Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

European Language Grid:
Eine KI-Plattform für flexible Sprachtechnologien

Georg Rehm

DFKI GmbH, Berlin
georg.rehm@dfki.de

DEUTSCHE NATIONAL BIBLIOTHEK **Fachtagung 2021 Maschinelle Verfahren in der Erschließung**

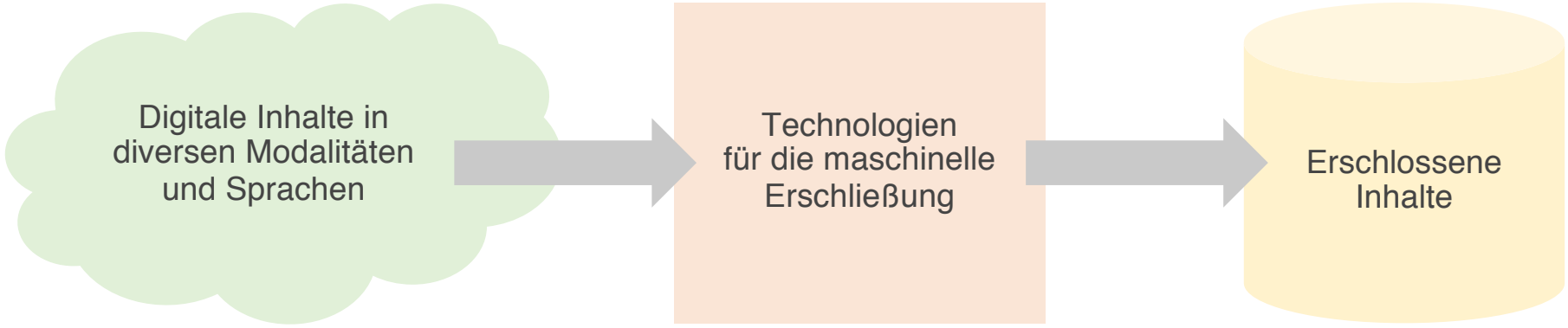
18. November 2021

Inhalt

1. Einleitung: Maschinelle Erschließung
2. QURATOR – Intelligente Kuratierungstechnologien
3. European Language Grid (ELG)
4. European Language Equality (ELE)
5. Zusammenfassung

Maschinelle Erschließung

- Maschinelle Verarbeitung digitaler Inhalte unterschiedlicher Modalitäten
- Primär Text, zunehmend Bild (Zeitungen, Briefe etc.), Fotos, Video, Audio
- Einsatz von NLP (speziell Textanalytik, Text Mining), ASR, OCR, CV etc.
- Herausforderung: Nicht-deutschsprachige sowie mehrsprachige Inhalte



Kuratierungstechnologien

- Smarte KI-basierte Technologien für die Kuratierung von Content mit Fokus auf professionellen Anwendungskontexten
- BMBF-Projekt **DKT** – Digitale Kuratierungstechnologien (2015-2017)
- BMBF-Projekt **QURATOR** – Curation Technologies (2018-2022)
- BMBF-Projekt **PANQURA** – QURATOR für die Pandemie (2021-2022)
- EU-Projekt **Lynx** – Legal Knowledge Graph (2018-2021)

Qurator

Curation Technologies

- **Zehn Teilprojekte – ein Verbundprojekt**
- **Technologieplattform** folgt dem **Baukastenprinzip**
- In Teilprojekten entwickelte Verfahren sind individuell nutzbar – sie bilden die **flexibel kombinierbaren Services** der **QURATOR-Plattform**

DFKI GmbH Kuratierungstechnologien – Flexible KI-Plattform für die adaptive Analyse und kreative Generierung digitaler Inhalte in branchenübergreifenden Kontexten

3pc GmbH Neue Kommunikation Entwicklung generischer Kuratierungstechnologien für interaktives Storytelling

Ada Health GmbH Werkzeuge und Technologie für die Kuratierung biomedizinischen Wissens

ART+COM AG
Kuratierungs-Tools für interaktive Multimedia-Inhalte

Condat AG
Kuratierungstechnologien für den TV/Medien Bereich

Fraunhofer Gesellschaft – FOKUS
Corporate Smart Insights (CSI)

Semtation GmbH Business Process Modelling – intelligente Navigation durch Wissensräume

Stiftung Preußischer Kulturbesitz, Staatsbibliothek zu Berlin
Automatisierte Kuratierungstechnologien für das digitale kulturelle Erbe

uberMetrics Technologies GmbH Kuratierungstechnologien für das Monitoring von Online-Inhalten und die Risikobeobachtung

Wikimedia Deutschland e.V.
Qualitätssteigerung in Wikidata

QURATOR Auftaktveranstaltung, 19. November 2018

WACHSTUMSKERNE
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Branchenlösungen

MEDIZIN **Medical Content Curator**

KULTUR **Smarte Exponate**

MEDIEN **Medien-Kurator**

INDUSTRIE **Intelligente Navigation**

INDUSTRIE **Risiko-Monitoring**

KULTUR **Next Reality Storytelling**

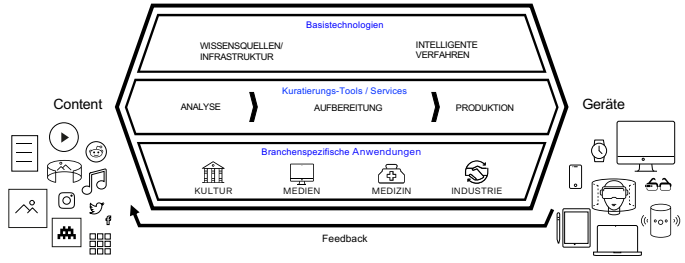
Showcases

Content Curation Engine

Corporate Smart Insights

Semantische Anreicherung

Fokus auf Qualität in Wikidata



Technologieplattform

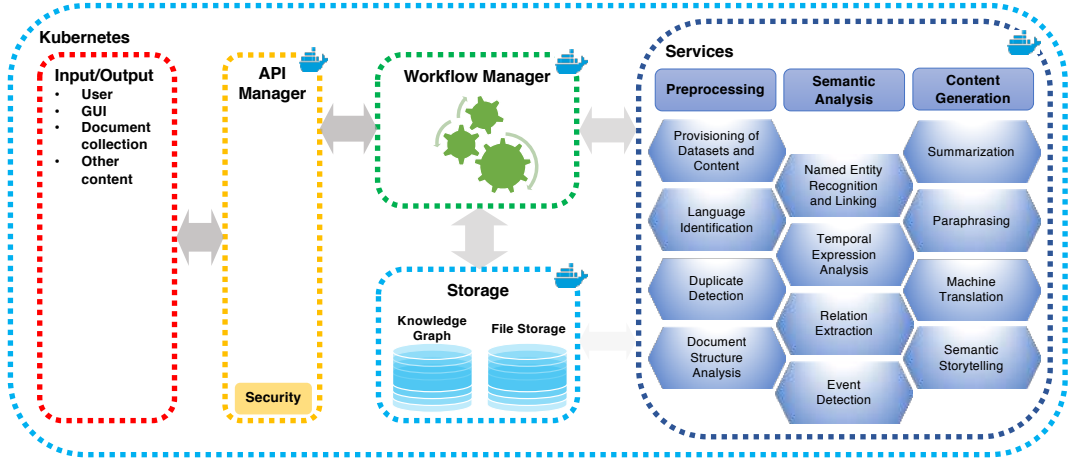
Qurator
Curation Technologies

Technologieplattform

- Flexible, robuste, skalierbare QURATOR-Plattform, gefüllt mit einer Vielzahl von QURATOR-Services
- Flexible Orchestrierung von Workflows
- Verfahren zur Erkennung unterschiedlicher Klassen von Inhalten, um sie in spezifische Workflows einzuspeisen.
- Modulare Inhalte einfacher in neuen Produktionen, Anwendungen, Nutzungskontexten aggregieren.
- Interoperabilität durch generische APIs (SaaS) und wenig Integrationsaufwand.

Vorgehensweise

- QURATOR-Plattform wird mit Services gefüllt.
- QURATOR hat 100+ Services entwickelt
- Basiert auf Kubernetes und Containern (Docker).



Beispiel: Semantische Relationen I

- Identifikation von Diskursrelationen zwischen Textsegmenten
- Diskursrelationen aus der Penn Discourse Tree Bank (PDTB2)
- Modell: Siamese BERT

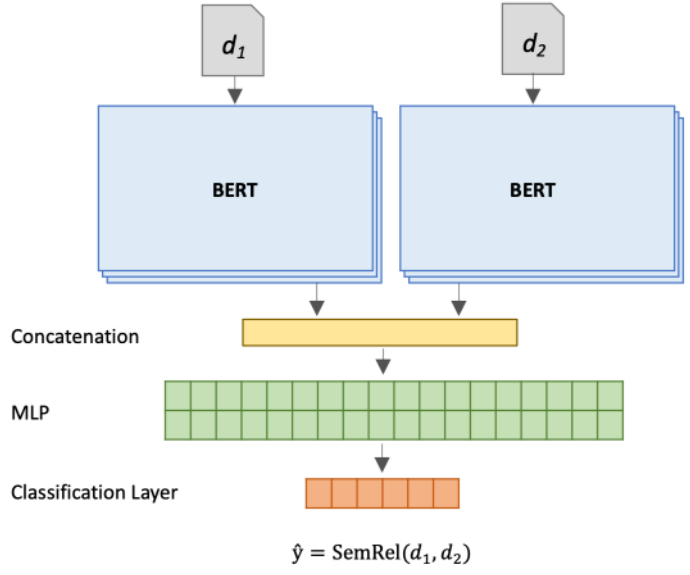


Table 1: Results of training multi-class prediction based on BERT in a 80-20 train-test-split.

