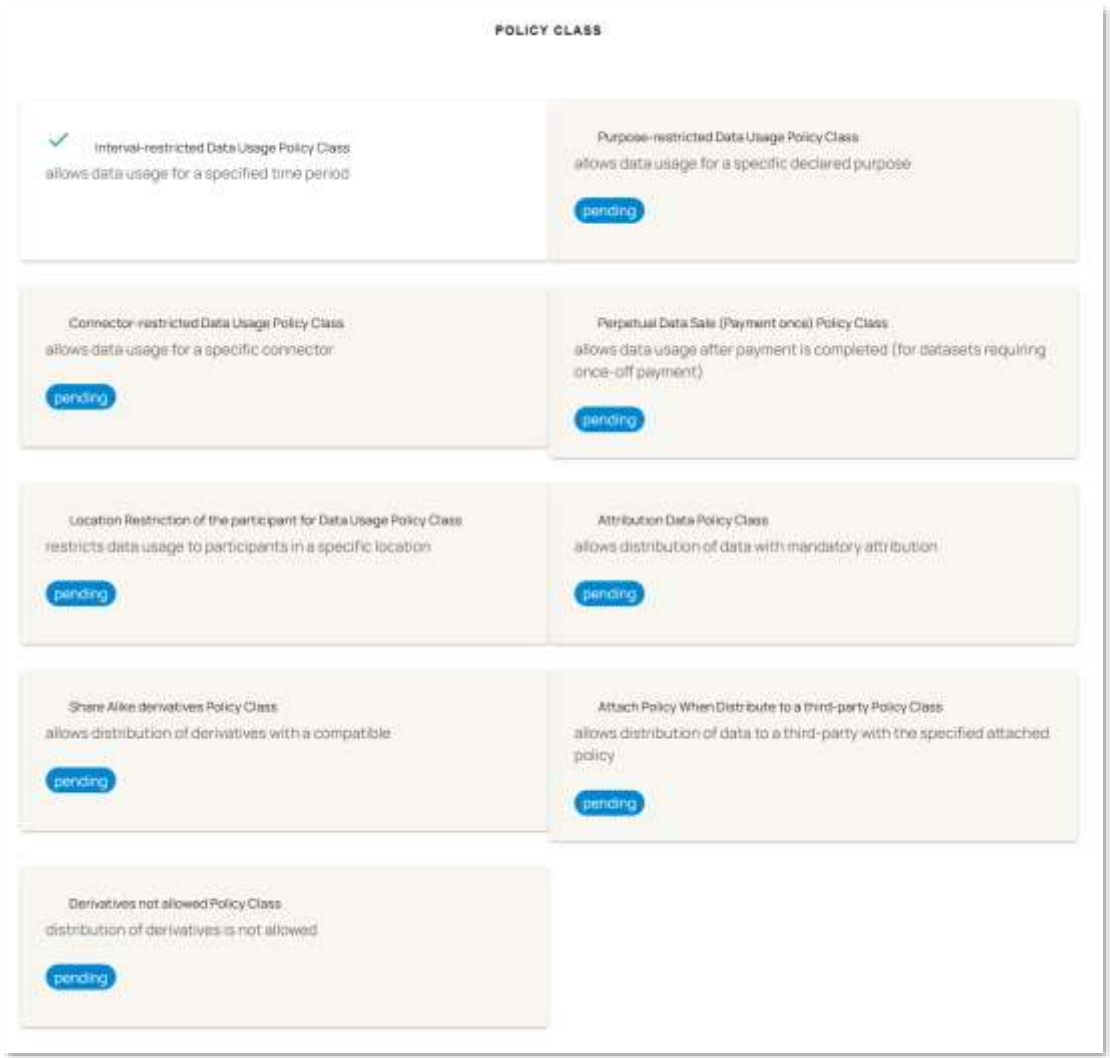


Create new data asset



Current LDS Prototype

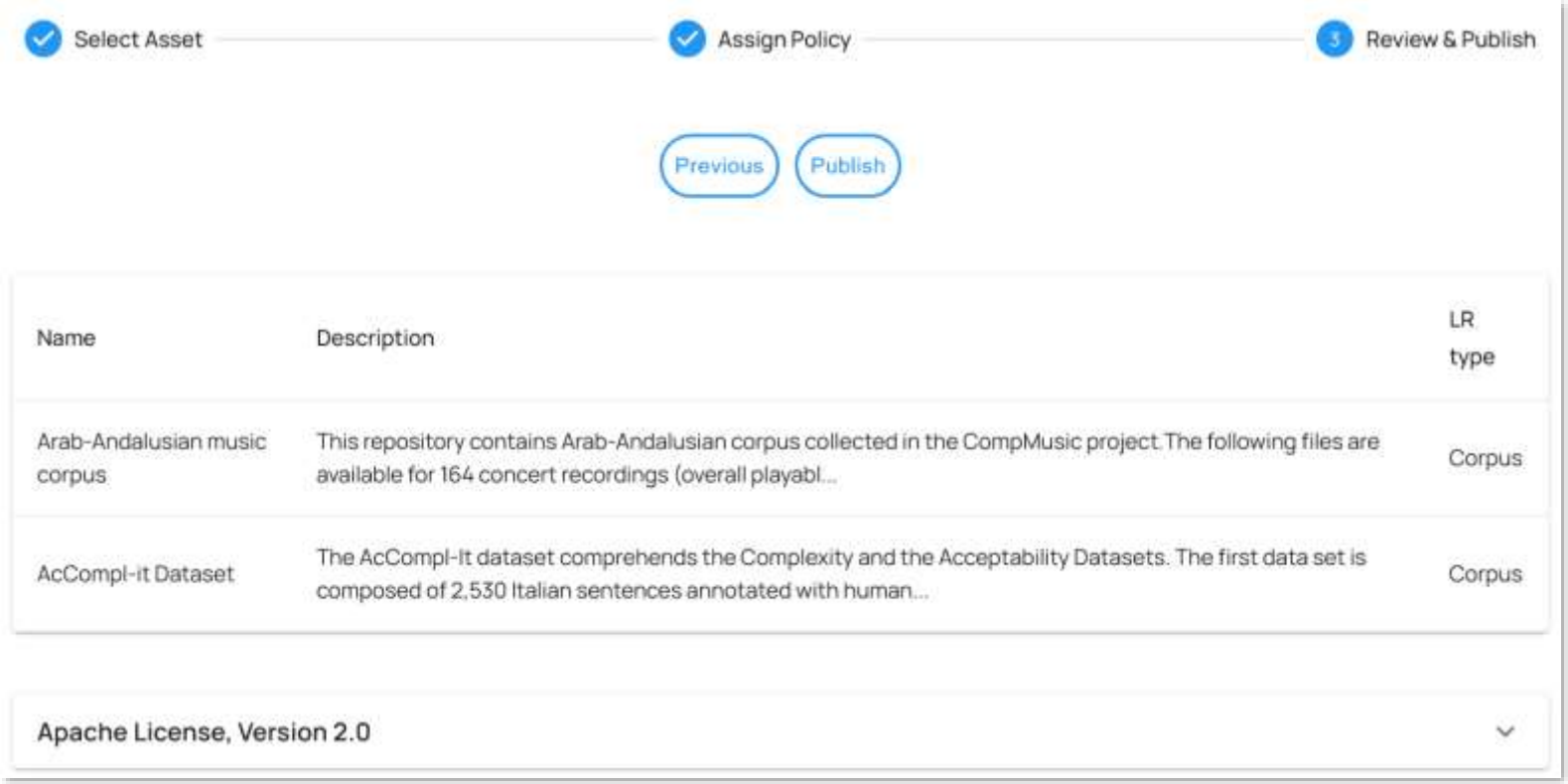
Adjust and attach policy

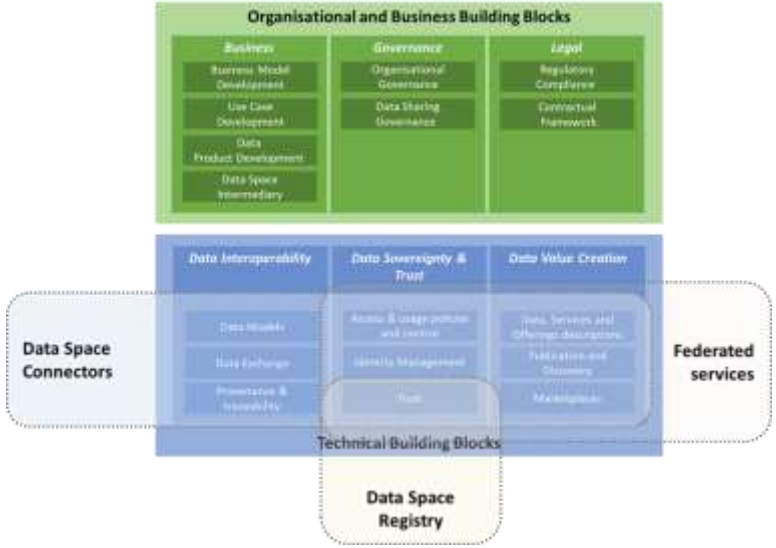


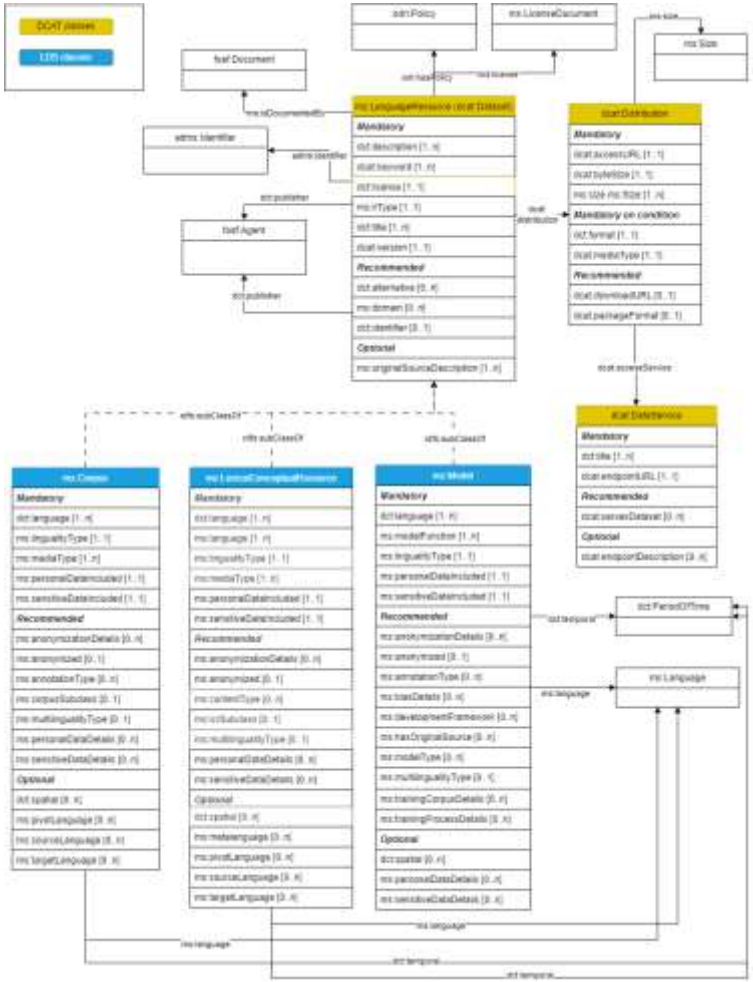
Current LDS Prototype

Create and publish offer

Current LDS Prototype







Built on Existing Solutions

- Following DSSC (see above)
- Eclipse Data Space Components (EDC)
- DCAT-AP, Language DCAT-AP (see right), ODRL
- Mappers from existing platforms

European Language Grid

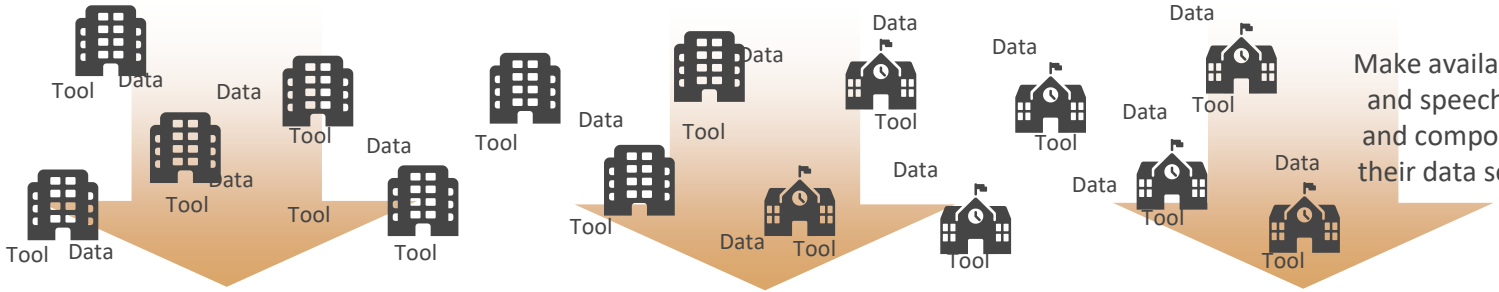
EUROPEAN LANGUAGE GRID

Kick-off meeting, 22/23 January 2019

Objectives (Selection)

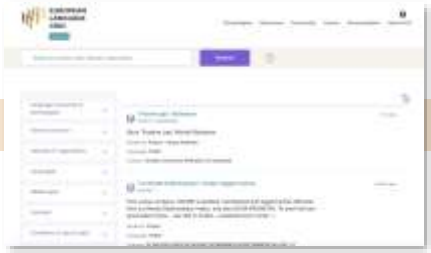
1. Establish the ELG as the primary LT platform and marketplace in Europe to tackle the fragmentation of the European LT landscape.
2. ELG as a platform for commercial and non-commercial, industry-related LTs (functional and non-functional).
3. Enable the European LT community to upload services and data sets, to deploy them and to connect with, and make use of those resources made available by others.
4. Enable businesses to grow and benefit from scaling up.
5. Unleash enormous potential for innovation.

