Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Herausforderungen und Lösungen für die europäische Sprachtechnologie-Forschung und -Entwicklung

Dr. Georg Rehm

DFKI GmbH

georg.rehm@dfki.de

Research Fellow-Präsentation – Berlin, 30. Oktober 2018

14. Juli 2017

1. Herausforderung

Das mehrsprachige Europa:
Sprachtechnologien für alle europäischen Sprachen?

2. Herausforderung

Online-Desinformationskampagnen:
Technische Lösungsansätze gegen „Fake News“?

3. Herausforderung

Digitaler Content:
Technologien für die effiziente Content-Kuratierung?

- **Mehrsprachigkeit ist der Kern der europäischen Idee**
- **24 EU Sprachen – alle besitzen den gleichen Status**
- **Dutzende von regionalen und Minderheitensprachen sowie Sprachen von Migranten und Handelspartnern**
- **Viele ökonomische, gesellschaftliche und technische Herausforderungen:**
  - **Mehrsprachigkeit für den digitalen Binnenmarkt**
  - **Technologien für sprachübergreifende sowie auch kulturübergreifende Kommunikation**
  - **Sprachtechnologien *gegen* den wachsenden Nationalismus und *für* die europäische Identität?**

META-NET

60 Forschungszentren in **34** Ländern.
Vorsitzender: Jan Hajic (CUNI)
Vizev.: J. van Genabith (DFKI), A. Vasiljevs (Tilde)
Generalsekretär: Georg Rehm (DFKI)

META

Multilingual Europe Technology Alliance.
900+ Mitglieder in **67** Ländern

veröffentlicht 2013

31 Bände, veröffentlicht 2012

T4ME (META-NET) CESAR META-NORD METANET4U

CRACKER

Cracking the Language Barrier

Coordination, Evaluation and Resources for European MT Research

Coordination and Support Action, H2020-ICT17, 2015–2017, 36 Monate – http://www.cracker-project.eu

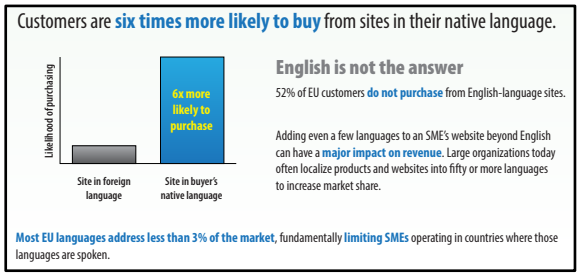
1	DFKI	Deutschland	Georg Rehm
2	CUNI	Tschechien	Jan Hajic
3	ELDA	Frankreich	Khalid Choukri
4	FBK	Italien	Marcello Federico
5	ATHENA RC	Griechenland	Stelios Piperidis
6	UEDIN	UK	Philipp Koehn
7	USFD	UK	Lucia Specia

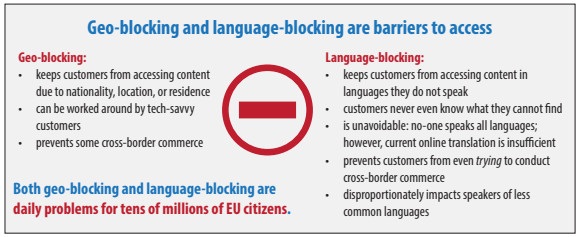
Zusammenführung und Integration von Communitys

- META-NET mit META-SHARE und META
- MT Evaulierungsinitiativen (WMT, IWSLT, MT Marathon)
- MT und andere LT-Industrien
- Sprachressourcen – META-SHARE, ELRA
- HT/MT Evaluationswerkzeuge – translate5
- Übersetzungsindustrie
- Nutzer maschineller Übersetzung

Strategic Agenda for the Multilingual Digital Single Market

- Version 0.5 präsentiert beim META-FORUM 2015
- Version 0.9 präsentiert beim META-FORUM 2016
- Version 1.0 präsentiert beim META-FORUM 2017



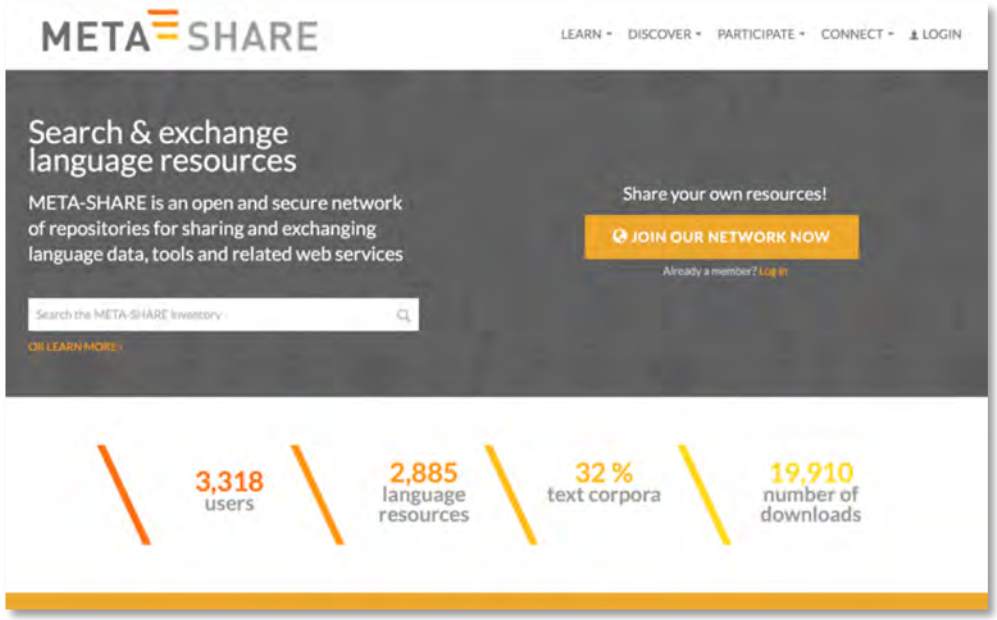


META-SHARE

META NET

- META-SHARE adressiert Technologie- bzw. Infrastrukturbedarfe in Bezug auf die
  - Sichtbarkeit
  - Dokumentation
  - Identifizierung
  - Verfügbarkeit
  - Langzeitspeicherung
  - Interoperabilität

  von LRs und LTs



- 35 META-SHARE-Mitglieder und Organisationen in 25 Ländern
- http://www.meta-share.org

Stelios Piperidis, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini, and Christian Girardi. "META-SHARE: One year after." In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014), pages 1532-1538, Reykjavik, Iceland, May 2014.

Georg Rehm. "The Language Resource Life Cycle: Towards a Generic Model for Creating, Maintaining, Using and Distributing Language Resources". In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016), pages 2450-2454, Portorož, Slovenia, May 2016.


Baskisch
Bulgarisch*
Deutsch*
Dänisch*
Englisch*
Estnisch*
Finnisch*
Französisch*
Galizisch
Griechisch*
Irisch*
Isländisch
Italienisch*
Katalanisch
Kroatisch*
Lettisch*
Litauisch*
Maltesisch*
Niederländisch*
Norwegisch
Polnisch*
Portugiesisch*
Rumänisch*
Schwedisch*
Serbisch
Slowakisch*
Slowenisch*
Spanisch*
Tschechisch*
Ungarisch*
Walisisch
http://www.meta-net.eu/whitepapers
* Offizielle EU-Sprache

MT

exzellente	gute	moderate	fragmentarische	schwache/keine Unterstützung
	Englisch	Französisch, Spanisch	Katalanisch, Niederländisch, Deutsch, Ungarisch, Italienisch, Polnisch, Rumänisch	Baskisch, Bulgarisch, Kroatisch, Tschechisch, Dänisch, Estnisch, Finnisch, Galizisch, Griechisch, Isländisch, Irisch, Lettisch, Litauisch, Maltesisch, Norwegisch, Portugiesisch, Serbisch, Slowakisch, Slowenisch, Schwedisch, Walisisch