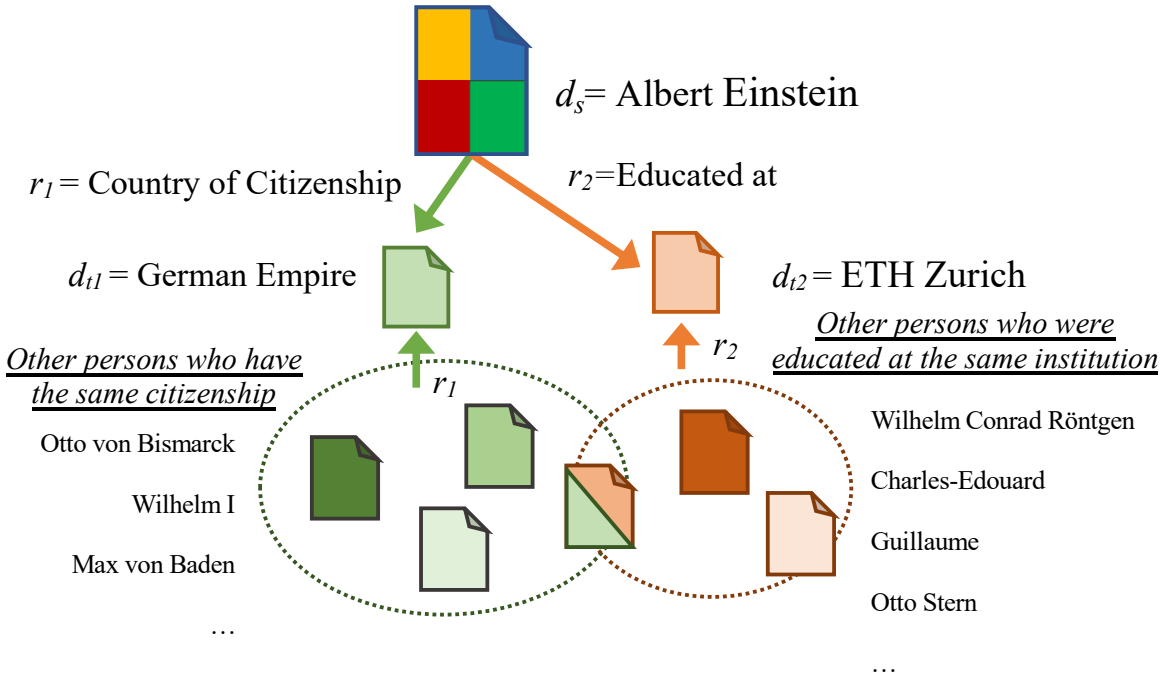
PDTB Relation	Precision	Recall	F1-score	Support
Comparison	0.50	0.47	0.48	1598
Contingency	0.38	0.65	0.48	1582
Expansion	0.50	**0.79**	**0.61**	2993
Temporal	**0.51**	0.55	0.53	869
None	0.49	0.73	0.59	1078
Micro avg	0.47	0.67	0.55	8120
Macro avg	0.48	0.64	0.54	8120

Beispiel: Semantische Relationen II

- Identifizierung von semantischen Relationen zwischen Wikipedia-Artikeln
- Relationen: Wikidata-Propertys, die Wikipedia-Artikel verknüpfen (z.B., *country of citizenship*, *different from*, *educated at*, *has effect*, *opposite of*, *facet of*, …)

Table 4: Results as micro avg. F1-score with standard deviation in 4-fold cross-validation for the tested methods.

Methods	F1-score	Std.
AvgGloVe [57]	0.875	± 0.0036
Paragraph Vectors [46]	0.845	± 0.0019
Siamese BERT [35, 60]	0.870	± 0.0067
Siamese XLNet [60, 74]	0.864	± 0.0096
BERT [35]	**0.933**	± 0.0039
XLNet [74]	0.926	± 0.0016



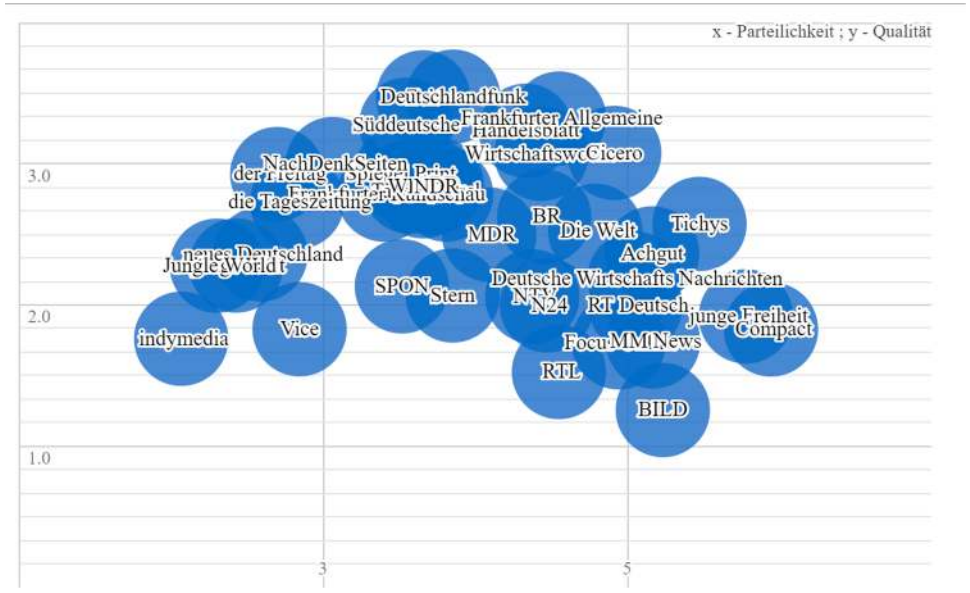
Beispiel: Politische Orientierung

Qurator
Curation Technologies

Ermittlung der politischen Orientierung von z.B. Nachrichtenartikeln (von linksextrem bis rechtsextrem etc.)

Model	BOW	TF-IDF	BERT
Logistic Regression	0.3132	0.2621	0.3389
Naive Bayes	**0.4243**	0.2234	0.3637
Random Forest	0.4007	**0.4303**	**0.3836**
EasyEnsemble	0.4197	0.4070	0.3432

Class	Precision	Recall	F_1	Support
Far-left	0.59	0.40	0.48	215
Centre-left	0.34	0.38	0.36	1,159
Centre	0.31	0.23	0.27	1,349
Centre-right	0.51	0.55	0.53	1,754
Far-right	0.46	0.58	0.51	671
Total	**0.44**	**0.43**	**0.43**	**5,148**



x Achse – Parteilichkeit (kleiner 3.5 links, 3.5 bis 4.5 relativ neutral, größer 4.5 rechts)
y Achse – Qualität (1 Clickbait – 5 Komplex)

https://medienkompass.org/deutsche-medienlandschaft/

Beispiel: Fact-Checking

Qurator
Curation Technologies

The sentences, containing potentially dubious claims, are highlighted accordingly:

- False
- Mostly False
- Mixture
- MostlyTrue
- True
- Missing info/Unclear

The latest CDC #COVIDView report shows that the hospitalization rates for adults are similar or higher than those seen at comparable points during recent flu seasons while those for children are much lower. For younger people, seasonal flu is in many cases a deadlier virus than COVID-19. More and more studies show that kids are actually stoppers of the disease and they don't get it and transmit it themselves. Prevalence of asymptomatic infections in children correlates with the overall incidence of COVID-19 in the local population, new JAMA Pediatrics study finds. Children ages 5 to 9 are not affected by the coronavirus. That is why no country in the world has started vaccinating children. Children are almost immune from Covid-19. However, COVID-19 is associated with additional complications like blood clots and multisystem inflammatory syndrome in children. That is why the U.S. CDC encourages the use of a COVID-19 flu shot on them.

SoNAR (IDH)
Interfaces to Data for Historical Social Network Analysis and Research

Qurator
Curation Technologies

Qualität in der Inhaltserschließung

Lydia Pintscher, Peter Bourgonje, Julián Moreno Schneider, Malte Ostendorff, Georg Rehm

Wissensbasen für die automatische Erschließung und ihre Qualität am Beispiel von Wikidata

1 Einführung

Wikidata[1] ist eine freie Wissensbasis, die allgemeine Daten über die Welt zur Verfügung stellt. Sie wird von Wikimedia entwickelt und betrieben, wie auch das Schwesterprojekt Wikipedia. Die Daten in Wikidata werden von einer großen Community von Freiwilligen gesammelt und gepflegt, wobei die Daten sowie die zugrundeliegende Ontologie von vielen Projekten, Institutionen und Firmen als Basis für Applikationen und Visualisierungen, aber auch für das Training von maschinellen Lernverfahren genutzt werden. Wikidata nutzt MediaWiki[2] und die Erweiterung Wikibase[3] als technische Grundlage der kollaborativen Arbeit an einer Wissensbasis, die verlinkte offene Daten für Menschen und Maschinen zugänglich macht.

Ende 2020 beschreibt Wikidata über 90 Millionen Entitäten (siehe Abb. 1) unter Verwendung von über 8 000 Eigenschaften, womit insgesamt mehr als 1,15 Milliarden Aussagen über die beschriebenen Entitäten getroffen werden. Die Datenobjekte dieser Entitäten sind mit äquivalenten Einträgen in mehr als 5 500 externen Datenbanken, Katalogen und Webseiten verknüpft, was Wikidata zu einem der zentralen Knotenpunkte des Linked Data Web macht. Mehr als 11 500 aktiv Editierende[4] (siehe Abb. 2) tragen neue Daten in die Wissensbasis ein und pflegen sie. Diese sind in Wiki-Projekten organisiert, die jeweils bestimmte Themenbereiche oder Aufgabengebiete adressieren. Die Daten werden in mehr als der Hälfte der Inhaltsseiten in den Wikimedia-Projekten genutzt und unter anderem mehr als 6,5 Millionen Mal am Tag über den SPARQL-Endpoint[5] abgefragt, um sie in externe Applikationen und Visualisierungen einzubinden.

1 https://www.wikidata.org (17.12.2020).
2 https://www.mediawiki.org (17.12.2020).
3 https://wikiba.se (17.12.2020).
4 Aktiv Editierende sind Editierende, die in den letzten 30 Tagen fünf oder mehr Änderungen vorgenommen haben.
5 https://query.wikidata.org (17.12.2020).

https://doi.org/10.1515/9783110691597-005

Clemens Neudecker, Karolina Zaczynska, Konstantin Baierer, Georg Rehm, Mike Gerber, Julián Moreno Schneider

Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten

1 Einleitung

Durch die systematische Digitalisierung der Bestände in Bibliotheken und Archiven hat die Verfügbarkeit von Bilddigitalisaten historischer Dokumente rasant zugenommen. Das hat zunächst konservatorische Gründe: Digitalisierte Dokumente lassen sich praktisch nach Belieben in hoher Qualität vervielfältigen und sichern. Darüber hinaus lässt sich mit einer digitalisierten Sammlung eine wesentlich höhere Reichweite erzielen, als das mit dem Präsenzbestand allein jemals möglich wäre. Mit der zunehmenden Verfügbarkeit digitaler Bibliotheks- und Archivbestände steigen jedoch auch die Ansprüche an deren Präsentation und Nachnutzbarkeit. Neben der Suche auf Basis bibliothekarischer Metadaten erwarten Nutzer:innen auch, dass sie die Inhalte von Dokumenten durchsuchen können.