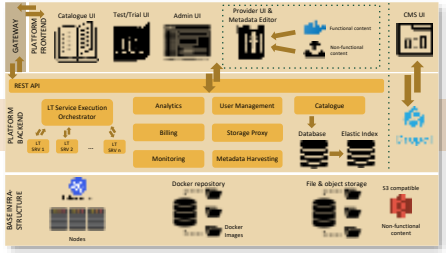
Developers of Language Technologies: Companies, Universities, Research Centres (approx. 1750-2000 organisations in total in Europe)

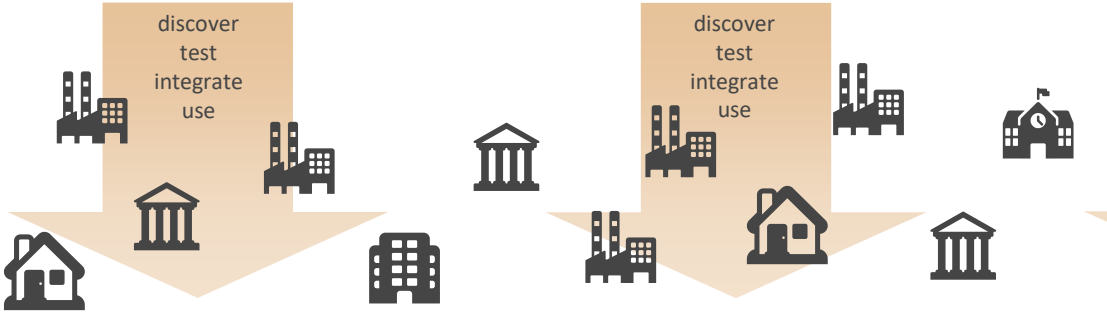


Make available their LT, NLP and speech tools, services and components as well as their data sets, corpora etc.

EUROPEAN LANGUAGE GRID



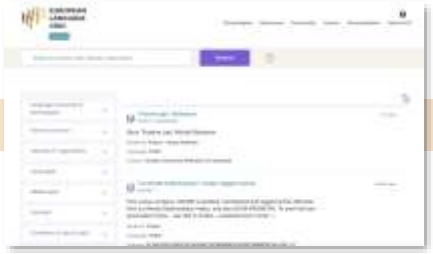


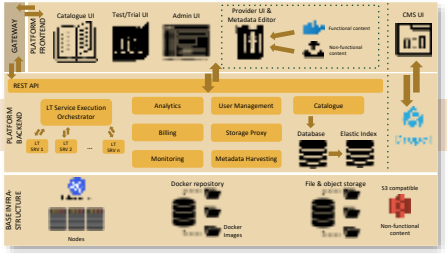


Discover, download, purchase, use (for example, through the ELG Python SDK or through the web UI) language technologies or language resources

Users/Consumers of LTs: Companies, Public Administrations, NGOs, Universities, Research Centres (hundreds of thousands of orgs.)

EUROPEAN LANGUAGE GRID





ELG is a joint tool and resource sharing technology platform as well as marketplace for the whole European LT community (approx. 1750-2000 organisations)

ELG makes all European Language Technologies and Language Resources available in a one-stop-shop.

EUROPEAN LANGUAGE GRID

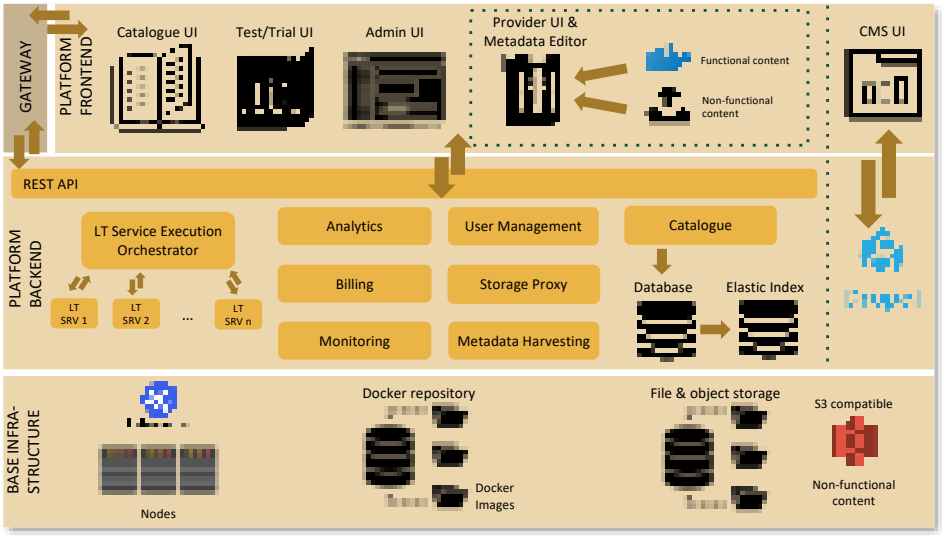
ELG Release 3 (May 2024)

+ ELG / ELE	(10081)
+ ELRC-SHARE	(3099)
+ ELRA Catalogue of Language Resources	(1178)
+ Zenodo	(1107)
+ LINDAT/CLARIAH-CZ	(596)
+ CLARIN.SI	(527)
+ Hugging Face	(385)
+ CLARIN-PL	(351)
+ Quantum Stat Datasets	(262)
+ LREC Shared LRs (ELRA)	(144)
+ META-SHARE/ILSP	(69)
+ META-SHARE/DFKI	(2)