Textanalytik

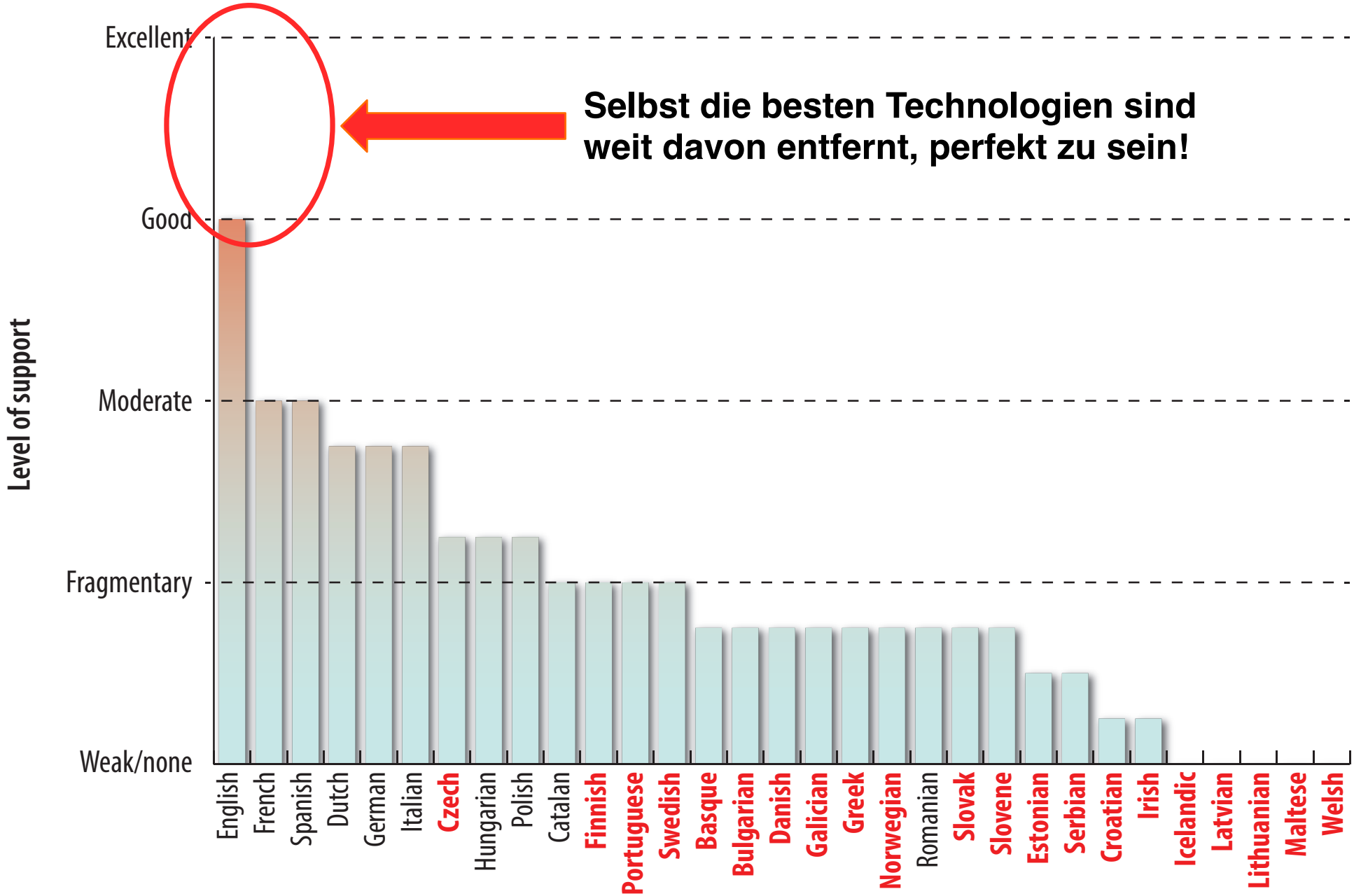
exzellente	gute	moderate	fragmentarische	schwache/keine Unterstützung
	Englisch	Niederländisch, Französisch, Deutsch, Italienisch, Spanisch	Baskisch, Bulgarisch, Katalanisch, Tschechisch, Dänisch, Finnisch, Galizisch, Griechisch, Ungarisch, Norwegisch, Polnisch, Portugiesisch, Rumänisch, Slowakisch, Slowenisch, Schwedisch	Kroatisch, Estnisch, Isländisch, Irisch, Lettisch, Litauisch, Maltesisch, Serbisch, Walisisch

Speech

exzellente	gute	moderate	fragmentarische	schwache/keine Unterstützung
	Englisch	Tschechisch, Niederländisch, Finnisch, Französisch, Deutsch, Italienisch, Portugiesisch, Spanisch	Baskisch, Bulgarisch, Katalanisch, Dänisch, Estnisch, Galizisch, Griechisch, Ungarisch, Irisch, Norwegisch, Polnisch, Serbisch, Slowakisch, Slowenisch, Schwedisch	Kroatisch, Isländisch, Lettisch, Litauisch, Maltesisch, Rumänisch, Walisisch

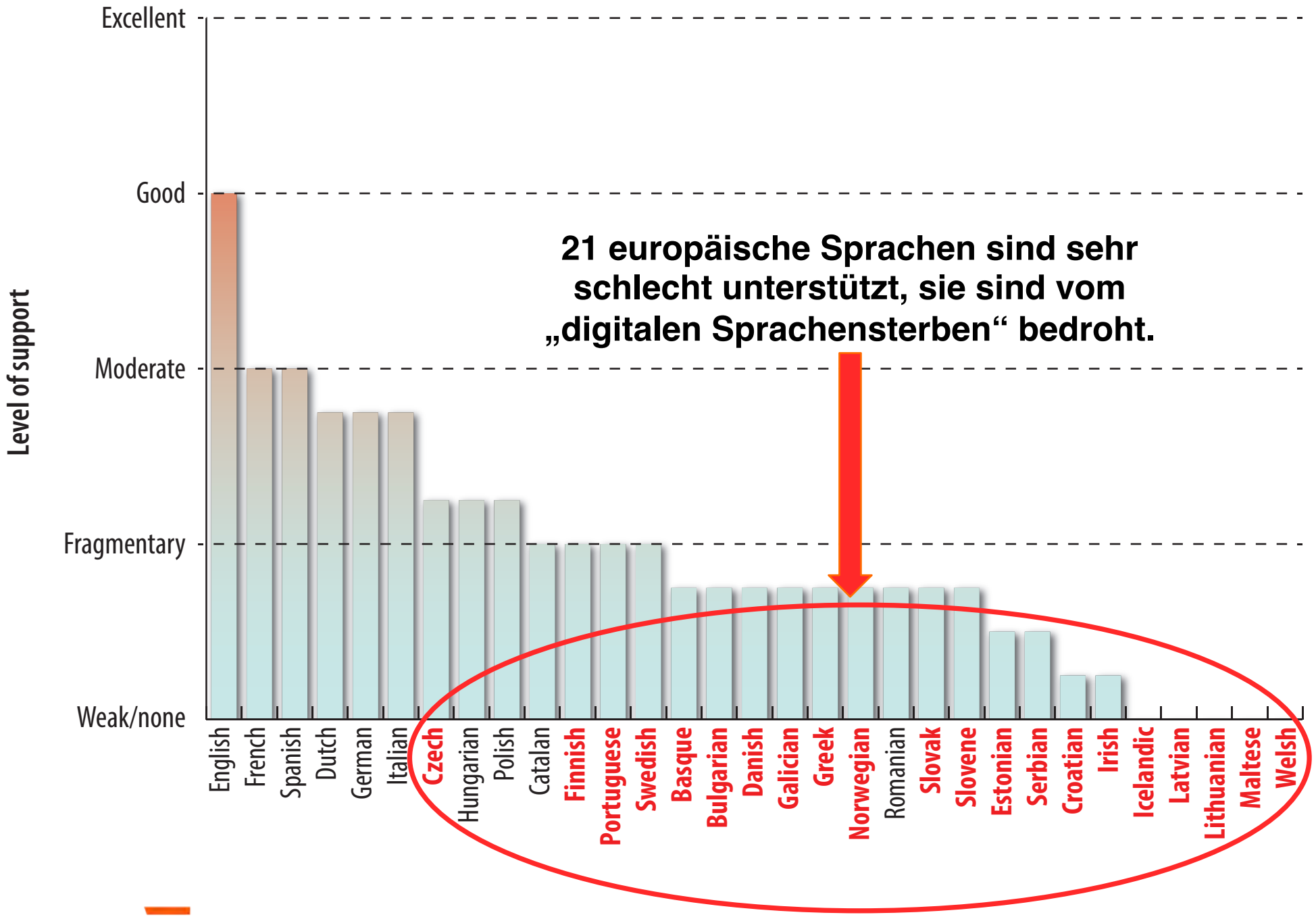
Ressourcen

exzellente	gute	moderate	fragmentarische	schwache/keine Unterstützung
	Englisch	Tschechisch, Niederländisch, Französisch, Deutsch, Ungarisch, Italienisch, Polnisch, Spanisch, Schwedisch	Baskisch, Bulgarisch, Katalanisch, Kroatisch, Dänisch, Estnisch, Finnisch, Galizisch, Griechisch, Norwegisch, Portugiesisch, Rumänisch, Serbisch, Slowakisch, Slowenisch	Isländisch, Irisch, Lettisch, Litauisch, Maltesisch, Walisisch