Im wissenschaftlichen Bereich werden mit maschinellen, quantitativen Analysen von Textmaterial große Erwartungen an neue Möglichkeiten für die Forschung verbunden. Neben der Bilddigitalisierung wird daher immer häufiger auch eine Erfassung des Volltextes gefordert. Diese kann entweder manuell durch Transkription oder automatisiert mit Methoden der *Optical Character Recognition* (OCR) geschehen (Engl et al. 2020). Der manuellen Erfassung wird im Allgemeinen eine höhere Qualität der Zeichengenauigkeit zugeschrieben. Im Bereich der Massendigitalisierung fällt die Wahl aus Kostengründen jedoch meist auf automatische OCR-Verfahren.

Die Einrichtung eines massentauglichen und im Ergebnis qualitativ hochwertigen OCR-Workflows stellt Bibliotheken und Archive vor hohe technische Herausforderungen, weshalb dieser Arbeitsschritt häufig an dienstleistende Unternehmen ausgelagert wird. Bedingt durch die Richtlinien für die Vergabepraxis und fehlende oder mangelhafte Richtlinien der digitalisierenden Einrichtungen bzw. entsprechender Förderinstrumente führt dies jedoch zu einem hohen Grad an Heterogenität der Digitalisierungs- bzw. Textqualität sowie des Umfangs der strukturellen und semantischen Auszeichnungen. Diese Heterogenität erschwert die Nachnutzung durch die Forschung, die neben einheitlichen

https://doi.org/10.1515/9783110691597-009

Sina Menzel, Hannes Schnaitter, Josefine Zinck, Vivien Petras, Clemens Neudecker, Kai Labusch, Elena Leitner, Georg Rehm

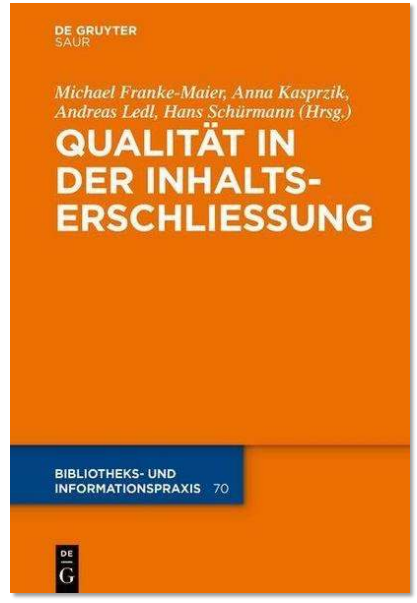
Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten

1 Einführung

Named Entities[1] (benannte Entitäten) – wie Personen, Organisationen, Orte, Ereignisse und Werke – sind wichtige inhaltstragende Komponenten eines Dokuments und sind daher maßgeblich für eine gute inhaltliche Erschließung. Die Erkennung von Named Entities, deren Auszeichnung (Annotation) und Verfügbarmachung für die Suche sind wichtige Instrumente, um Anwendungen wie z. B. die inhaltliche oder semantische Suche in Texten, dokumentübergreifende Kontextualisierung oder das automatische Textzusammenfassen zu verbessern. Inhaltlich präzise und nachhaltig erschlossen werden die erkannten Named Entities eines Dokuments allerdings erst, wenn sie mit einer oder mehreren Quellen verknüpft werden (Grundprinzip von Linked Data, Berners-Lee 2006), die die Entität eindeutig identifizieren und gegenüber gleichlautenden Entitäten disambiguieren (vergleiche z. B. *Berlin* als Hauptstadt Deutschlands mit dem Komponisten *Irving Berlin*). Dazu wird die im Dokument erkannte Entität mit dem Entitätseintrag einer Normdatei oder einer anderen zuvor festgelegten Wissensbasis (z. B. Gazetteer für geografische Entitäten) verknüpft, gewöhnlich über den persistenten Identifikator der jeweiligen Wissensbasis oder Normdatei. Durch die Verknüpfung mit einer Normdatei erfolgt nicht nur die Disambiguierung und Identifikation der Entität, sondern es wird dadurch auch Interoperabilität zu anderen Systemen hergestellt, in denen die gleiche Normdatei benutzt wird, z. B. die Suche nach der Hauptstadt Berlin in verschiedenen Datenbanken bzw. Portalen. Die Entitätenverknüpfung (*Named Entity Linking*, NEL) hat zudem den Vorteil, dass die Normdateien oftmals Relationen zwischen Entitäten enthalten, sodass Dokumente, in denen Named Entities erkannt wurden, zusätzlich auch im Kontext einer größeren Netzwerkstruktur von Entitäten verortet und suchbar gemacht werden können (z. B. die Ausweitung einer Suche

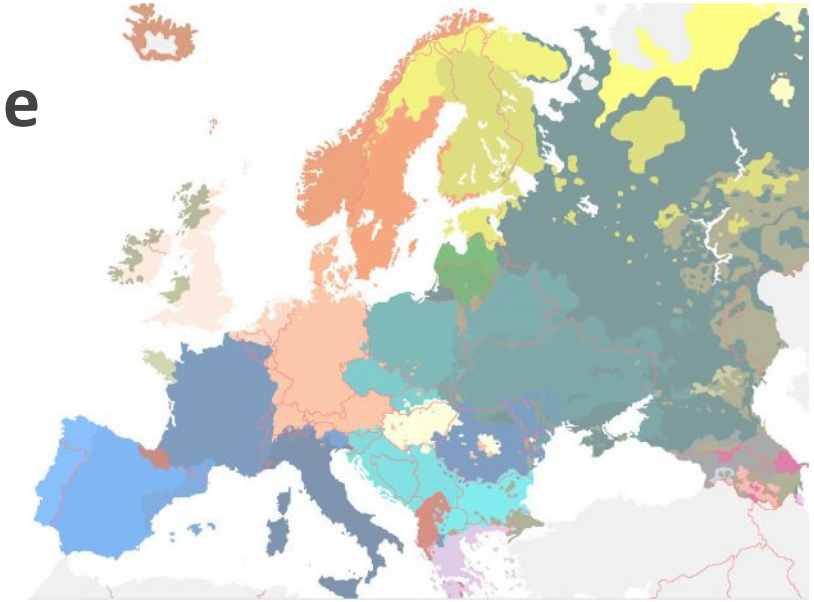
1 Wir verwenden in diesem Beitrag nicht den deutsch- (*Entitys*), sondern den englischsprachigen Plural (*Entities*).

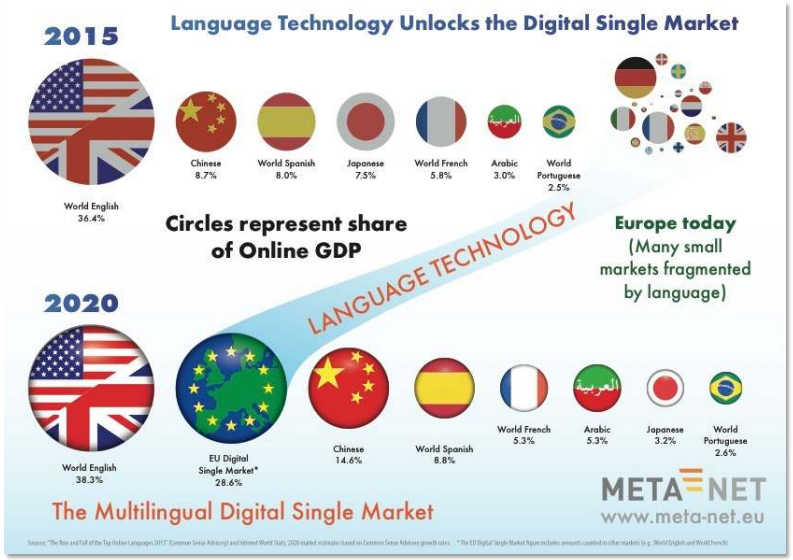
https://doi.org/10.1515/9783110691597-012



Point of Departure: Multilingualism in Europe

- Multilingualism is at the heart of the European idea
- 24 official EU languages – they all have the same status
- Dozens of co-official, regional and minority languages as well as languages of immigrants and trade partners
- Many economic, social and technical challenges
  - The Digital Single Market needs to be multilingual
  - Cross-border, cross-lingual, cross-cultural communication
  - Fragmentation of the LT market and landscape





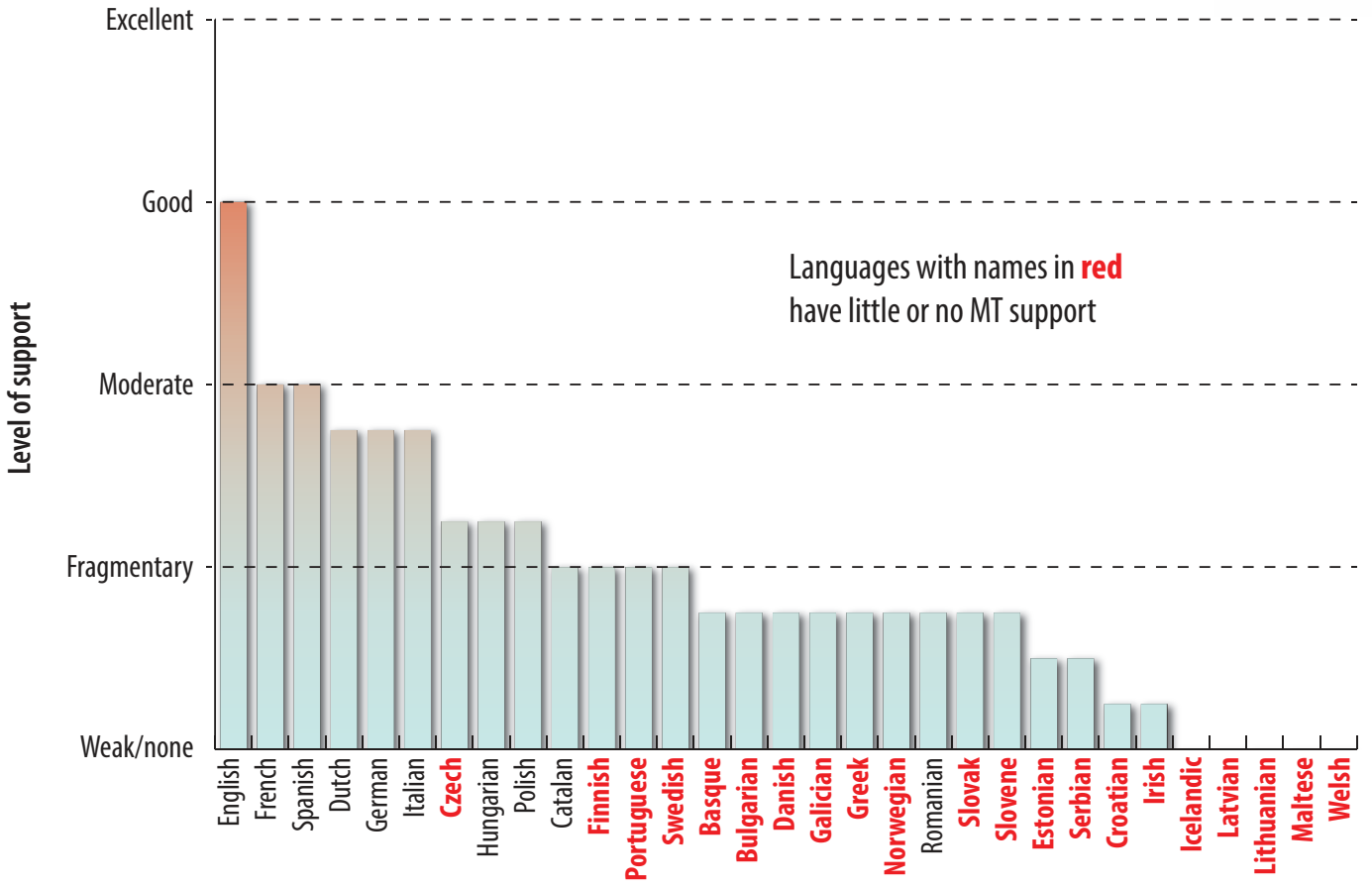
2012

META-NET Language White Papers "Europe's Languages in the Digital Age"