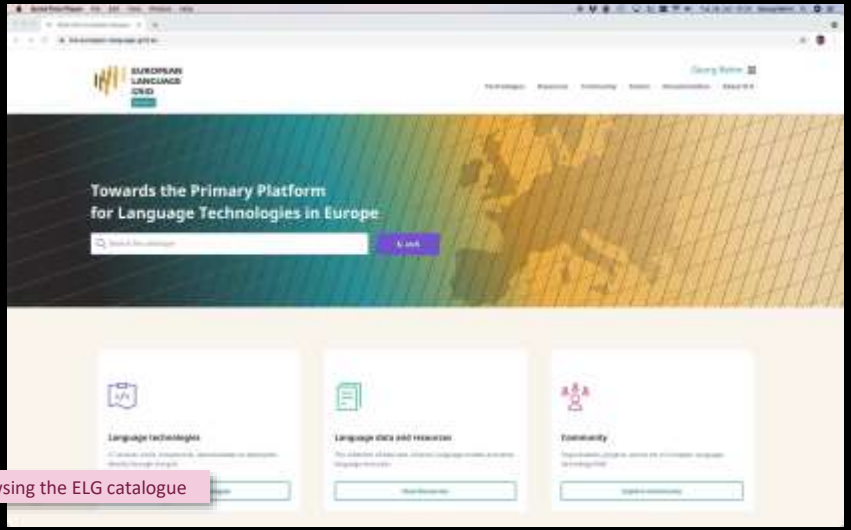
Sources of data sets

17,800+ Resources without organisations

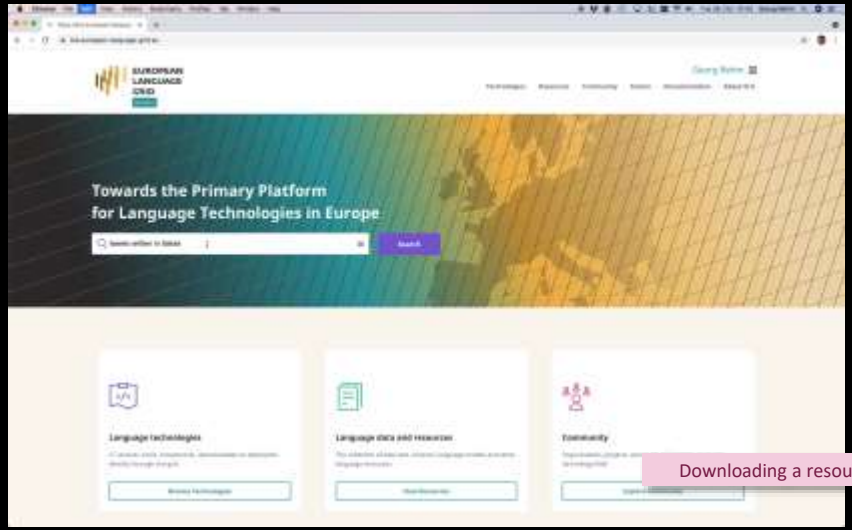
- 8090 corpora and data sets
- 3860 services and tools (1142 services integrated)
- 2830 lexical/conceptual resources
- 513 models, grammars, lang. descriptions
- 2000 organisations (research orgs., companies)



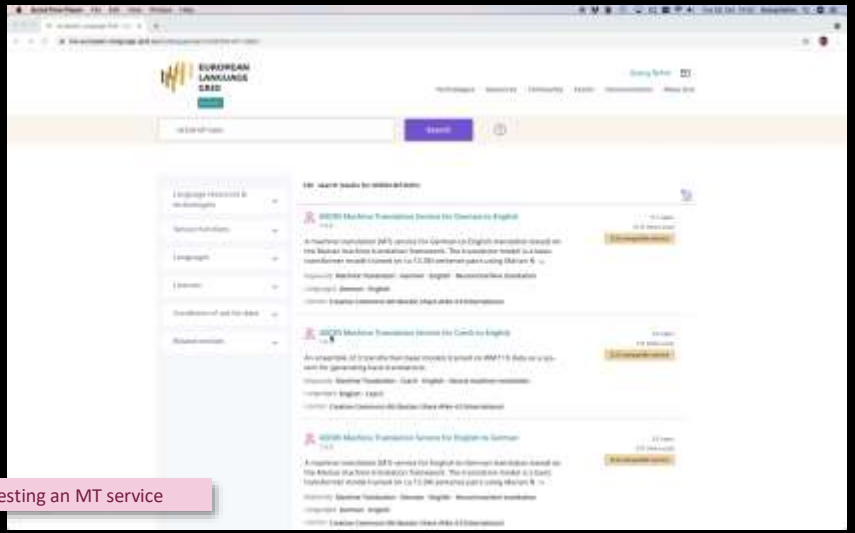
Users can connect to the ELG cloud platform via ELG APIs, remote APIs, ELG GUI, Python SDK, download of containers or source code.