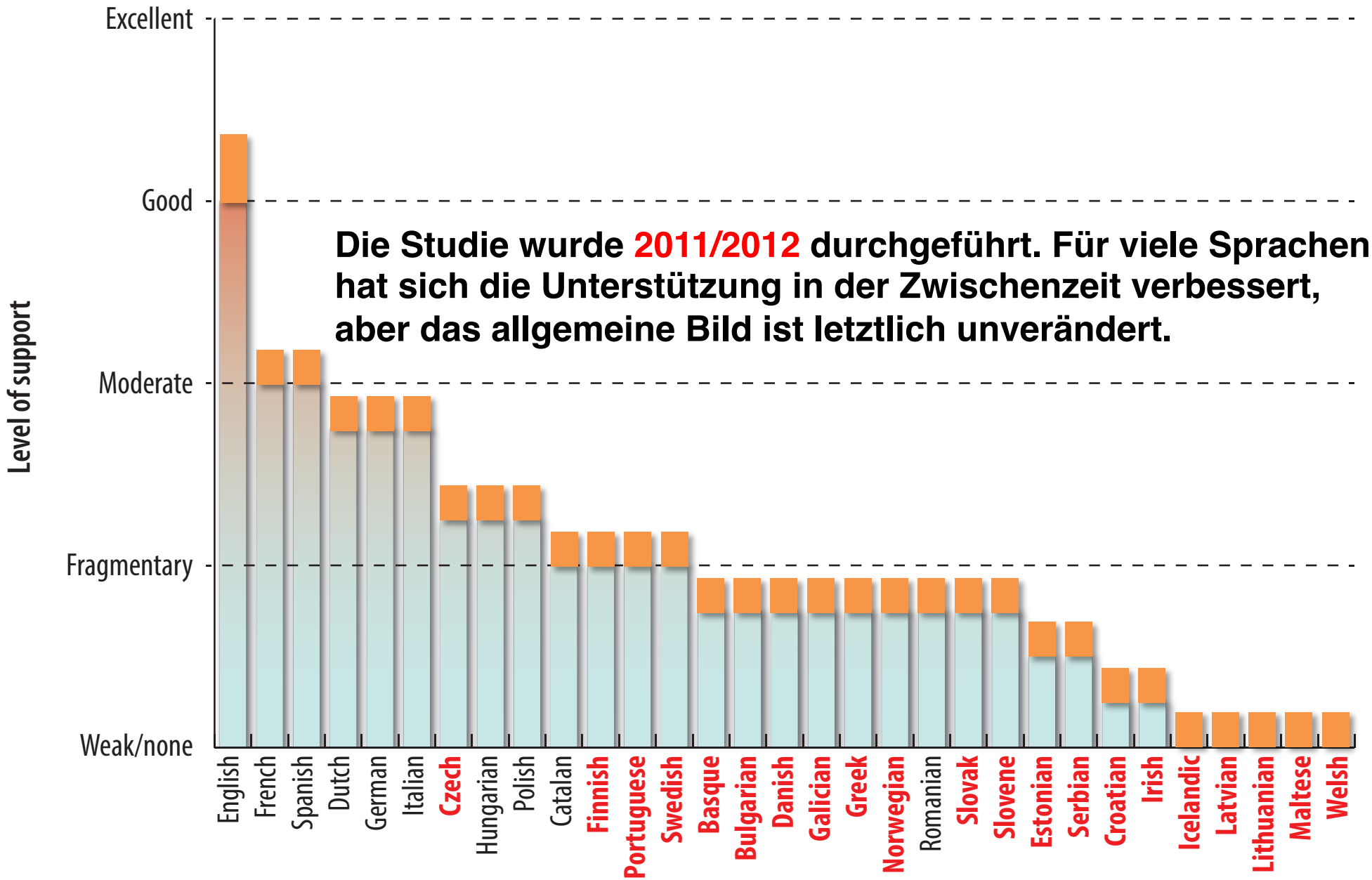
META-NET

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)



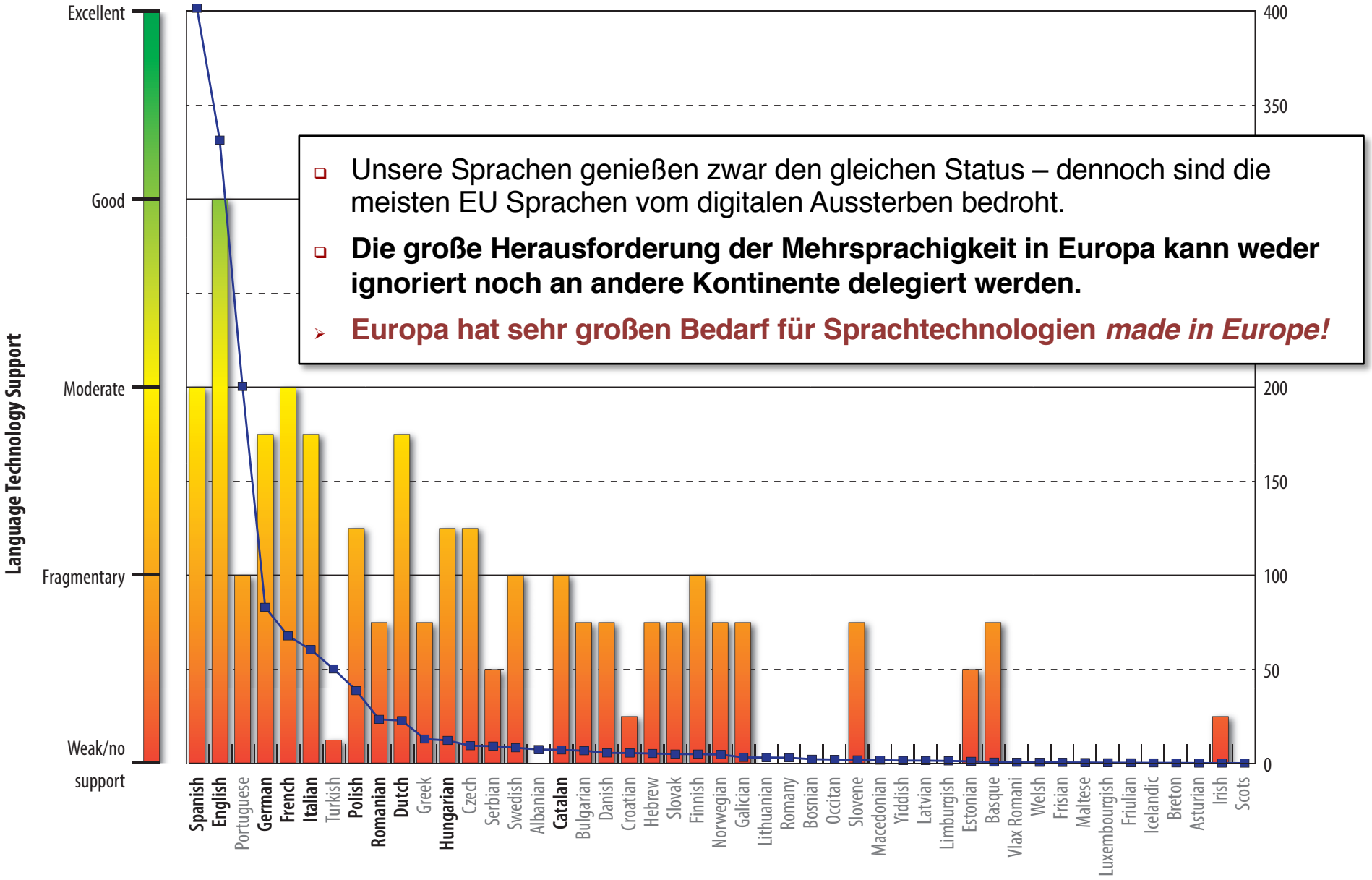
META-NET

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)



META-NET

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)



Georg Rehm, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Váradi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, and Sigve Gramstad. An Update and Extension of the META-NET Study “Europe's Languages in the Digital Age”. In Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014), Reykjavik, Iceland, May 2014.

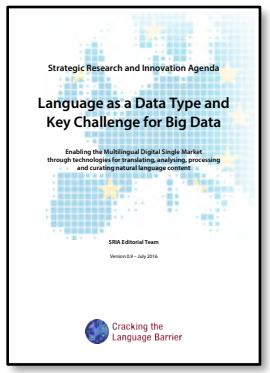
Georg Rehm, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen García-Mateo, Josef van Genabith, Jan Hajič, Inma Hernáez, John Judge, Svetla Koeva, Simon Krek, Cvetana 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 Skadiņa, Koenraad De Smedt, Marko Tadić, Paul Thompson, Dan Tufiş, Tamás Váradi, Andrejs Vasiļjevs, Kadri Vider, and Jolanta Zabarskaite. *The Strategic Impact of META-NET on the Regional, National and International Level.* Language Resources and Evaluation, 50(2):351-374, 2016.

META-NET SRA, veröffentlicht im Frühjahr 2013

- Erste strategische Forschungsagenda unseres Gebiets
- Komplexer Prozess der Sammlung von Technologievisionen
- Etwa 200 Forscherinnen und Forscher haben mitgewirkt

SRIA V0.5 präsentiert beim META FORUM 2015

- Basiert auf Strategiepapieren und Roadmaps, die von diversen EU-Projekten erstellt wurden, inklusive META-NET SRA (s.o.)



SRIA V0.9 präsentiert beim META FORUM 2016

- Vorbereitet, präsentiert und unterstützt von der Cracking the Language Barrier Föderation
- Erläutert, wie die LT-Community für Mehrsprachigkeit im digitalen Binnenmarkt sorgen kann.

SRIA V1.0 präsentiert beim META FORUM 2017

- Unterstützt und komplementiert die STOA-Studie.
- Wichtigste Empfehlung: Das Human Language Project initiieren.

Georg Rehm and Hans Uszkoreit, editors. *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer, Heidelberg, New York, Dordrecht, London, 2013.

Georg Rehm, editor. *Language Technologies for Multilingual Europe: Towards a Human Language Project*. Strategic Research and Innovation Agenda. Dec. 2017. Version 1.0. Unveiled at META-FORUM 2017 in Brussels, Belgium, on Nov. 13/14, 2017. Prepared by the Cracking the Language Barrier federation, supported by CRACKER.