Basque	Galician	Polish*
Bulgarian*	German*	Portuguese*
Catalan	Greek*	Romanian*
Croatian*	Hungarian*	Serbian
Czech*	Icelandic	Slovak*
Danish*	Irish*	Slovene*
Dutch*	Italian*	Spanish*
English*	Latvian*	Swedish*
Estonian*	Lithuanian*	Welsh
Finnish*	Maltese*	
French*	Norwegian	

* Official EU language

2012



META-NET

Source: META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, September 2012. Georg Rehm and Hans Uszkoreit (series editors)

2018

European Parliament Resolution (2018)

EP Resolution “Language equality in the digital age”
P8_TA(2018)0332 – partially based on the STOA study

Voting (11 Sept. 2018): **592 yes** – 45 no

Recommendations addressed by European Language Grid:

29. **Create a European LT platform for sharing of services**

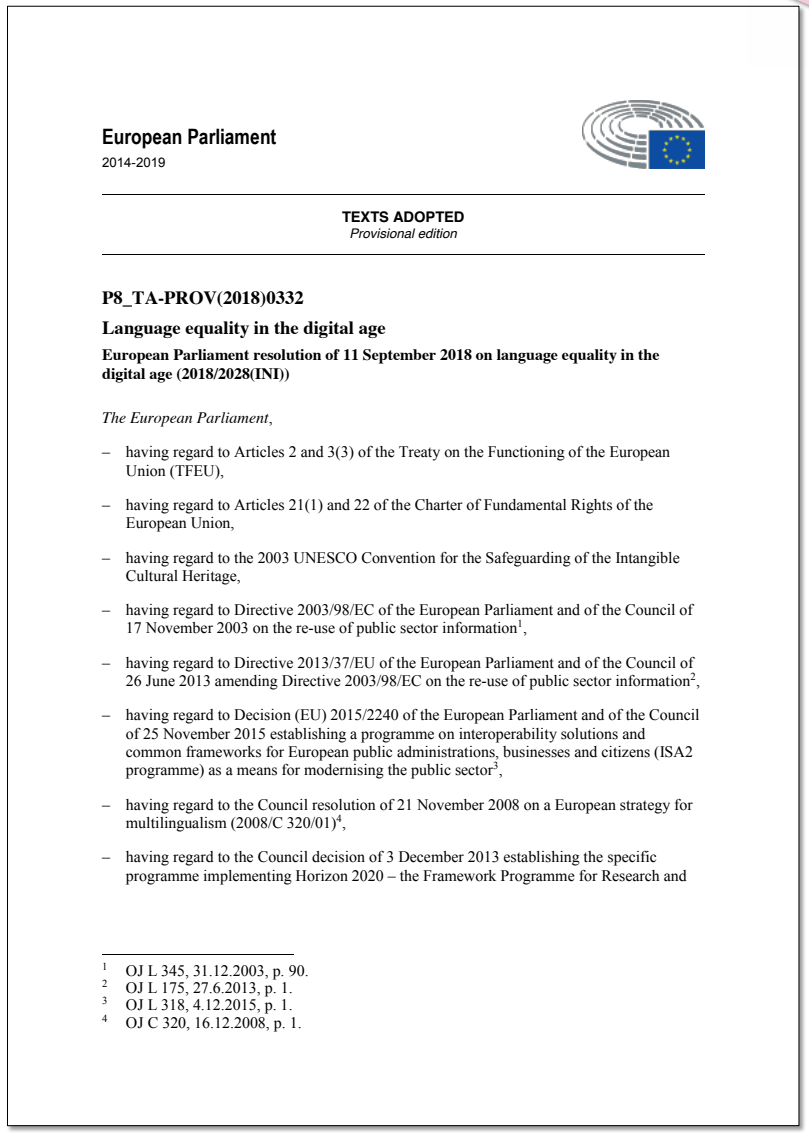
41. **Enable and empower European SMEs to use LTs**

26. **ICT integrators should be given economic incentives for LT**

27. **Europe has to secure its leadership in language-centric AI**

32. **Set up LT financing platform; emphasise R&D in Deep NLU**

40. **Develop investment instruments and accelerator programs**



European Parliament
2014-2019

TEXTS ADOPTED
Provisional edition

P8_TA-PROV(2018)0332

Language equality in the digital age

European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))

The European Parliament,

– having regard to Articles 2 and 3(3) of the Treaty on the Functioning of the European Union (TFEU),

– having regard to Articles 21(1) and 22 of the Charter of Fundamental Rights of the European Union,

– having regard to the 2003 UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage,

– having regard to Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information[1],

– having regard to Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information[2],

– having regard to Decision (EU) 2015/2240 of the European Parliament and of the Council of 25 November 2015 establishing a programme on interoperability solutions and common frameworks for European public administrations, businesses and citizens (ISA2 programme) as a means for modernising the public sector[3],

– having regard to the Council resolution of 21 November 2008 on a European strategy for multilingualism (2008/C 320/01)[4],

– having regard to the Council decision of 3 December 2013 establishing the specific programme implementing Horizon 2020 – the Framework Programme for Research and

1 OJ L 345, 31.12.2003, p. 90.
2 OJ L 175, 27.6.2013, p. 1.
3 OJ L 318, 4.12.2015, p. 1.
4 OJ C 320, 16.12.2008, p. 1.

EUROPEAN LANGUAGE GRID

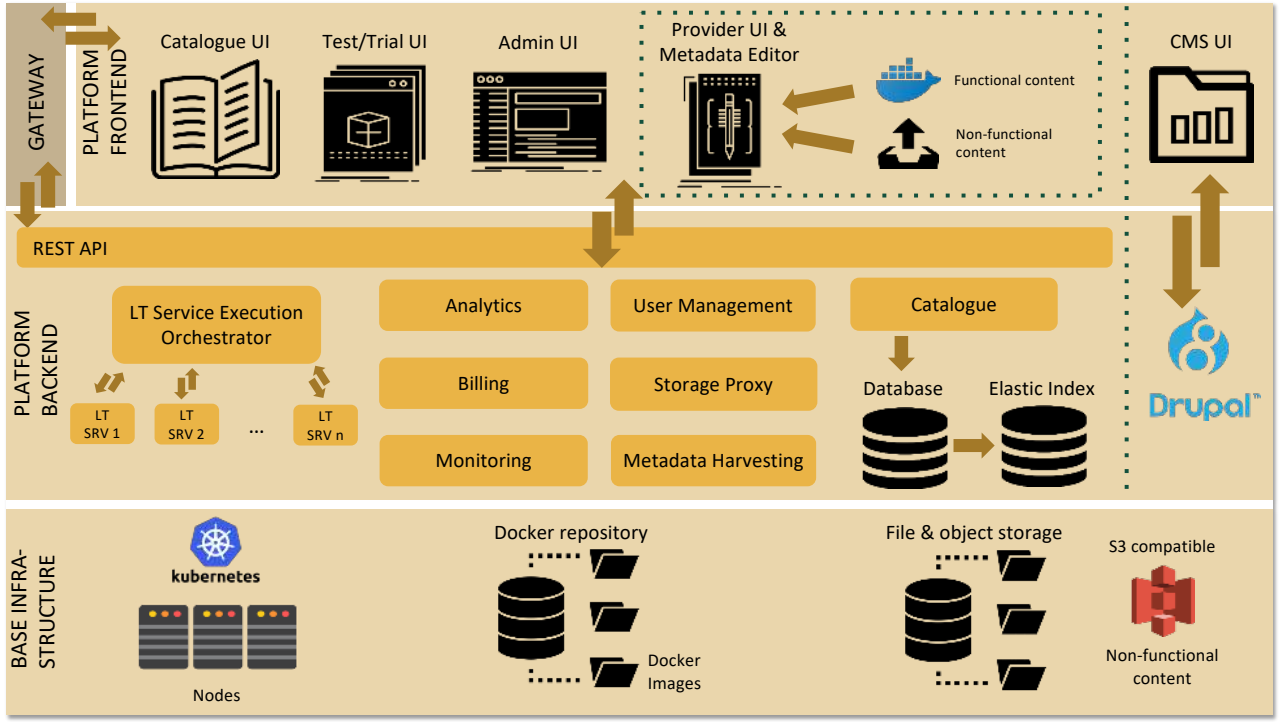
Kick-off meeting, 22/23 January 2019

Objectives (Selection)

1. Establish the ELG as the primary Language Technology platform and market place 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.