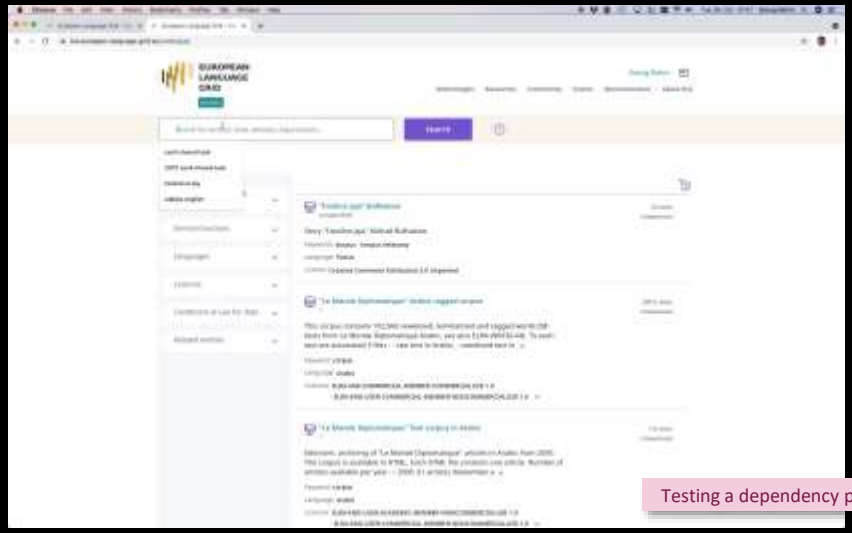
Browsing the ELG catalogue



Downloading a resource

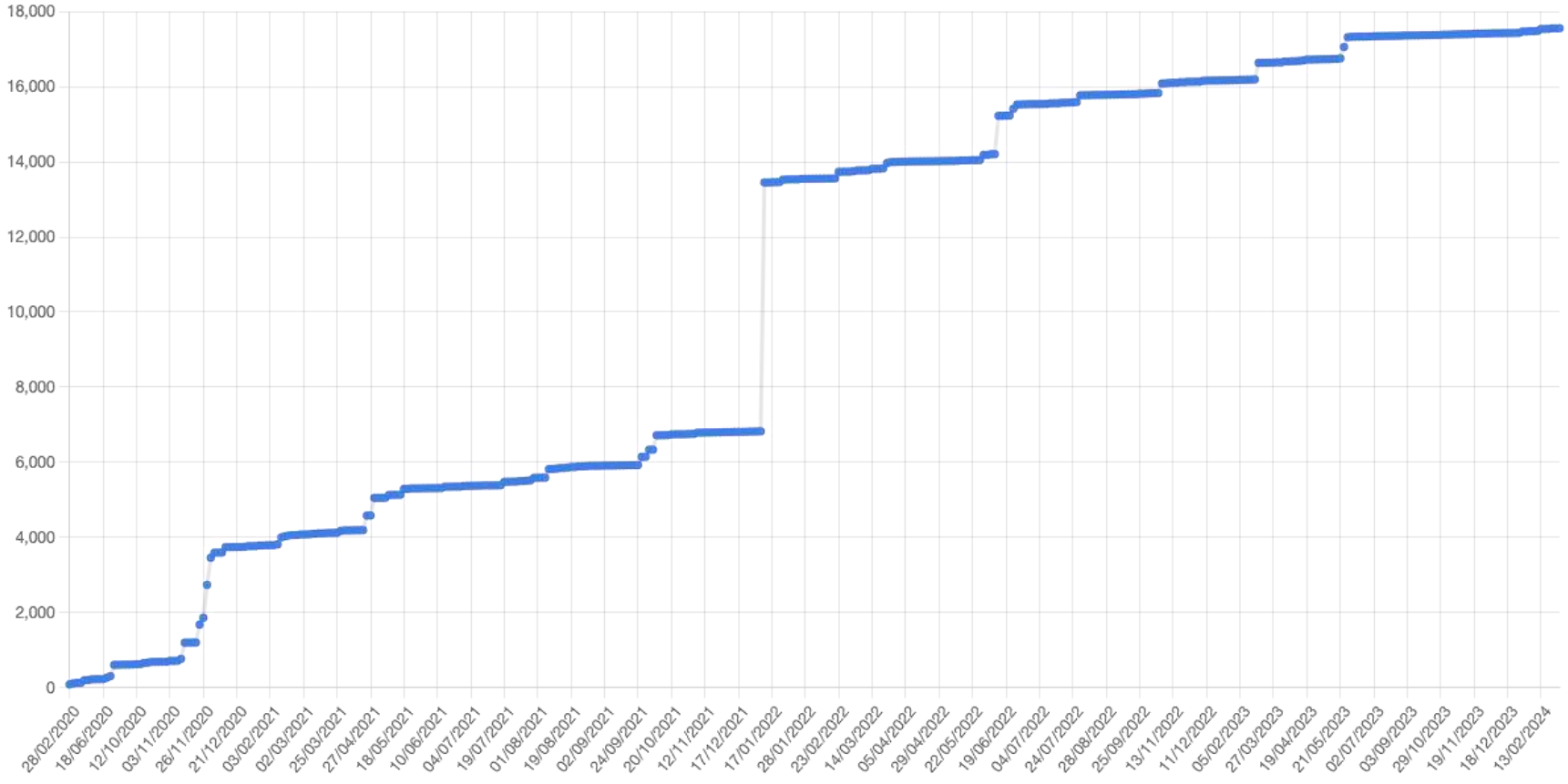


Testing an MT service

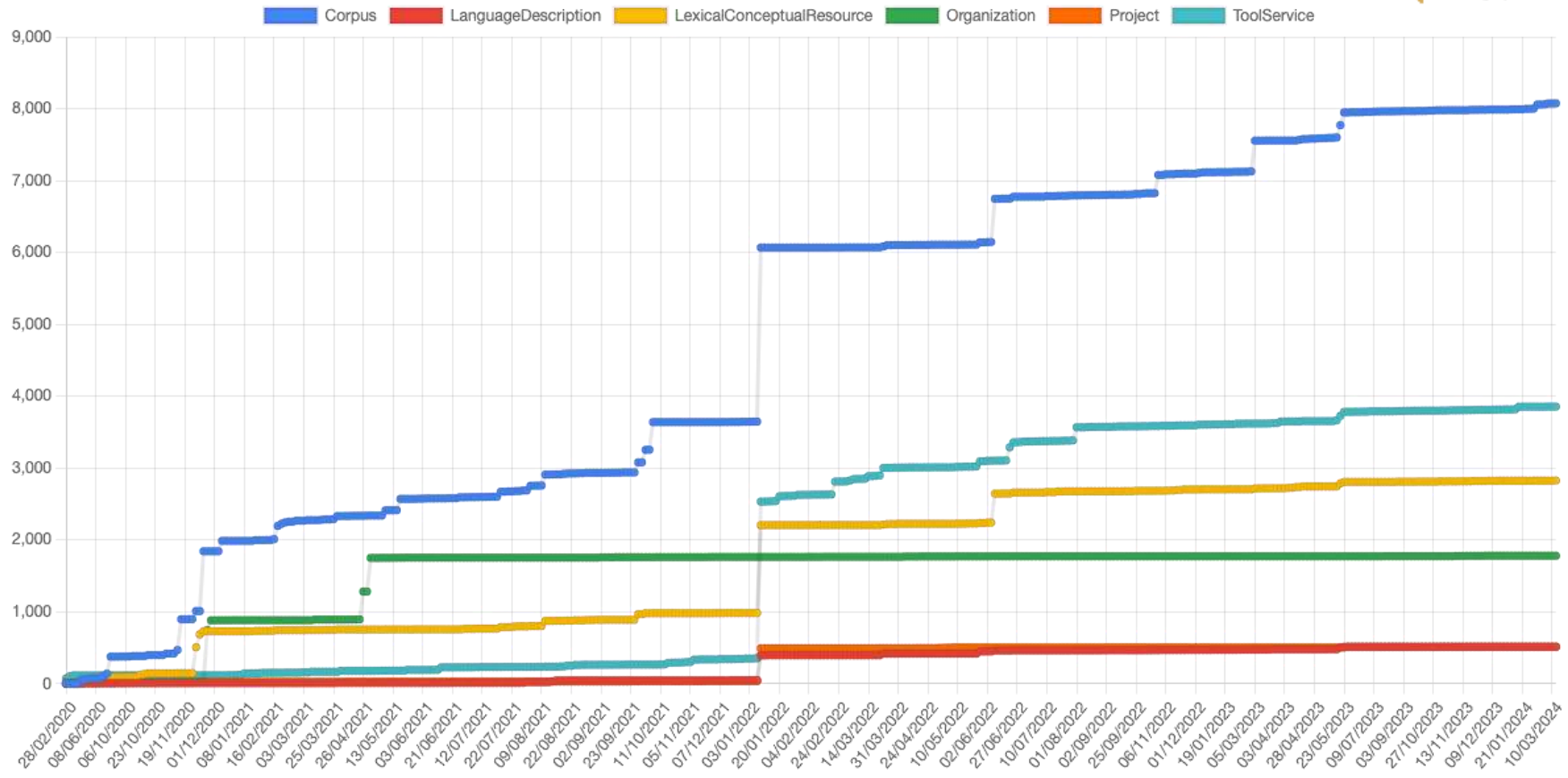


Testing a dependency parser

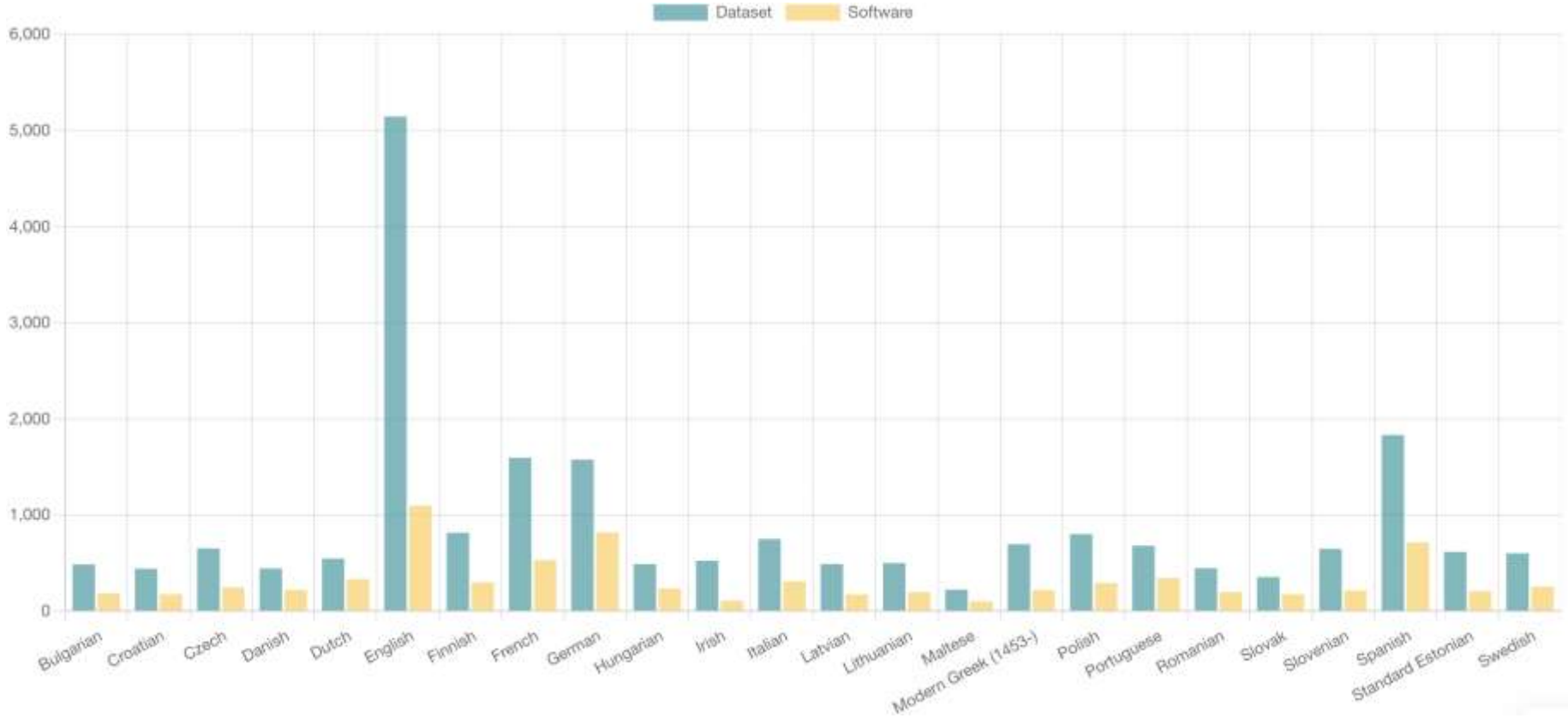
Number of Resources over Time



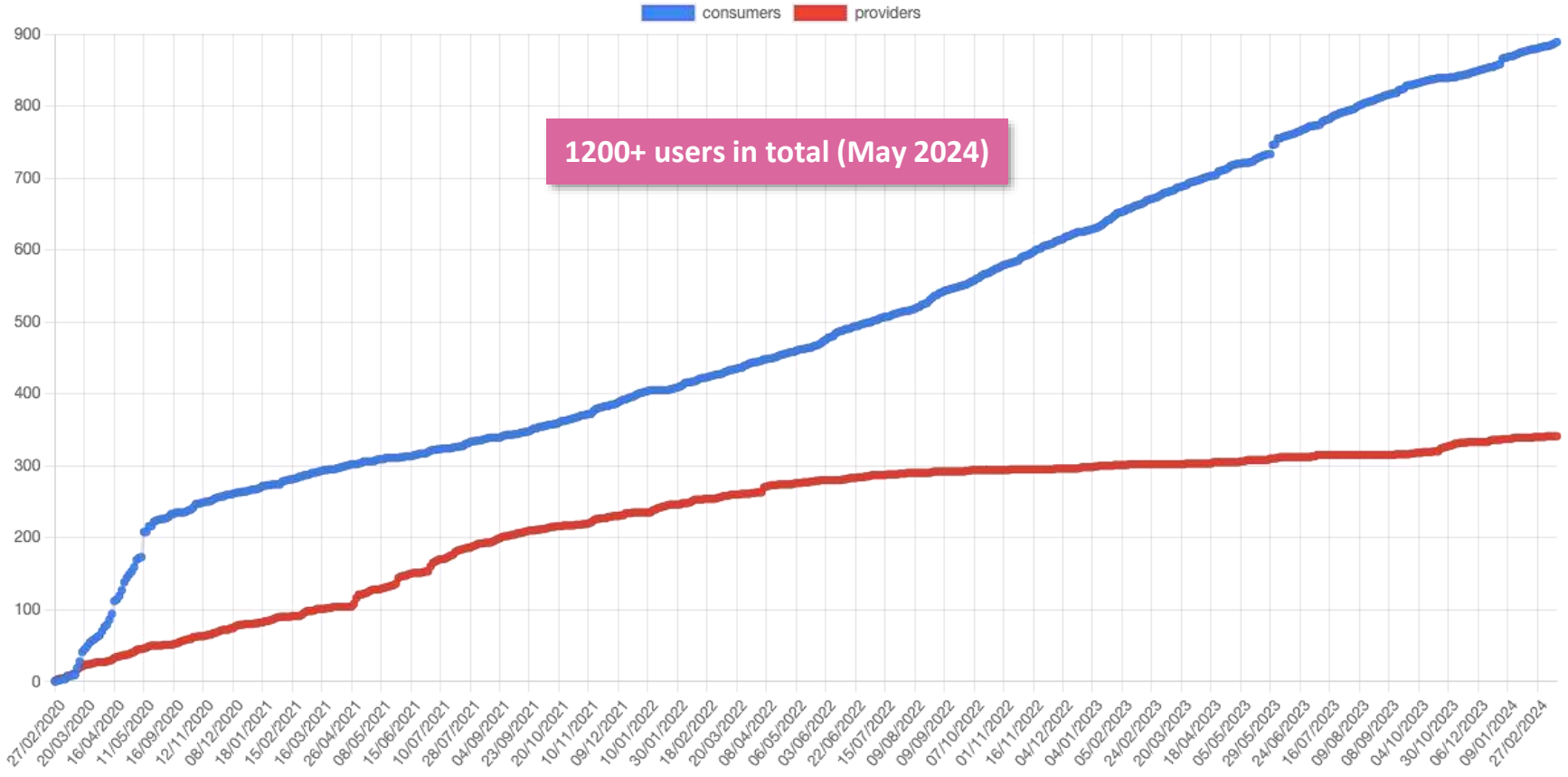
Number of Resources over Time

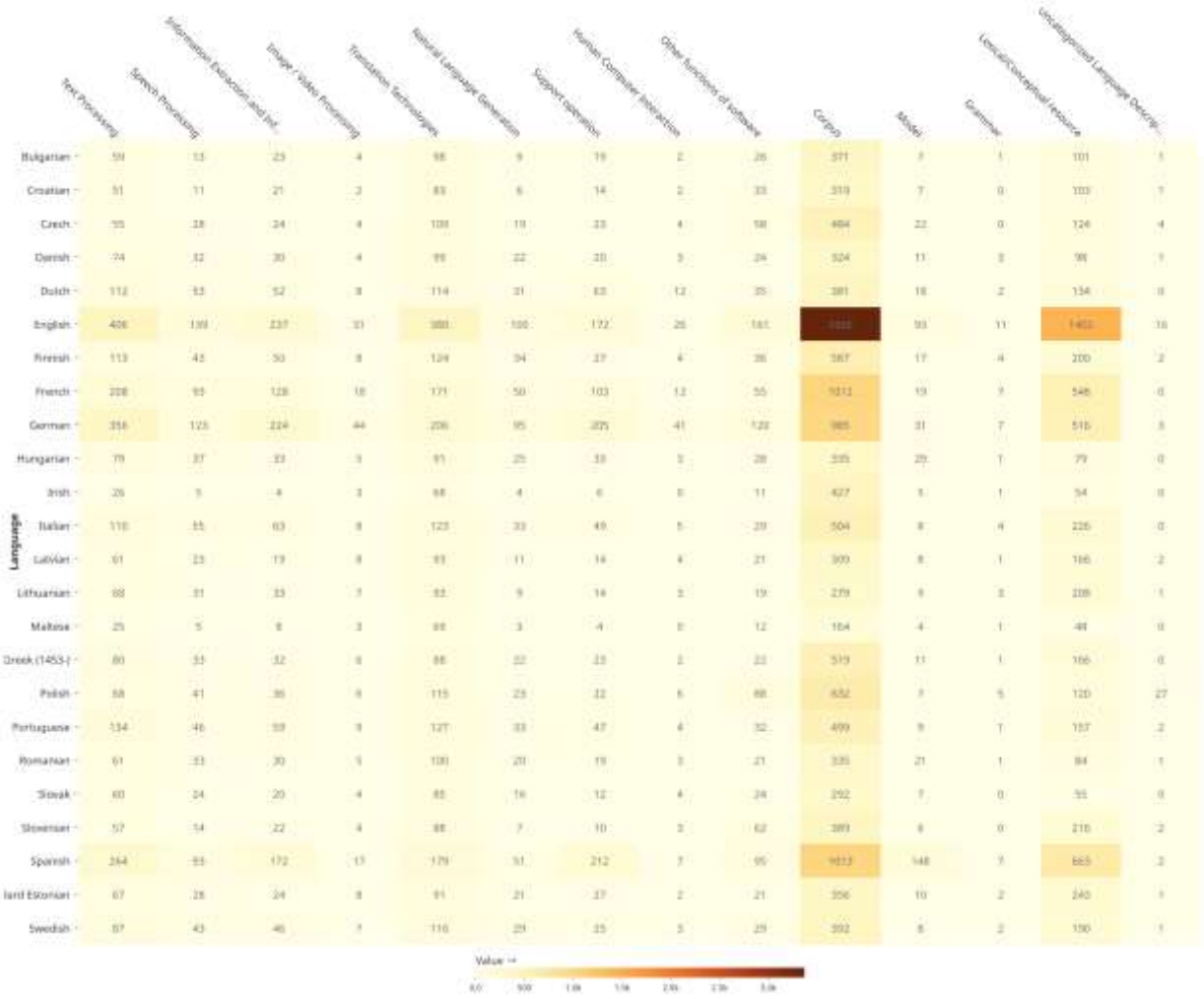


Number of LRTs available in ELG (24 EU languages only)



Number of Users (Consumers of LT, Providers of LT)