Georg Rehm, editor. *Language as a Data Type and Key Challenge for Big Data*. Strategic Research and Innovation Agenda for the Multilingual Digital Single Market. CRACKER and Cracking the Language Barrier federation, July 2016. Version 0.9. 04 July 2016. Supported by CRACKER and LT_Observatory.

Georg Rehm, editor. *Strategic Agenda for the Multilingual Digital Single Market – Technologies for Overcoming Language Barriers towards a truly integrated European Online Market*. CRACKER and LT_Observatory, April 2015. Version 0.5. 22 April 2015. Prepared by the EU-funded projects CRACKER and LT_Observatory.

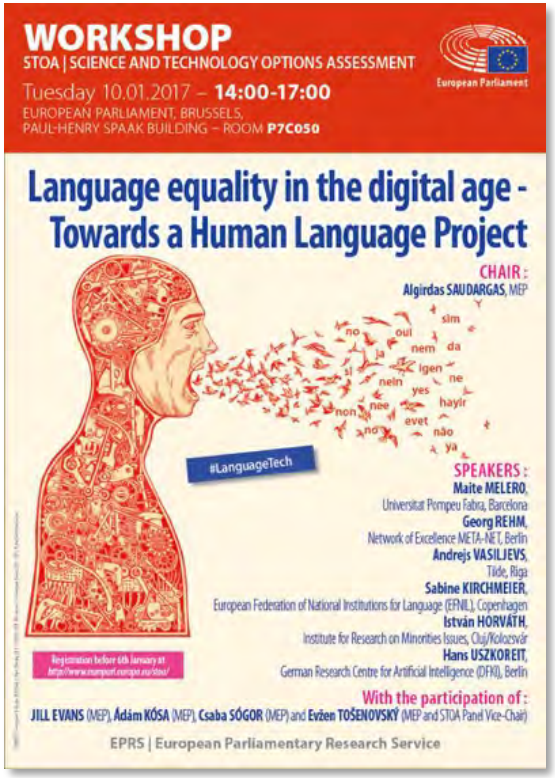
“Multilingual Europe”-Umfrage

- Durchgeführt im Mai/Juni 2017
- 29 Fragen (16 offene, 13 multiple choice)
- 634 Teilnehmer aus 52 Ländern
- Sehr hohe Komplettierungsrate (27%)
- Durchschnitt: 35,48 Minuten (!)
- 97% unterstützen das HLP
- 87% glauben, dass „tiefes maschinelles Sprachverstehen bis 2030“ ein adäquates wissenschaftliches Ziel ist.

Georg Rehm, Jan Hajic, Josef van Genabith, and Andrejs Vasiļjevs. “Fostering the Next Generation of European Language Technology: Recent Developments – Emerging Initiatives – Challenges and Opportunities.” In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016), pages 1586-1592, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

Georg Rehm and Stefanie Hegele. “Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs.” In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), pages 3282-3289, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

STOA Workshop
Europaparlament
10. Januar 2017



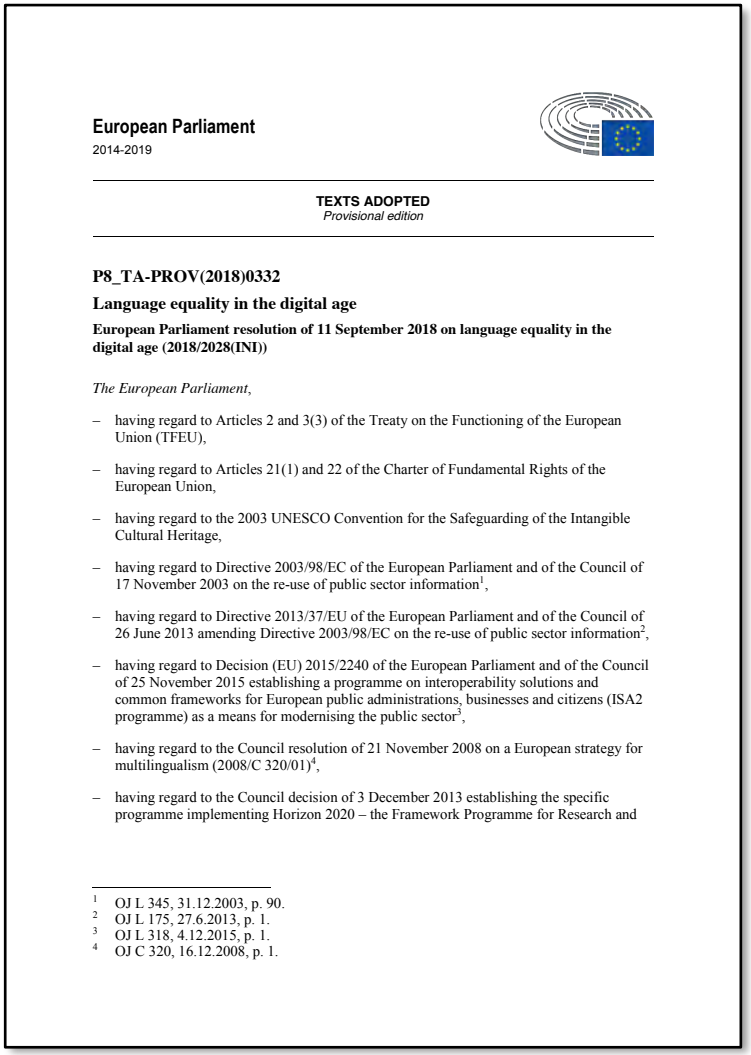
- STOA-Studie – veröffentlicht im März 2017
- Empfiehlt der EC, das Human Language Project (HLP) zu initiieren
- Drei wesentliche Vorschläge im Bereich Forschungspolitik:
  - Forschung stärken und auf das HLP fokussieren
  - Europäische LT-Plattform von Daten und Services aufbauen
  - Technologiekluft zwischen den europäischen Sprachen überbrücken

STOA
Science and Technology Options Assessment

European Parliament

“Language equality”-Resolution

- EP-Resolution “Language equality in the digital age” P8_TA(2018)0332 – basiert partiell auf der STOA-Studie
- Abstimmung im EP am 11. Sept. 2018: 592 Ja- vs. 45 Nein-Stimmen!
- **Unsere aktuellen Initiativen adressieren sehr viele der 45 Empfehlungen**
  - 25. **Establish a large-scale, long-term LT funding programme**
  - 27. Europe has to secure its leadership in language-centric AI
  - 29. **Create a European LT platform for sharing of services**
  - 31. Recommends an update of the META-NET white paper series



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.

META FORUM 2015

META SHARE

META FORUM 2013

META FORUM 2012

METANET4U

META FORUM 2011

META NET

2018

HLP Prep Final Proposal Submission

Successful submission of proposal on September 18, 2018.

2017

Strategic Research and Innovation Agenda (Version 1.0)

Launch of the Strategic Research and Innovation Agenda titled "Language Technologies for Multilingual Europe – Towards a Human Language Project" (December 2017)

"Language equality in the digital age: Towards a Human Language Project"

Launch of the study on "Language equality in the digital age: Towards a Human Language Project", commissioned by the EU Parliament (March 2017)

2016

Strategic Research and Innovation Agenda (Version 0.9)

Launch of the Strategic Research and Innovation Agenda titled "Language as a Data Type and Key Challenge for Big Data" (July 2016)

2015

Riga Summit on the Multilingual Digital Single Market

Summit carrying the motto "Shape the future of the multilingual digital single market" (April 27–29, 2015)

META-FORUM 2015

Conference "Technologies for the Multilingual Digital Single Market" (Riga – April 27, 2015)

CRACKER

EU-funded project CRACKER (Horizon2020) pushing towards an improvement of MT research in terms of efficiency and effectiveness (2015 – 2017)

META-SHARE

Initiated in 2013, META-SHARE has functioned as an open and secure network of repositories for sharing and exchanging language data, tools and related web services.

META-FORUM 2013

Conference "Connecting Europe for New Horizons", (Berlin –September 19/20, 2013)

META-FORUM 2012

Conference "A Strategy for Multilingual Europe" (Brussels – June 20/21, 2012)

METANET4U

EU-funded project (ICT Policy Support Programme) to enhance the European Linguistic Infrastructure (2011 – 2013)

META-FORUM 2011

Conference "Solutions for Multilingual Europe" (Budapest – June 27/28, 2011)

META-NET

Established in 2010, META-NET is a network of Excellence consisting of 60 research centres from 34 countries, dedicated to building the technological foundations of a multilingual European information society

?