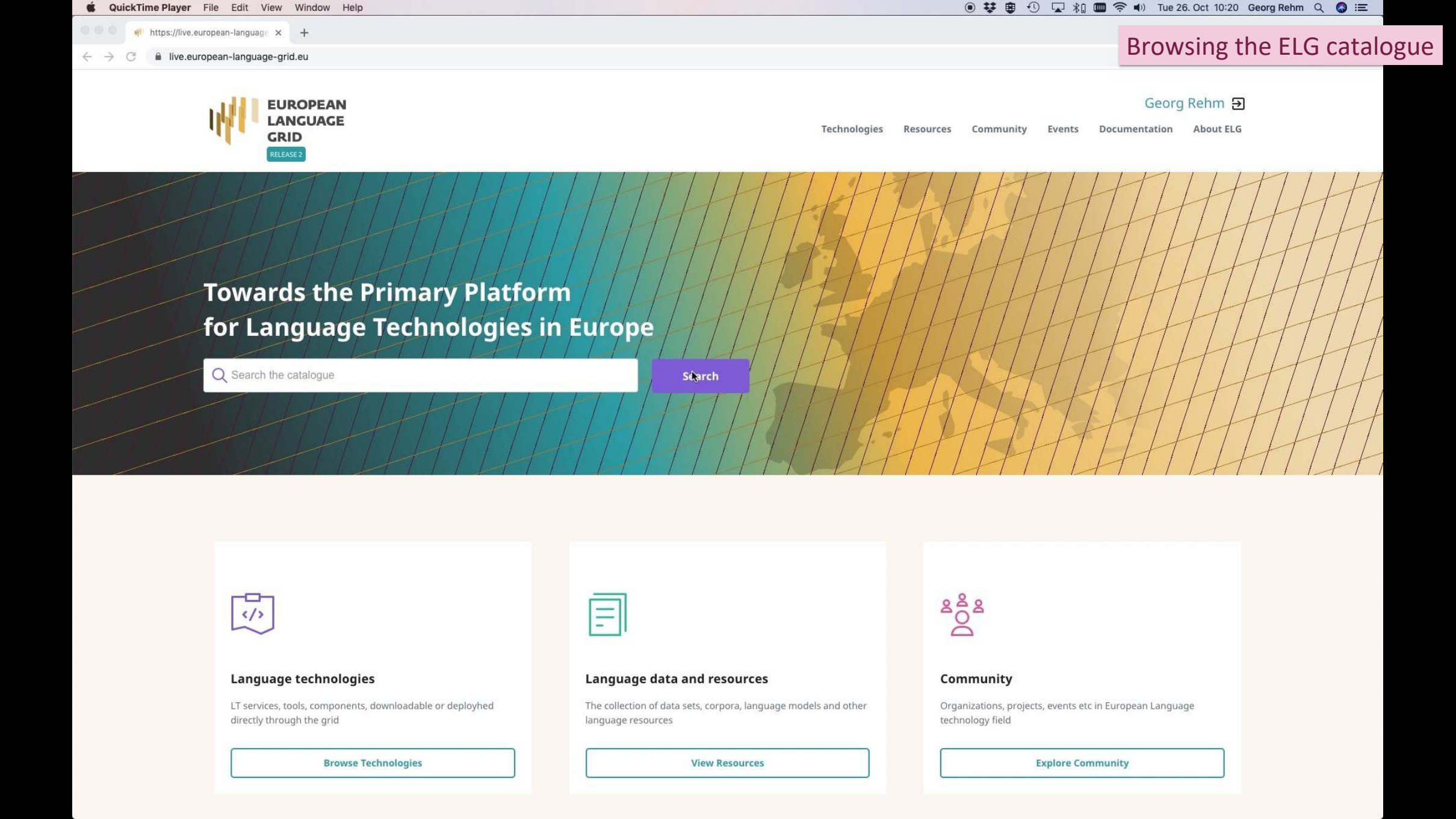
European Language Grid – Current State of Play (November 2021)

- User registration, authentication, authorisation
- User categories, respective rights and policies
- LT metadata upload
- Metadata conversion and harvesting
- LT service registration, integration
- LT service try out and execution
- LT data browsing, searching, downloading
- **Current state of play ELG Release 2 (Nov. 2021):**
  - 450 functional services and tools
  - 3563 corpora and data sets
  - 948 lexical/conceptual resources
  - 27 language descriptions, grammars, models
  - 1789 organisations (research orgs., companies)

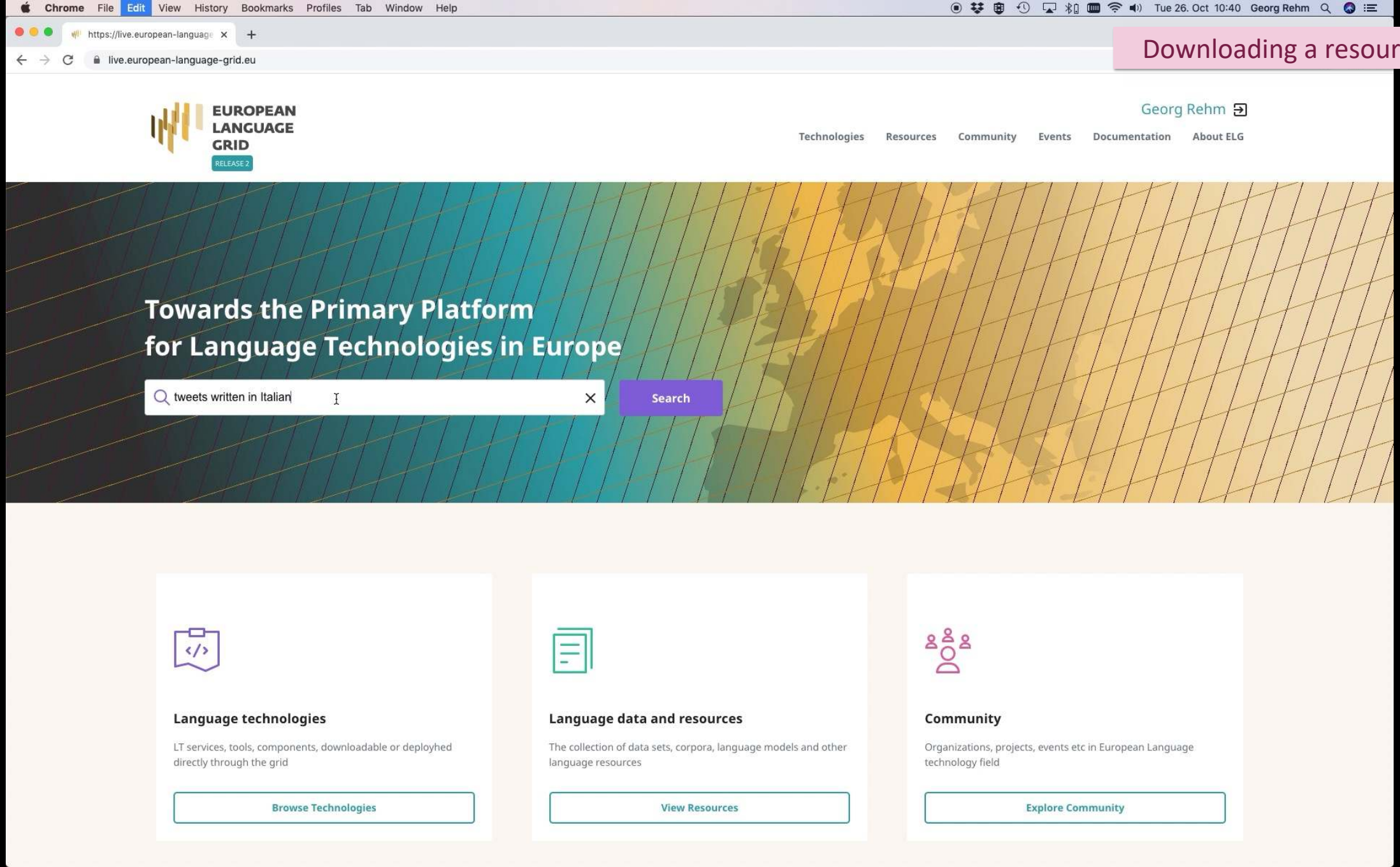


Users can connect to the ELG cloud platform via ELG APIs, remote APIs, ELG GUI, Python SDK, download of containers or source code.

ELG Release 3 to be made available in early 2022.


QuickTime Player File Edit View Window Help
Tue 26. Oct 10:20 Georg Rehm
https://live.european-language
live.european-language-grid.eu
Browsing the ELG catalogue
EUROPEAN LANGUAGE GRID
RELEASE 2
Georg Rehm
Technologies Resources Community Events Documentation About ELG
Towards the Primary Platform for Language Technologies in Europe
Search the catalogue
Search
Language technologies
LT services, tools, components, downloadable or deployhed directly through the grid
Browse Technologies
Language data and resources
The collection of data sets, corpora, language models and other language resources
View Resources
Community
Organizations, projects, events etc in European Language technology field
Explore Community

Downloading a resource



Testing an MT service

QuickTime Player File Edit View Window Help — Tue 26. Oct 10:52 Georg Rehm

European Language Grid - Cat

live.european-language-grid.eu/catalogue/search/UEDIN-MT-DeEn

EUROPEAN LANGUAGE GRID
RELEASE 2

Georg Rehm

Technologies Resources Community Events Documentation About ELG

UEDIN-MT-DeEn

Search

Language resources & technologies

Service functions

Languages

Licences

Conditions of use for data

Related entities

140 search results for UEDIN-MT-DeEn

UEDIN Machine Translation Service for German to English

1.0.0

111 views
2131 times used
ELG-compatible service

A machine translation (MT) service for German-to-English translation based on the Marian machine translation framework. The translation model is a basic transformer model trained on ca 13.3M sentence pairs using Marian N

Keywords: Machine Translation · German · English · Neural machine translation

Languages: German · English

Licence: Creative Commons Attribution Share Alike 4.0 International

UEDIN Machine Translation Service for Czech to English

1.0.0

25 views
152 times used
ELG-compatible service

An ensemble of 3 transformer-base models trained on WMT'19 data as a system for generating back-translations.

Keywords: Machine Translation · Czech · English · Neural machine translation

Languages: English · Czech

Licence: Creative Commons Attribution Share Alike 4.0 International

UEDIN Machine Translation Service for English to German

1.0.0

62 views
570 times used
ELG-compatible service

A machine translation (MT) service for English-to-German translation based on the Marian machine translation framework. The translation model is a basic transformer model trained on ca 13.3M sentence pairs using Marian N

Keywords: Machine Translation · German · English · Neural machine translation

Languages: German · English

Licence: Creative Commons Attribution Share Alike 4.0 International

Testing a dependency parser

EUROPEAN LANGUAGE GRID
RELEASE 2

Search for services, tools, datasets, organizations...