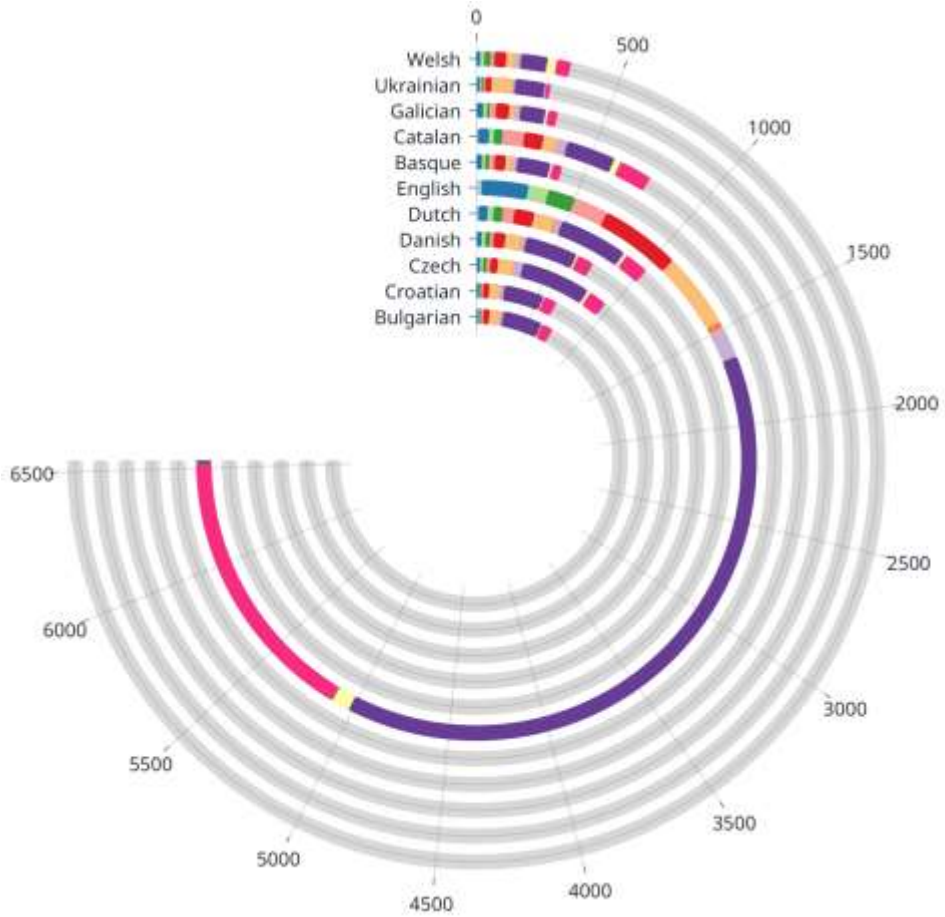


EUROPEAN LANGUAGE EQUALITY

Produced using the ELE Dashboard

https://live.european-language-grid.eu/catalogue/dashboard

LDS



Human Computer Interaction
Information Extraction and Information Retrieval
Natural Language Generation
Speech Processing
Support operation
Text Processing
Translation Technologies
Image / Video Processing
Other functions of software
Corpus
Model
Grammar
Lexical/Conceptual resource
Uncategorized Language Description

EUROPEAN LANGUAGE EQUALITY

Produced using the ELE Dashboard
https://live.european-language-grid.eu/catalogue/dashboard

LDS

European Language Grid and Data Spaces

- Interoperability between ELG and **LDS**: installing the LDS connector within ELG → discoverability of ELG resources through LDS.
- Interoperability between the emerging **DataBri-X** toolbox and ELG using a similar methodology, which will be compliant with IDSA.
- In **OpenGPT-X**, we adapt ELG to Gaia-X, e.g., by serving SDs to federated catalogues.