EP on report "Language equality in the digital age"

The European Parliament voted upon the joint CULT/ITRE report on "Language equality in the digital age" with 592 votes in favour, 45 against and 44 abstentions (September 18, 2018)

META-FORUM 2017

Conference "Towards a Human Language Project" (Brussels – November 13/14, 2017)

"Language Equality in the Digital Age"

Workshop on "Language Equality in the Digital Age", commissioned by the EU Parliament's Science and Technology Options Assessment Committee (STOA) (January 2017)

META-FORUM 2016

Conference "Beyond Multilingual Europe" (Lisbon – July 4/5, 2016)

Strategic Agenda for the Multilingual Digital Single Market (Version 0.5)

Launch of the Strategic Agenda for the Multilingual Digital Single Market titled "Technologies for Overcoming Language Barriers towards a truly integrated European Online Market" (April 2015)

Cracking the Language Barrier Federation

Founded in 2015, the federation has been assembling European research and innovation projects as well as all related community organisations working on multilingual technologies

2013

"State of the Art of Machine Translation – Current Challenges and Future Opportunities"

Workshop on "State of the Art of Machine Translation", commissioned by the EU Parliament (December 2013)

Strategic Research Agenda for Multilingual Europe 2020

Launch of the Strategic Research Agenda for Multilingual Europe 2020 (January 2013)

2012

META-NET White Papers

Release of 32 volumes on 31 languages, revealing that there is a severe threat of digital extinction for at least 21 European languages (December 2012)

META-NORD

EU-funded project META-NORD (ICT Policy Support Programme) establishing an open linguistic infrastructure in the Baltic and Nordic countries (2011 – 2013)

2011

CESAR

EU-funded project (ICT Policy Support Programme) functioning as a part of META-NET to standardise language resources and tools (2011 – 2013)

2010

META-FORUM 2010

Conference "Challenges for Multilingual Europe" (November 17/18, 2010)

T4ME

EU-funded project (Seventh Framework Programme) working on technologies for the Multilingual European Information Society (2010 – 2013)

META FORUM 2017

META FORUM 2016

Cracking the Language Barrier

META NORD

CESAR

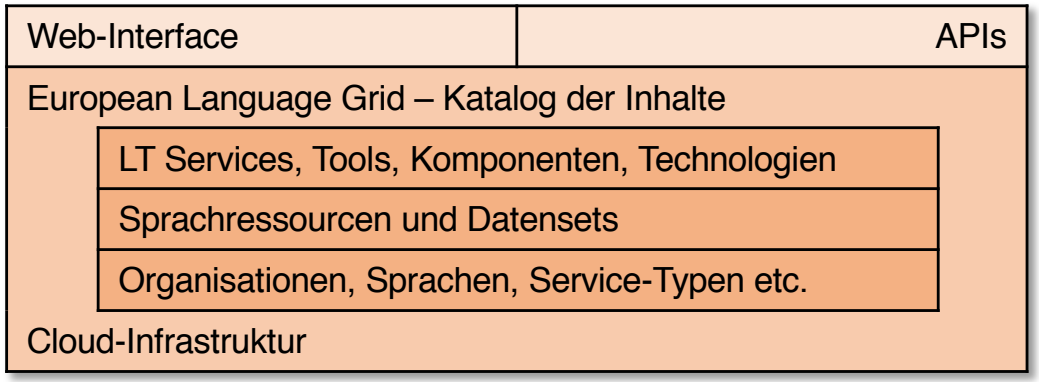
META FORUM 2010

T4ME

ELG – Die Sprachtechnologie-Cloud-Plattform für Europa

1 Jan. 2019 – 31 Dez. 2021

- **Entwicklung einer funktionalen Sprachtechnologie-Cloud-Plattform für Europa**
- Marktplatz für die europäische Sprachtechnologie-Industrie
- Hunderte von LT-Services und -Ressourcen für alle europäischen Sprachen
- 30+ nationale Kompetenzzentren als starkes europäisches Netzwerk
- Stärkung des mehrsprachigen digitalen Binnenmarktes
- Evaluation durch 15-20 Pilotprojekte
- Interoperabilität von Services durch Containerisierung
- Gründung eines Spin-offs, um den langfristigen Betrieb und das Wachstum der Plattform zu gewährleisten



Konsortium

- DFKI GmbH (Koordinator) (DE)
- ILSP, R.C. “Athena“ (GR)
- Universität Sheffield (UK)
- Charles Universität (CZ)
- ELDA (FR)
- Tilde (LV)
- SAIL LABS GmbH (AT)
- Expert System Iberia (ES)
- Universität Edinburgh (UK)

- ICT-29-2018: **Multilingual Next Generation Internet**
  - Zwei Sub-Topics (Gesamtbudget 25M€)
- ICT-29 a) **European Language Grid**
  - Eine Innovation Action (7M€)
- ICT-29 b) **Domain-specific/challenge-oriented HLT**
  - Sechs Research and Innovation Actions (je 3M€)

Human Language Project

- **Ziel: Tiefes maschinelles Sprachverstehen bis 2030**
- **Alle offiziellen europäischen und viele weitere Sprachen**
- **Breite Abdeckung, hohe Qualität, hohe Präzision**
- **Ganz neue Ansätze, Algorithmen, Datensets**
- **Alle Modalitäten:** Text, Texttypen, Speech, Video etc.
- **Alle Plattformen:** Messaging, Telefonie, Social, Mobil, IoT, Roboter, smarte Geräte, persönliche Assistenten etc.
- **Kulturübergreifend:** Wissen, Bräuche, Formalitäten, Humor, Emotion, Subjektivität, Meinungen, Filterblase etc.
- **Wie? Als das nächste EU FET Flagship-Projekt!**

HLP Prep Antrag

- FET Flagship Projekte: 1 Mrd. € Förderung für 10 Jahre
- Aktuelle Flagships: HBP, Graphene, Quantum (2019)
- H2020 FETFLAG-01-2018 für kleine Projekte zur Vorbereitung von Anträgen (1M€, 12 Monate Laufzeit)
- Unser Antrag: “Human Language Project Preparation”
- Konsortium mit 16 Partnern, koordiniert vom DFKI
- Bislang 375+ Unterstützungsbriefe, inklusive 16 Ministerien und 24 nationalen Sprachinstitutionen
- Mehr Informationen: http://human-language-project.eu

HLP Prep

Das HLP ist ein großes, langfristig angelegtes Forschungs-, Entwicklungs- und Innovationsprogramm, in dem Grundlagen- und angewandte Forschung und Entwicklung sowie Innovation und Kommerzialisierung eng zusammen arbeiten, um bahnbrechende Technologien für das ***tiefe maschinelle Sprachverstehen bis 2030*** zu entwickeln.

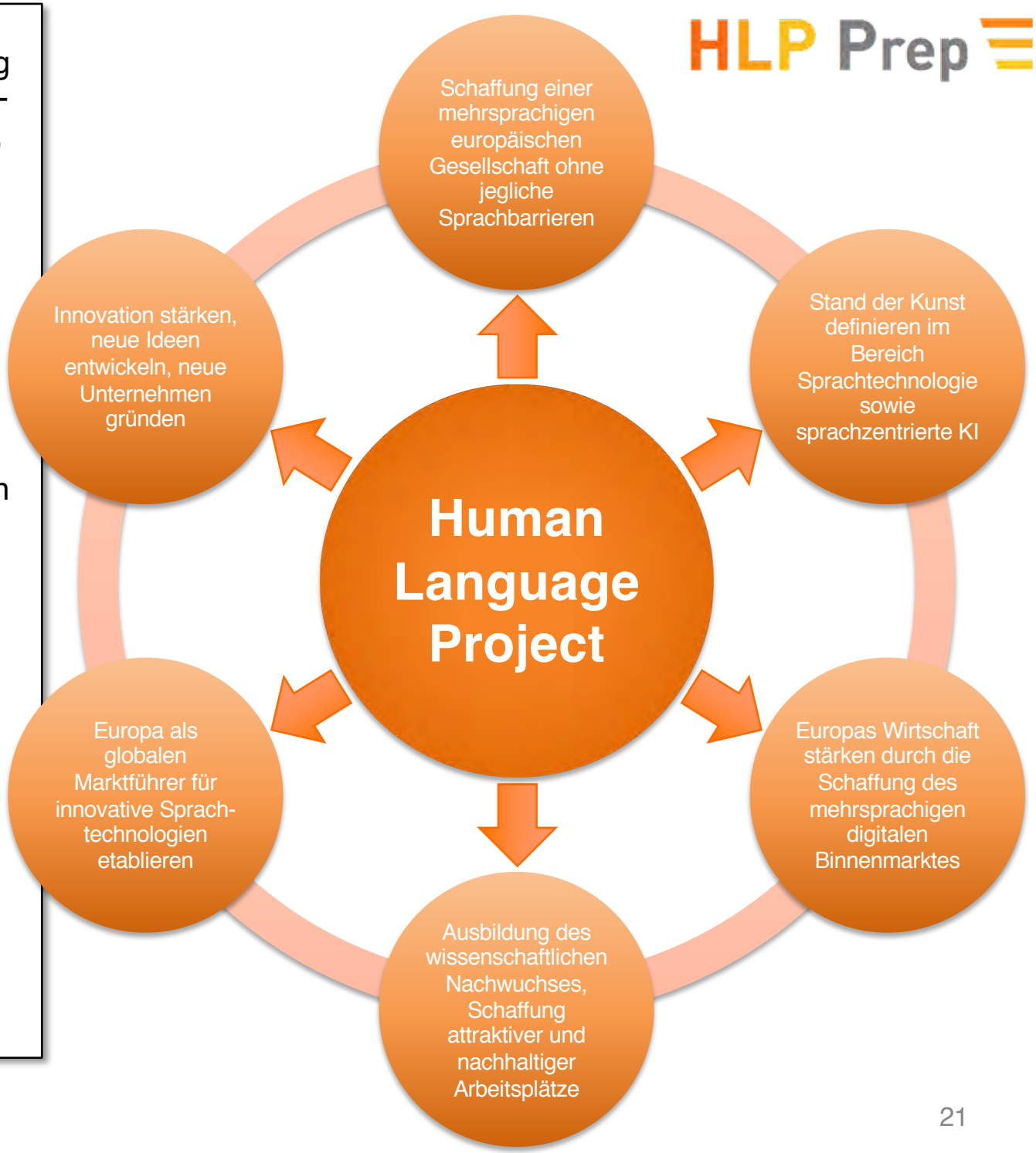
Im HLP werden u.a. die folgenden Gebiete kollaborieren:

Primäre Gebiete:

- Computerlinguistik & LT
- Linguistik
- Künstliche Intelligenz
- Wissenstechnologien

Sekundäre Gebiete:

- Gesellschaftswissenschaften und Digital Humanities
- Informatik
- Kognitionswissenschaft



HLP Prep

HLP Core Project
- Koordination des Flagships (CP und PPs)
- Weiterentwicklung der Roadmap
- Allgemeine Forschung und Technologieentwicklung
- Daten, Ressourcen, Computing- und Kollaborationsinfrastruktur

HLP Core Project

PP

PP

HLP Partnering Project: Baltische Sprachen

HLP Partnering Project: Nieder-ländisch

HLP Partnering Project: Polnisch

HLP Partnering Project: Deutsch

HLP Partnering Project: Spanisch

PP

PP

HLP Partnering Project: Grie-chisch

HLP Partnering Project: Italienisch

European Parliament

HLP Partnering Projects
- Sprachspezifische und/oder regionale Konsortien, die Forschung für ihre eigenen Sprachen betreiben
- Enge Kooperation mit dem Core Project
- CP und PP teilen sich einige Partner

Wichtige Daten und nächste Schritte
- 11. Sep. 2018: EP “Language equality” Beschluss
- 18. Sep. 2018: HLP Prep Hauptantrag eingereicht
- 27. Sep. 2018: “Language equality” Konferenz (EP)
- 29. Nov. 2018: Bekanntgabe der Ergebnisse
- 04. Dez. 2018: Ankündigung der Gewinner bei der ICT 2018
- 01. Mrz. 2019: Ggf. Start des HLP Prep Vorbereitungsprojekts
- 29. Feb. 2020: Ggf. Ende des HLP Prep Vorbereitungsprojekts

1. Herausforderung

Das mehrsprachige Europa:
Sprachtechnologien für alle europäischen Sprachen?

2. Herausforderung

Online-Desinformationskampagnen:
Technische Lösungsansätze gegen „Fake News“?

3. Herausforderung

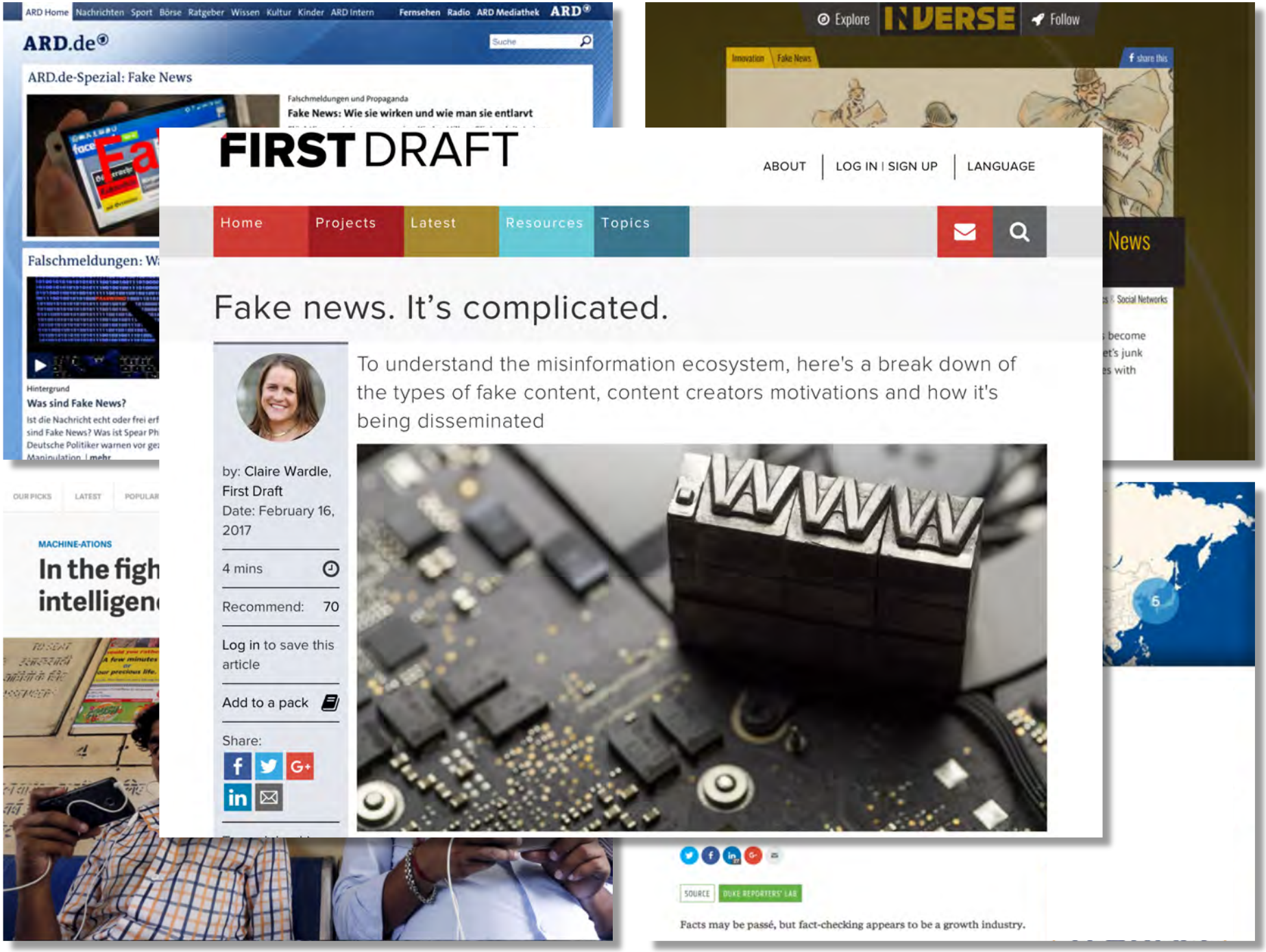
Digitaler Content:
Technologien für die effiziente Content-Kuratierung?


ARD Home Nachrichten Sport Börse Ratgeber Wissen Kultur Kinder ARD Intern Fernsehen Radio ARD Mediathek ARD
ARD.de
Suche
ARD.de-Spezial: Fake News
Fake
Falschmeldungen und Propaganda
Fake News: Wie sie wirken und wie man sie entlarvt
Flüchtlinge urinieren gegen eine Kirche, Hillary Clinton leitet einen Kinderporno-Ring und es gehen Koran-CDs mit Gift um. Das sind Fake News - also Lügenmärchen, die gezielt verbreitet werden. Sie beeinflussen das gesellschaftliche Klima und können sich auf Wahlen auswirken. Auch für die kommende Bundestagswahl wird mit Fake News und Social Bots gerechnet. Das ARD.de-Spezial zeigt, wer dahinter steckt, was Fake News bewirken und wie man sie erkennt.
Falschmeldungen: Was steckt dahinter?
Hintergrund
Was sind Fake News?
Ist die Nachricht echt oder frei erfunden? Was sind Fake News? Was ist Spear Phishing? Deutsche Politiker warnen vor gezielter Manipulation. | mehr
Wie eine Ente im Netz entsteht
Das Geschäft mit den Fake News
Die Falschmeldung, der Papst unterstütze Donald Trump, wurde 960.000 Mal auf Facebook geteilt. Wer sind die Menschen, die gezielt unwahre Artikel streuen? | hr
Wahlkampf
Die Macht der Social Bots
Computerprogramme, die in sozialen Medien wie Nutzer agieren, mischten schon im Ukraine-Konflikt, dem Brexit-Referendum und im US-Wahlkampf mit. | dradio


Explore
INVERSE
Follow
Innovation
Fake News
share this
Artificial Intelligence is Going to Destroy Fake News
But A.I. can also cause the volume of fake news to explode.
By Rosalie Chan on February 21, 2017
Filed Under AI, Politics & Social Networks
With the rise of email came the rise of spam filling inboxes. Email has become sophisticated faster than spamming technology and now, the internet's junk mail is often caught in a folder; out of sight and out of mind are messages with the subject line "Kindly get back to me urgently" and the greeting "Dear Beneficiary."
The worlds may be virtual
but the prize is very real.