Search

conll shared task
2007 conll shared task
hosted at elg
udpipe english

Service functions
Languages
Licences
Conditions of use for data
Related entities

"Fatalne jaja" Bułhakow
unspecified
33 views
0 downloads

Story "Fatallne Jaja" Michaił Bułhakow

Keywords: korpus · korpus tekstowy
Language: Polish
Licence: Creative Commons Attribution 3.0 Unported

"Le Monde Diplomatique" Arabic tagged corpus
1
28612 views
0 downloads

This corpus contains 102,960 vowelised, lemmatised and tagged words (58 texts from Le Monde Diplomatique Arabic, see also ELRA-W0036-04). To each text are associated 3 files : - raw text in Arabic, - vowelized text in

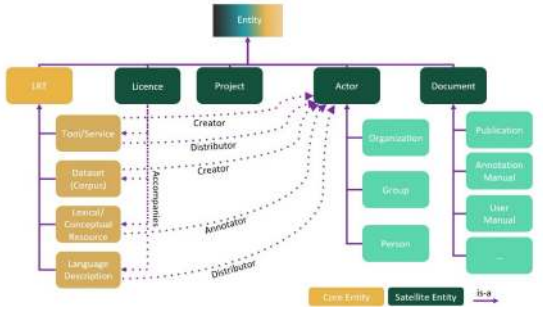
Keyword: corpus
Language: Arabic
Licences: ELRA-VAR-COMMERCIAL-MEMBER-COMMERCIALUSE-1.0
· ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0

"Le Monde Diplomatique" Text corpus in Arabic
1
116 views
0 downloads

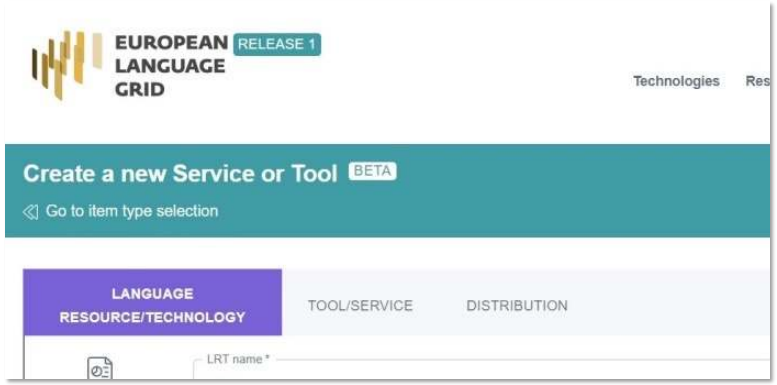
Electronic archiving of "Le Monde Diplomatique" articles in Arabic from 2000. The corpus is available in HTML. Each HTML file contains one article. Number of articles available per year : • 2000: 61 articles (November a

Keyword: corpus
Language: Arabic
Licences: ELRA-END-USER-ACADEMIC-MEMBER-NONCOMMERCIALUSE-1.0
· ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0

European Language Grid – What can you do with it?



- **Data consumers** can search and browse the ELG catalogue
  - for different types of data, language processing services, projects and organisations
  - download data (depending on access conditions)
- **Service consumers** can try out and test language processing services
  - call a service from the command line, integrate it into workflow
  - view code samples
  - current API support: MT, IE, ASR, TTS, text classification



- **Data and service consumers** can use a Python-based API for accessing the ELG catalogue, searching and directly fetching datasets to feed them into, e.g., their model training pipeline
- **Data providers** can upload resources and register metadata descriptions using a semantic schema
  - a dashboard provides an overview of own resources including their status (*draft* → *published*)
- **Language Technology providers** can provide their service/tool as a Docker container
  - Different options: (a) Docker image in ELG, can be run from ELG; (b) Docker image in ELG as proxy talks to remote service

Stakeholders and Users

Companies that

- … *develop* Language Technologies
- … *integrate* Language Technologies
- … *purchase* Language Technologies

Universities and research centres that

- … *develop* Language Technologies
- … *use* Language Technologies

META-FORUM 2019 (8/9 October) – Brussels, Belgium

Public administrations that *purchase* or *use* Language Technologies

Other organisations (e.g., NGOs) that *purchase* or *use* Language Technologies

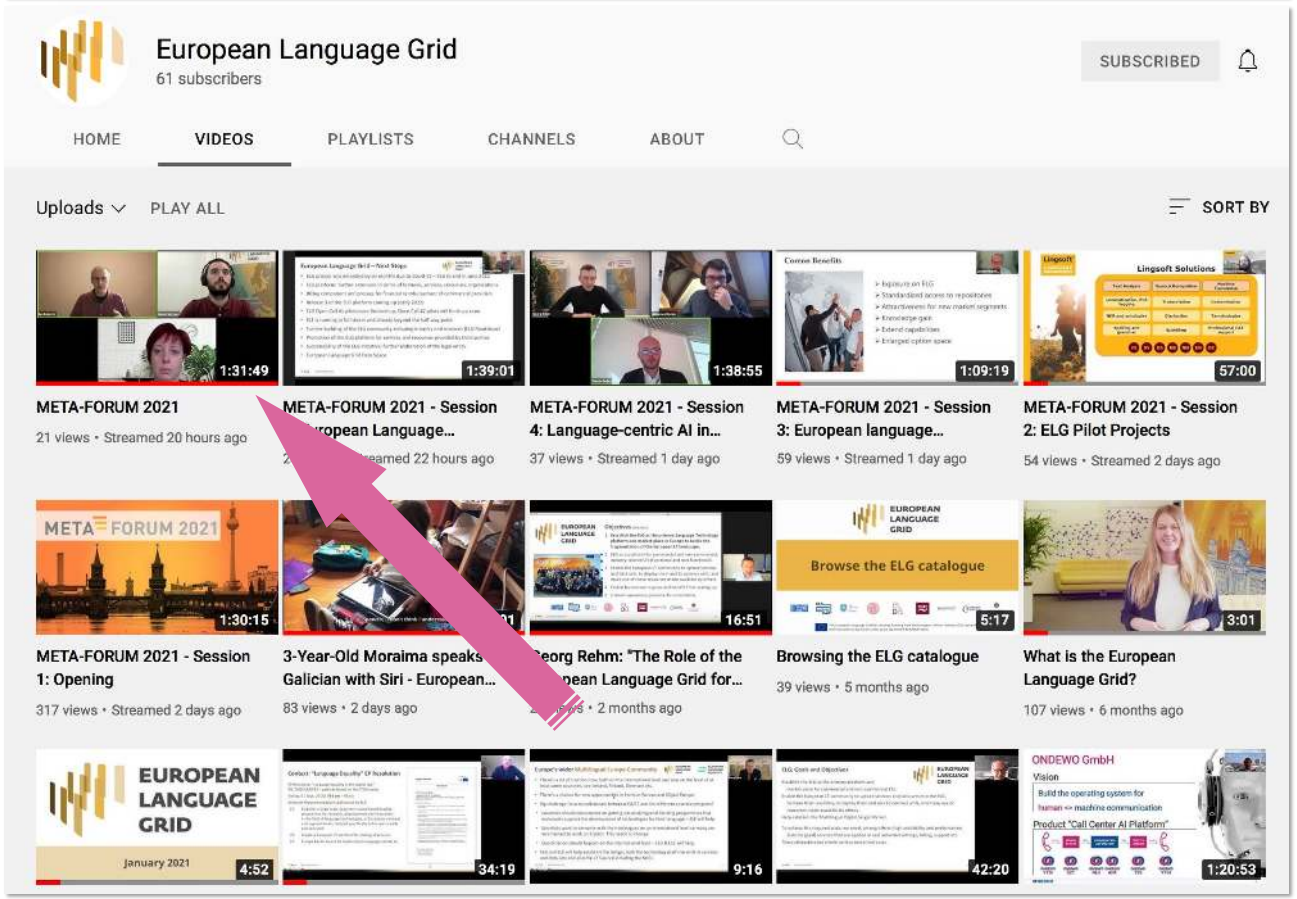
Funding agencies that support the development of Language Technologies

META-FORUM Conference Series

META-FORUM 2022 – June 08-10, Brussels, Belgium
Save the date!
META-FORUM 2021 – November 15-17, *virtual conference*
Using the European Language Grid
META-FORUM 2020 – December 01-03, *virtual conference*
Piloting the European Language Grid
META-FORUM 2019 – October 08/09, Brussels, Belgium
Introducing the European Language Grid
META-FORUM 2017 – November 13/14, Brussels, Belgium
Towards a Human Language Project
META-FORUM 2016 – July 04/05, Lisbon, Portugal
Beyond Multilingual Europe
META-FORUM 2015 – April 27, Riga, Latvia
Technologies for the Multilingual Digital Single Market
META-FORUM 2013 – September 19/20, Berlin, Germany
Connecting Europe for New Horizons
META-FORUM 2012 – June 20/21, Brussels, Belgium
A Strategy for Multilingual Europe
META-FORUM 2011 – June 27/28, Budapest, Hungary
Solutions for Multilingual Europe
META-FORUM 2010 – November 17/18, Brussels, Belgium
Challenges for Multilingual Europe

All META-FORUM 2020 and META-FORUM 2021 sessions are available on the European Language Grid YouTube channel:

https://www.youtube.com/channel/UCarEHmsWT2JslcvvWkbhL4A



Example: Integration of the results of the project QURATOR into the ELG platform.

Qurator
Curation Technologies

ART+COM

ada

SEMTATION

ubermetrics

3pc
Neue Kommunikation

SBB

condat

WIKIMEDIA DEUTSCHLAND

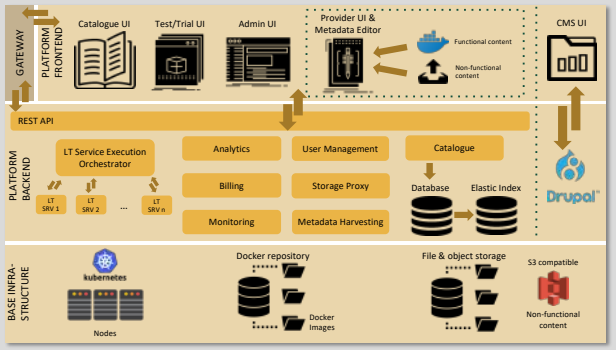
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