NFDI for Data Science and AI

https://www.nfdi4datascience.de

NFDI for Data Science and Artificial Intelligence

(max. runtime of project: 10 years)

- NFDI: German National Research Data Infrastructure
- NFDI4DS develops a research data infrastructure for DS and AI
- Foundation: Multiple existing tools and infrastructures – including ELG and LDS
- The NFDI4DS infrastructure will eventually also be connected to and made interoperable with the emerging national NFDI infrastructure
- Research topics: scholarly information processing, knowledge extraction and the construction of research knowledge graphs such as ORKG
- DFKI co-coordinates two Task Areas: Infrastructure & Services, Transfer & Application
- Most recent event: Natural Scientific Language Processing Workshop (NSLP 2024), 27 May 2024, proceedings to be published with Springer (LNAI) soon.
- Upcoming event: 4th Int. Workshop on Scientific Knowledge – Representation, Discovery, and Assessment (Sci-K 2024), 11/12 Nov. 2024, Baltimore.

https://nfdi4ds.github.io/nslp2024/

Natural Scientific Language Processing Workshop
(NSLP 2024), 27 May 2024

Summary and Conclusions

Summary and Conclusions

European Language Grid: One Year After

Georg Rehm[1], Stelios Piperidis[2], Dimitrios Galanis[2], Penny Labropoulou[2], Maria Giagkou[2], Miltos Deligiannis[2], Leon Voukoutis[2], Martin Courtois[1], Julian Moreno-Schneider[1] and Katrin Marheinecke[1]
[1]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
[2]R. C. "Athena", Institute for Language and Speech Processing (ILSP), Athens, Greece
Corresponding author: georg.rehm@dfki.de

Abstract
The European Language Grid (ELG) is a cloud platform for the whole European Language Technology community. While the EU project that developed the platform successfully concluded in June 2022, the ELG initiative has continued. This article provides a description of the current state of ELG in terms of user adoption and number of language resources and technologies available in early 2024. It also provides an overview of the various activities with regard to ELG since the end of the project and since the publication of the ELG book, especially the co-authors' attempt to integrate the ELG platform into various data space initiatives. The article also provides an overview of the Digital Language Equality (DLE) dashboard and the current state of DLE in Europe.

Keywords: LR Infrastructures and Architectures, LR National/International Projects, Tools, Systems

1. Introduction

The European Language Grid (ELG) is a joint cloud platform for the whole European Language Technology community.[1] It provides access to thousands of language resources (LRs) and language technologies (LTs) as well as to related information such as academic and commercial organisations active in the European LT space. The platform is thoroughly documented in the form of several articles (e.g., Labropoulou et al., 2020; Rehm et al., 2021) and, most recently, a comprehensive book that provides detailed technical and conceptual information concerning all aspects of the EU project and the technology platform (Rehm, 2023).

In a development which occurred in parallel with and independently from ELG and its EU project, several different communities have started fleshing out the concept of *data spaces*, which have by now reached a certain level of relevance in Europe. Most importantly, the European Union (EU) kickstarted the development of a number of official, EU-mandated data spaces in various sectors to support the EU Data Strategy. These activities are supported through the EU project Data Spaces Support Centre (DSSC),[2] in which the main initiatives – Gaia-X,[3] IDSA,[4] FIWARE,[5] BDVA[6] – collaborate with other partners and other data space projects to develop the underlying technical and organisational specifications of the European data space ecosystem. In terms of data provisioning, data exchange, data sharing and data markets, it is safe to predict that the emerging DSSC blueprint will gain more and more importance and relevance in the coming years. Because they did not exist back when the ELG project was prepared, one crucial aspect could not have been foreseen to be further investigated in the ELG EU project and that was the relation of ELG to data spaces.

This article describes recent developments around the European Language Grid, especially developments started in 2023 that we have not been able to describe in the ELG book. The remainder of this article is structured as follows. First, Section 2 provides a brief general overview of the ELG platform while Section 3 presents an overview of current statistics in terms of users and resources provided. Section 4 describes the dashboard implemented for monitoring the situation towards Digital Language Equality (DLE) in Europe. Section 5 reports on our attempts to integrate ELG into various data space initiatives including the Common European Language Data Space (Section 5.1), IDSA (Section 5.2), and Gaia-X (Section 5.3); Section 5.4 discusses interoperability across data spaces at the level of metadata. Section 6 describes related work and Section 7 concludes the paper.

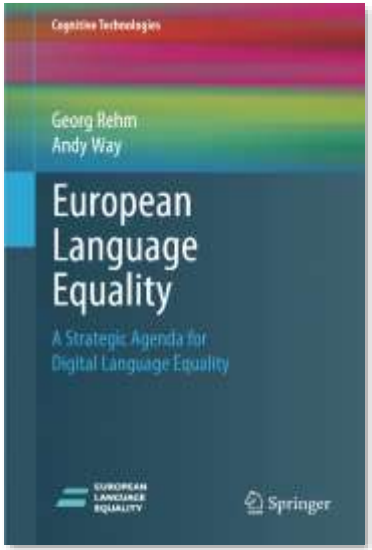
[1]https://www.european-language-grid.eu
[2]https://dssc.eu
[3]https://gaia-x.eu
[4]https://internationaldataspaces.org
[5]https://www.fiware.org
[6]https://www.bdva.eu

2. ELG: A Brief Overview

The European Language Grid (ELG) project was funded by the EU under the Horizon 2020 research

- **European Language Grid**
  - ELG is the most comprehensive NLP/LT platform in Europe.
  - There's still an extremely strong imbalance in terms of technology support of Europe's languages – all languages except English under-resourced!
  - ELG can be used to measure the LT support for Europe's languages.
  - The EU concentrates on data spaces and establishing a data economy in Europe.
  - We're integrating ELG into various data space initiatives.
  - A follow-up project is very much needed to adapt ELG to modern state of the art technologies, especially LLMs.
- **European Language Data Space**
  - Project is in full swing: technical development, promotion, dissemination, governance etc.
  - Collaborations with DSSC, Simpl and ALT-EDIC; European projects, e.g., HPLT, OpenGPT-X, OpenWebSearch; other data spaces, especially Media and Cultural Heritage
  - Adoption of LDS by industry and other organisations → grow the LDS User Group
  - Make available new and fresh language data, esp. from industry, covering all European languages and modalities
- ***Still <u>the</u> key open challenge: interoperability between platforms and infrastructures!***

Thank you!



A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

Prof. Dr. Georg Rehm (DFKI GmbH, Germany)
georg.rehm@dfki.de

06-06-2024 Inria and ECDF Event
https://language-data-space.ec.europa.eu