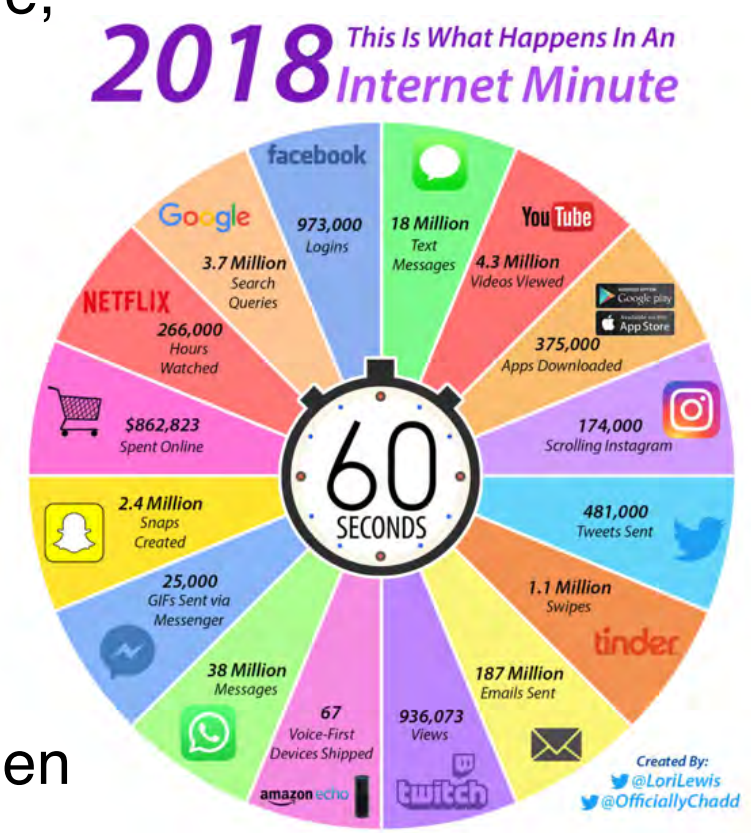

OUR PICKS
LATEST
POPULAR
QUARTZ
OBSESSIONS
MACHINE-ATIONS
In the fight against fake news, artificial intelligence is waging a battle it cannot win


9
34
8
34
5
2
2
9
FACT-CHECKING
There are now 114 fact-checking initiatives in 47 countries
By Alexios Mantzarlis • February 28, 2017
SOURCE
DUKE REPORTERS' LAB
Facts may be passé, but fact-checking appears to be a growth industry.


ARD Home Nachrichten Sport Börse Ratgeber Wissen Kultur Kinder ARD Intern Fernsehen Radio ARD Mediathek ARD
ARD.de
ARD.de-Spezial: Fake News
Falschmeldungen und Propaganda
Fake News: Wie sie wirken und wie man sie entlarvt
Falschmeldungen: W
Hintergrund
Was sind Fake News?
Explore INVERSE Follow
Innovation Fake News
share this
News
Social Networks
become
et's junk
es with
FIRST DRAFT
ABOUT LOG IN | SIGN UP LANGUAGE
Home Projects Latest Resources Topics
Fake news. It's complicated.
To understand the misinformation ecosystem, here's a break down of the types of fake content, content creators motivations and how it's being disseminated
by: Claire Wardle, First Draft
Date: February 16, 2017
4 mins
Recommend: 70
Log in to save this article
Add to a pack
Share:
OUR PICKS LATEST POPULAR
MACHINE-ATIONS
In the figh
intelligen
SOURCE DUKE REPORTERS' LAB
Facts may be passé, but fact-checking appears to be a growth industry.

Relevanz Digitaler Medien

- Immens zunehmende (global)politische, gesellschaftliche und ökonomische Relevanz
- Facebook: >2 Mrd. Nutzer
- WhatsApp: >1 Mrd. Nutzer
- Instagram: >1 Mrd. Nutzer
- Öffentliche Debatten finden in erster Linie online statt.
- Diskussionen zu aktuellen Themen, Parteien, Wahlen, Personen etc. werden durch Social-Media-Kampagnen sehr geschickt beeinflusst und manipuliert.



Viralität und „Fake News“

- Inhalte werden ohne Kontrollinstanz publiziert, über soziale Medien entdeckt und, falls relevant, zügig geteilt
- Dies geschieht oft *ohne* Lektüre oder kritische Prüfung
- Ziel: Viralität ⇢ Reichweite ⇢ Klicks ⇢ Werbeeinnahmen
- Ziel: die Meinung bestimmter Personen manipulieren
- Nicht alle „journalistisch“ aussehenden Inhalte fühlen sich tatsächlich der Wahrheit verpflichtet
- Bürde der kritischen Prüfung liegt heute bei den Lesern
- „Fake News“: Etikett für mehrere Klassen von Inhalten

Georg Rehm. “An Infrastructure for Empowering Internet Users to handle Fake News and other Online Media Phenomena”. In Georg Rehm and Thierry Declerck, editors, *Language Technologies for the Challenges of the Digital Age: Proceedings of the GSCL Conference 2017*, Berlin, September 2017. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V. 13.-15. September 2017.

Sieben Klassen von Falsch-nachrichten

		Satire oder Parodie: führt Menschen aber dennoch manchmal auf das Glatteis	Falscher Zusam-menhang: wenn Titel und Fotos den Inhalt nicht stützen	Irreführen-der Inhalt: Nutzung von Informatio-nen, um etwas/jmd. In ein schlechtes Licht zu rücken	Falscher Kontext: wenn echte Inhalte im falschen Kontext präsentiert werden	Betrüge-rische Inhalte: wenn echte Quellen imitiert werden	Manipulier-ter Inhalt: Manipula-tion von Inhalten zum Zweck der Täuschung	Fabrizierter Inhalt: basiert zu 100% nicht auf Tatsachen, geschrie-ben um zu täuschen
Charakteristika	Clickbait		X	X	?		?	?
	Desinformation			X	X		X	X
	Politisch gefärbt		?	X	?		?	X
	Schlechter Journalismus		X	X	X			
Intentionen der Urheber	Parodie	X				?		?
	Provokation					X	X	X
	Profit	?	X			X		X
	Täuschung		X	X	X	X	X	X
	Politik beeinflussen			X	X		X	X
	Meinungen beeinflussen			X	X	X	X	X

Unterschiedliche Klassen von Falschnachrichten und ihre jeweiligen Charakteristika und Intentionen (nach Wardle, 2017; Walbrühl, 2017; Rubin et al., 2015; Holan, 2016; Weedon et al., 2017)

		Satire oder Parodie: führt Menschen aber dennoch manchmal auf das Glatteis	Falscher Zusam-menhang: wenn Titel und Fotos den Inhalt nicht stützen	Irreführen-der Inhalt: Nutzung von Informatio-nen, um etwas/jmd. In ein schlechtes Licht zu rücken	Falscher Kontext: wenn echte Inhalte im falschen Kontext präsentiert werden	Betrüge-rische Inhalte: wenn echte Quellen imitiert werden	Manipulier-ter Inhalt: Manipula-tion von Inhalten zum Zweck der Täuschung	Fabrizierter Inhalt: basiert zu 100% nicht auf Tatsachen, geschrie-ben um zu täuschen
Charakteristika	Clickbait		X	X	?		?	?
	Desinformation			X	X		X	X
	Politisch gefärbt		?	X	?		?	X
	Schlechter Journalismus							
Intentionen der Urheber	Parodie	X						
	Provokation					X	X	X
	Profit	?	X			X		X
	Täuschung		X	X	X	X	X	X
	Politik beeinflussen			X	X		X	X
	Meinungen beeinflussen			X	X	X	X	X

Parodie

Clickbait
Schlechter Journalismus

Desinformation
Bewusste Täuschung
Versuch der Beeinflussung von Meinungen und Politik

Unterschiedliche Klassen von Falschnachrichten und ihre jeweiligen Charakteristika und Intentionen (nach Wardle, 2017; Walbrühl, 2017; Rubin et al., 2015; Holan, 2016; Weedon et al., 2017)

Beispiel 1: Clickbait-Erkennung

- Automatische Prüfung arbiträrer Behauptungen bis auf Weiteres technisch nicht möglich
- Annäherung: Ermittlung der Haltung eines Textes zu einem Thema („Stance Detection“)

Annotierte Titel/Artikel-Paare	49.972	100%	
Klasse: *unrelated*	36.545	73%	**Schritt 1:** Klassifikation *related* vs. *unrelated* = Clickbait-Erkennung
Klasse: *discuss*	8.909	18%	**Schritt 2:** Nur wenn sich der Titel auf den Text *bezieht*, kann *discuss*, *agree*, *disagree* klassifiziert werden.
Klasse: *agree*	3.678	7%	
Klasse: *disagree*	840	2%	

DFKI-System	Relatedness	93,29
	Drei Klassen	88,36
	Gewichtet	**89,59**

Mit einer Präzision von 89,59 haben wir bei der ersten Fake News Challenge (FNC1) Platz 9 von 50 Teams erreicht.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. “From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles”. In Octavian Popescu and Carlo Strapparava, editors, *Proceedings of Natural Language Processing meets Journalism – EMNLP 2017 Workshop (NLPMJ 2017)*, Copenhagen, Denmark, September 2017. 7. September.