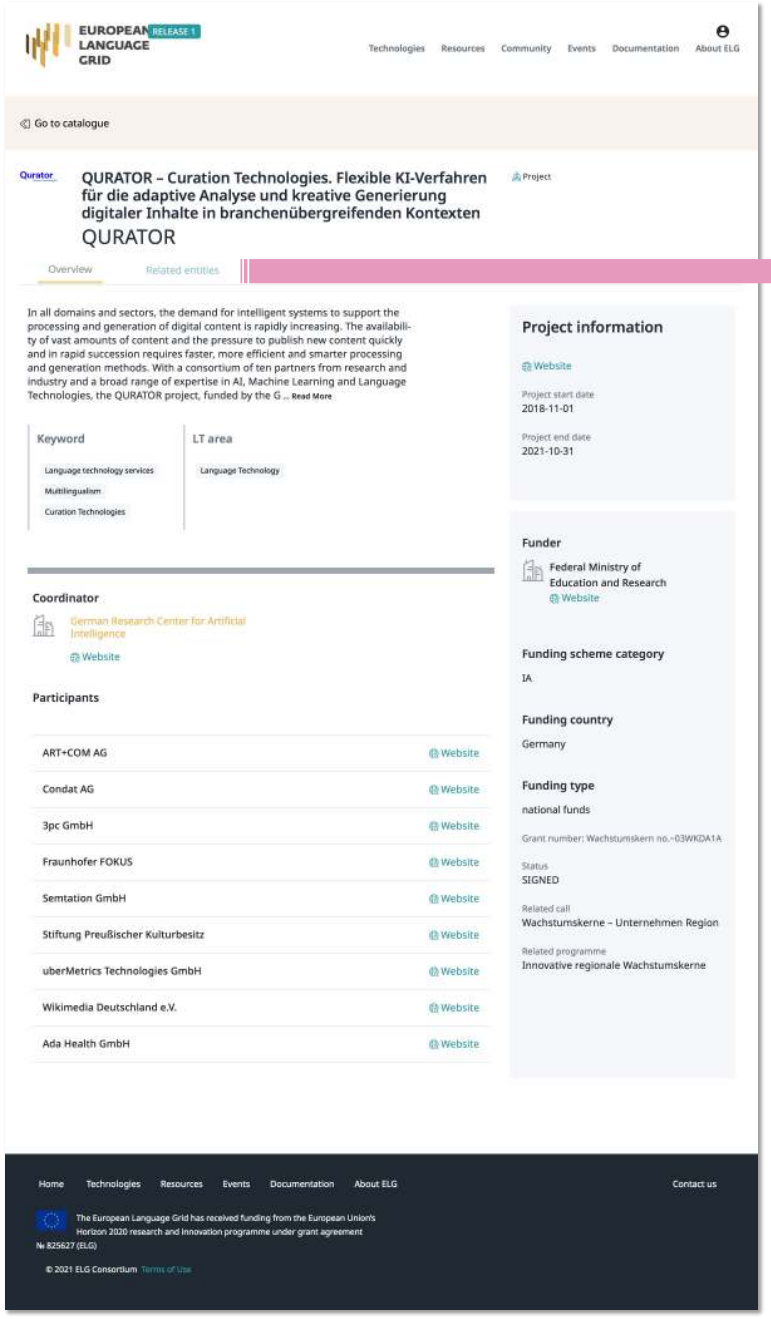
Fraunhofer FOKUS

What does that mean?

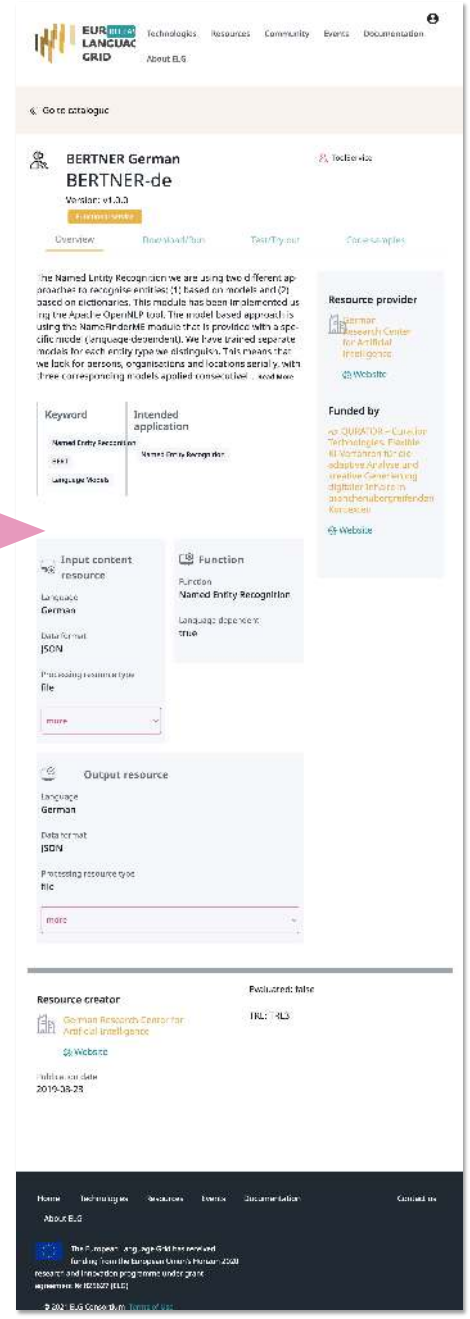
- We created a page for the QURATOR project in the ELG catalogue
- We created for all QURATOR partners, too
- Two QURATOR services have already been made available through the ELG platform – their metadata records are associated with QURATOR
- More QURATOR resources will be made available in the future

EUROPEAN LANGUAGE GRID





Similar integration of services or resources can be done by companies or research groups: all members of the European LT community can make their assets available via the ELG platform!



“At Least 21 European Languages in Danger of Digital Extinction”

2012

Press release on European Day of Languages (Sept. 26, 2012)

Huge global interest in the topic and our key findings!

600+ mentions in the press

40+ countries in 35+ different languages

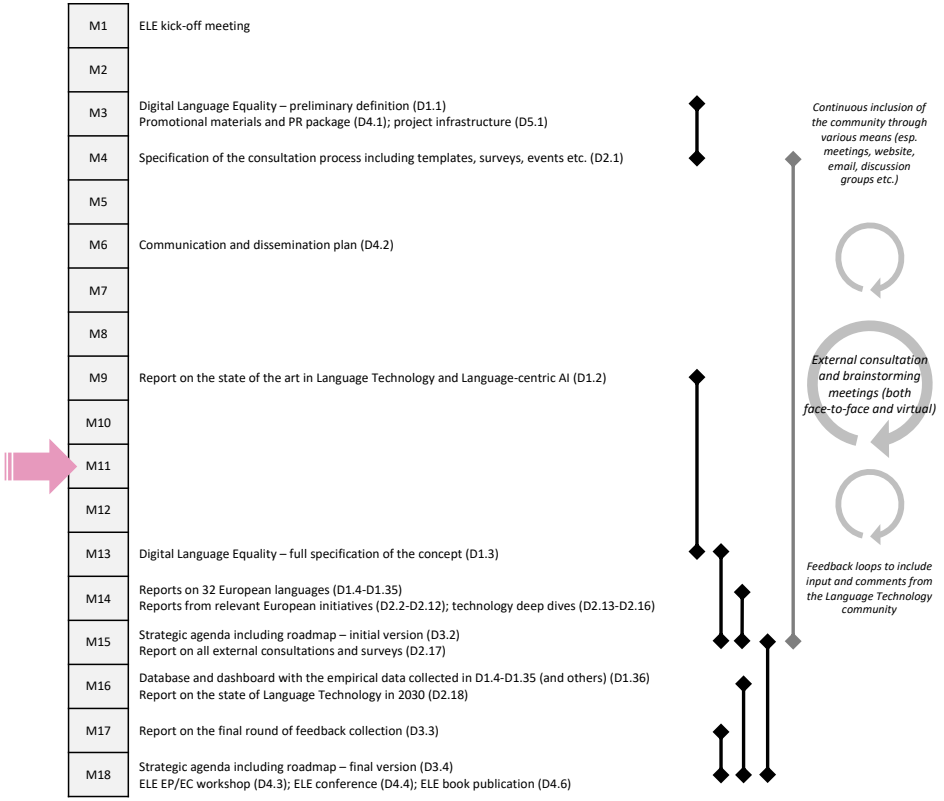
20+ television reports and 30+ interviews (radio, tv) with META-NET representatives

Two Parliamentary Questions in the EP on the “digital extinction of languages” topic

ELG

META-NET

EUROPEAN LANGUAGE EQUALITY



Consortium: 52 partners from all over Europe

Coordinator: ADAPT Centre (Dublin City University)

Objective: *development of a strategic research, innovation and deployment agenda to achieve digital language equality in Europe by 2030*

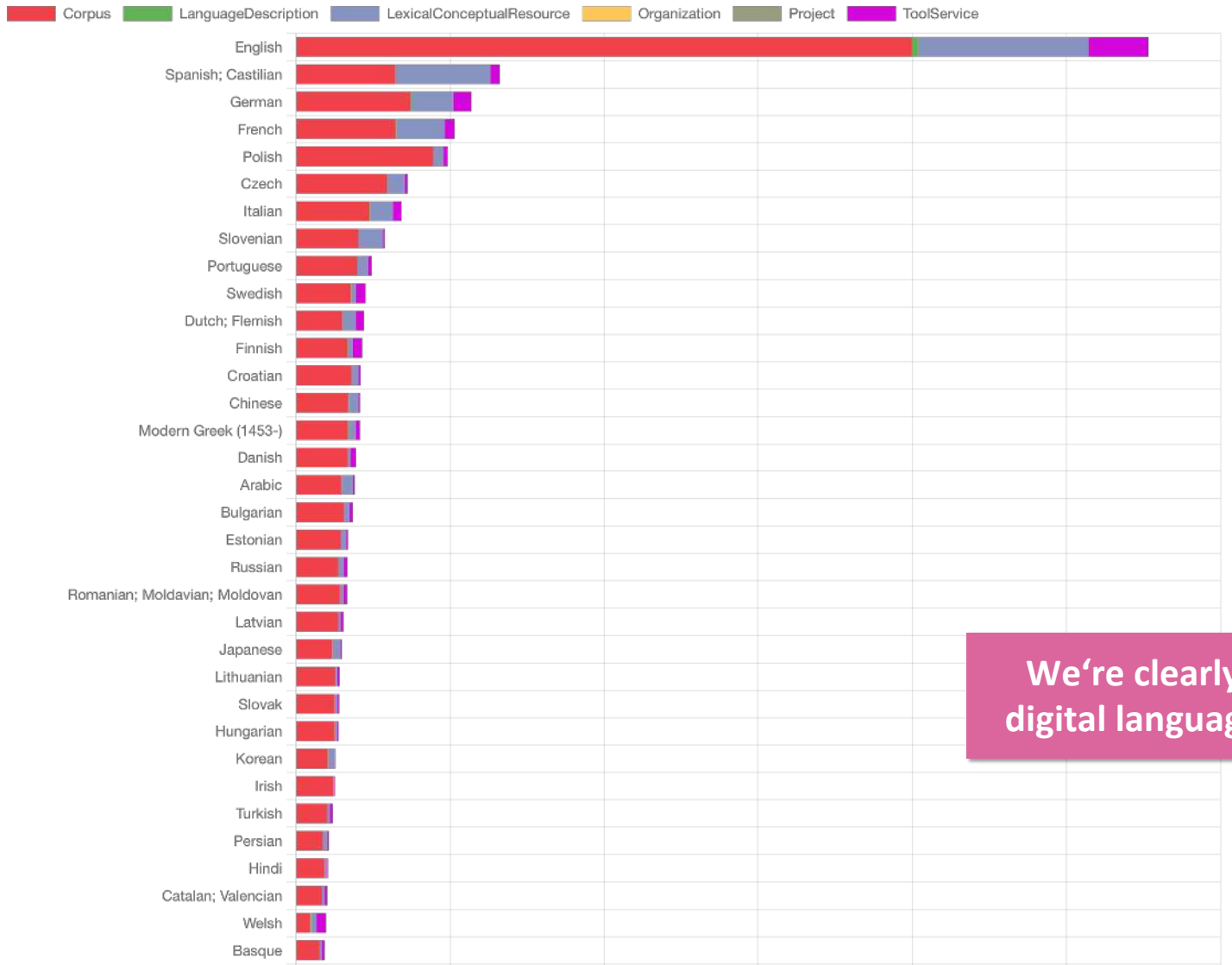
Runtime: 18 months – ELE & ELG will finish in June 2022

Start on **1 January 2021**

PP/PA (*not* H2020) – Pilot Project/Preparatory Action

Budget: 1,8M€ (Coordination and Support Action)

Languages of the Resources currently in ELG



We're clearly very far away from digital language equality in Europe.

WP1 European Language Equality: Status Quo in 2020/2021
Task 1.1: Defining Digital Language Equality
Task 1.2: Language Technologies and Language-centric AI – State of the Art
Task 1.3: Language Technology Support of Europe's Languages in 2020/2021

WP2 European Language Equality: The Future Situation in 2030
Task 2.1: The perspective of European LT developers (industry and research)
Task 2.2: The perspective of European LT users and consumers
Task 2.3: Science – Technology – Society: Language Technology in 2030

EUROPEAN LANGUAGE EQUALITY

Digital Language Equality: Definition of the concept

Language Technology and language-centric AI: State of the Art

32 reports on the technology support of 32 European languages (META-NET White Paper update)

Reports from networks, initiatives and associations

Deep dives (MT, speech, text analytics, data)

Report on external consultations and surveys

Forecast: Language Technology in 2030

WP3 Development of the Strategic Agenda and Roadmap
Task 3.1: Desk research – landscaping
Task 3.2: Consolidation and aggregation of all input received
Task 3.3: Final round of feedback collection

Existing strategic documents and projects in LT/AI

Strategic agenda and roadmap: initial version

Final round of feedback collection

Strategic agenda and roadmap: final version

WP4 Communication – Dissemination – Exploitation – Sustainability
Task 4.1: Overall project communication and dissemination
Task 4.2: Liaise with EP/EC – organisation of a targeted workshop
Task 4.3: Organisation of final ELE conference
Task 4.4: Production of PR materials and sustainable results

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

EP/EC Workshop

ELE Conference

ELE Strategic Agenda and Roadmap (print version, interactive version)

Final ELE Book Publication

WP5 Project Management
Task 5.1: Overall project management including Project Management Office
Task 5.2: Digital collaboration and document management infrastructure

Work Packages and main Deliverables

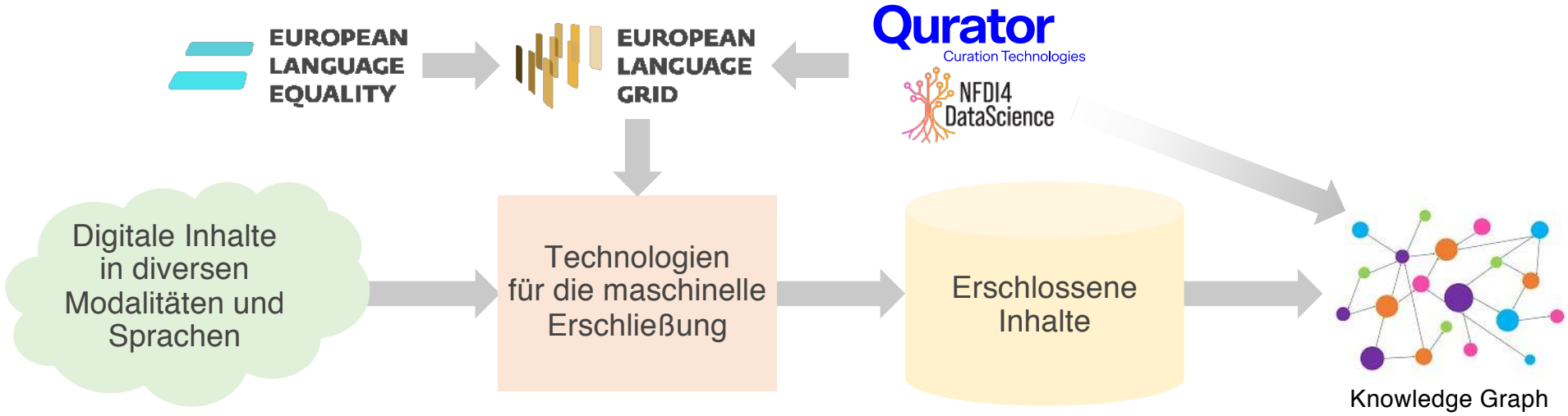
ELG and ELE: Summary and Next Steps

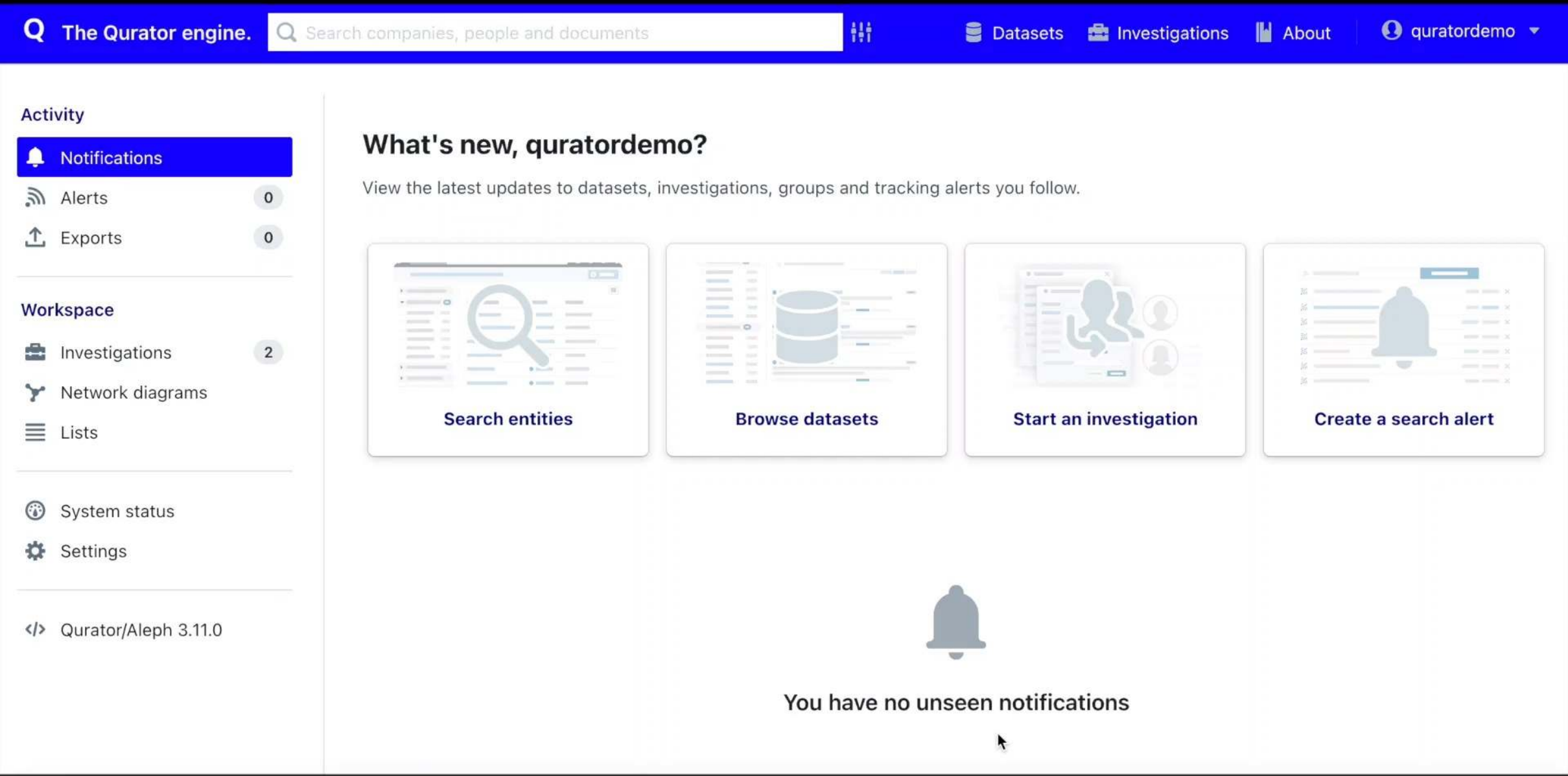
- **ELE:** plan and strategic agenda how to achieve digital language equality in Europe by 2030.
- We will recommend a **coordinated action plan** between the EU and Member States.
- Results and recommendations to be presented to the **European Parliament** in March 2022.
- Goal: develop resources, datasets and technologies *for all European languages*.
- **Establish ELG as the primary platform and marketplace for Language Technology in Europe.**
- ELG is 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**.
- ELG will also contribute to ELE's mission of **Digital Language Equality** in Europe by giving all our languages one virtual home and umbrella platform that collects **all** services and resources.
- **Next steps:** validation of **ELG products**; **raise awareness** of LT for Europe; develop **ELG Release 3** (early 2022); establish the **ELG legal entity** (late 2021).

Zusammenfassung

Zusammenfassung

- QURATOR entwickelt zahlreiche Services für Kuratierungstechnologien
- ELG als gemeinsame Plattform der europäischen Sprachtechnologie-Community
- Unser Kontext: ELG als Sammlung von Technologien für die maschinelle Erschließung
- Für die Erschließung insbesondere Textanalytik, ASR, OCR und MT
- Relevant vor allem für Library Labs und Experimente im Bereich maschinelle Erschließung




The Qurator engine.
Search companies, people and documents
Datasets
Investigations
About
quratordemo
Activity
Notifications
Alerts 0
Exports 0
Workspace
Investigations 2
Network diagrams
Lists
System status
Settings
Qurator/Aleph 3.11.0
What's new, quratordemo?
View the latest updates to datasets, investigations, groups and tracking alerts you follow.
Search entities
Browse datasets
Start an investigation
Create a search alert
You have no unseen notifications

Herzlichen Dank!

Prof. Dr. Georg Rehm

Principal Researcher + DFKI Research Fellow

Speech and Language Technology Lab
DFKI, Berlin, Germany

- **georg.rehm@dfki.de**
- **http://georg-re.hm**
- **http://de.linkedin.com/in/georgrehm**
- **https://www.slideshare.net/georgrehm**

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

https://twitter.com/EuroLangTech

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

Subscribe to our newsletter (2500+ subscribers already)