Beispiel 2: Beleidigende Sprache

- Beleidigende Beiträge verhindern konstruktive Online-Debatten
- Klassifikationsexperimente mit verschiedenen Datensets
- Englische Tweets: *neutral*, *rassistisch*, *sexistisch*
- Deutsche Tweets: *hasserfüllt* vs. *nicht hasserfüllt*
- Wikipedia-Talk-Seiten mit Nutzerkommentaren
  - A1: *Angriff auf eine Person* vs. *kein Angriff auf eine Person*
  - A2: *Aggression* vs. *keine Aggression*

	Tweets EN (15.979)	Tweets DE (469)	Wikipedia A1 (11.304)	Wikipedia A2 (11.304)
Precision	85,67	78,19	80,90	80,42
Recall	77,45	78,16	80,97	80,46

Schlussfolgerungen: Viel versprechende Ergebnisse, allerdings stellt die Aufgabe eine große Herausforderung dar – Teilnahme an SemEval 2019.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. "Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication". In Georg Rehm and Thierry Declerck, editors, *Language Technologies for the Challenges of the Digital Age: Proceedings of the GSCL Conference 2017*, Berlin, September 2017. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V. 13.-15. September 2017.

Julian Moreno Schneider, Roland Roller, Peter Bourgonje, Stefanie Hegele, and Georg Rehm. "Towards the Automatic Classification of Offensive Language and Related Phenomena in German Tweets". In Josef Ruppenhofer, Melanie Siegel, and Michael Wiegand, editors, Proceedings of the GermEval Workshop 2018 – Shared Task on the Identification of Offensive Language, pages 95-103, Vienna, Austria, September 2018. 21 September 2018.

Lösungsvorschlag: Infrastruktur

- Inhalte werden im/über das World Wide Web konsumiert
- Ziel: Leser im Umgang mit Inhalten unterstützen, Fakten prüfen, Täuschungsversuche erkennen etc.
- Im Browser z.B. Ampelmetaphorik: Rot, Gelb, Grün
- Somit Filterblasen- und Netzwerkeffekte ausbalancieren
- Kombination aus automatischen Werkzeugen und menschlicher Schwarmintelligenz
- 2019: Kollaboration mit W3C Credible Web CG
- 2019: Re-submission des EU-Antrags PROTECT-IT

Georg Rehm. “An Infrastructure for Empowering Internet Users to handle Fake News and other Online Media Phenomena”. In Georg Rehm and Thierry Declerck, editors, *Language Technologies for the Challenges of the Digital Age: Proceedings of the GSCL Conference 2017*, Berlin, September 2017. 13.-15.09.2017.

Georg Rehm, Julian Moreno Schneider, and Peter Bourgonje. “Automatic and Manual Web Annotations in an Infrastructure to handle Fake News and other Online Media Phenomena.” In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), pages 2416-2422, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Dezentrale Filter verarbeiten Inhalte automatisch und schicken Ergebnisse zum Browser (wichtig: Mehrsprachigkeit)

Website mit Inhalten

Service$_1$

Detektion von Hassrede

Klassifikation des politischen Spektrums

Faktenchecker

Service$_2$

Service$_3$

MGM

MGM

MGM

UGA

Web Annotations DB$_1$

andere Nutzer

UGM

UGM

Web Annotations DB$_2$

UGA

Dezentrale Repositorien speichern alle Annotationen

Web Annotations DB$_3$

UGA

Annotationen anderer Nutzer

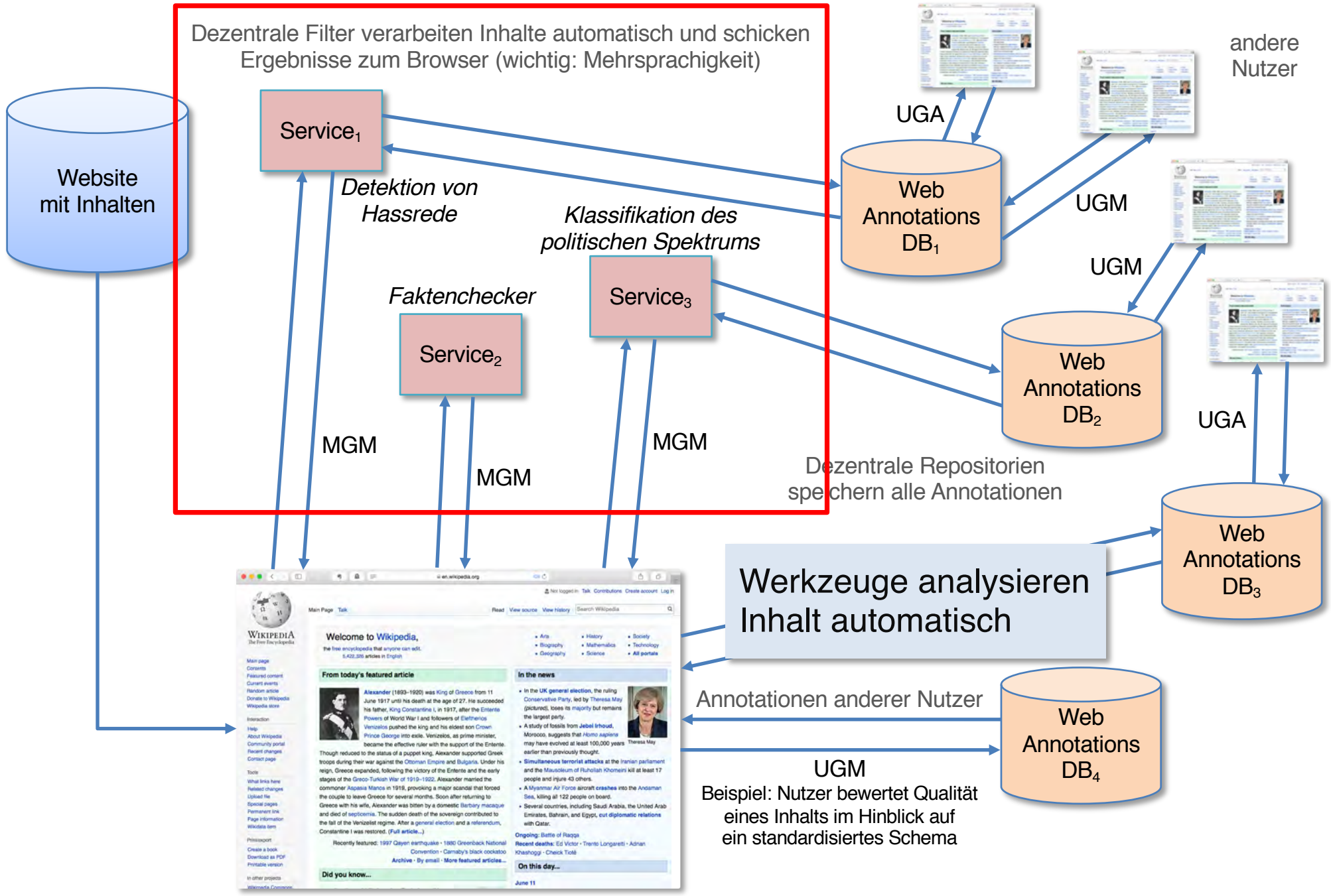
UGM

Beispiel: Nutzer bewertet Qualität eines Inhalts im Hinblick auf ein standardisiertes Schema

Web Annotations DB$_4$

Browser unterstützt Infrastruktur nativ und aggregiert unterschiedlichen Bewertungen, Kommentare und Meinungen über einen Inhalt in klare Botschaften oder Warnungen

UGA: User-generierte Annotationen (Freitext)
UGM: User-generierte Metadaten (standardisiert)
MGM: Maschinen-generierte Metadaten (standardisiert)


Dezentrale Filter verarbeiten Inhalte automatisch und schicken Ergebnisse zum Browser (wichtig: Mehrsprachigkeit)
Website mit Inhalten
Service1
Detektion von Hassrede
Klassifikation des politischen Spektrums
Faktenchecker
Service2
Service3
MGM
MGM
MGM
UGA
UGM
UGM
UGA
andere Nutzer
Web Annotations DB1
Web Annotations DB2
Web Annotations DB3
Web Annotations DB4
Dezentrale Repositorien speichern alle Annotationen
Werkzeuge analysieren Inhalt automatisch
Annotationen anderer Nutzer
UGM
Beispiel: Nutzer bewertet Qualität eines Inhalts im Hinblick auf ein standardisiertes Schema
Browser unterstützt Infrastruktur nativ und aggregiert unterschiedlichen Bewertungen, Kommentare und Meinungen über einen Inhalt in klare Botschaften oder Warnungen
UGA: User-generierte Annotationen (Freitext)
UGM: User-generierte Metadaten (standardisiert)
MGM: Maschinen-generierte Metadaten (standardisiert)

- Maschinelle Ergebnisse sowie auch Freitext-Anmerkungen werden als W3C Web Annotations gespeichert.

sch und schicken
sprachigkeit)

andere Nutzer

Website mit Inhalten

Detektion von Hassrede

Klassifikation des politischen Spektrums

Faktenchecker

Service$_2$

Service$_3$

UGA

UGM

UGM

Web Annotations DB$_1$

Web Annotations DB$_2$

Web Annotations DB$_3$

Web Annotations DB$_4$

MGM

MGM

MGM

UGA

Dezentrale Repositorien speichern alle Annotationen

UGA

Annotationen anderer Nutzer

UGM

Beispiel: Nutzer bewertet Qualität eines Inhalts im Hinblick auf ein standardisiertes Schema

Browser unterstützt Infrastruktur nativ und aggregiert unterschiedlichen Bewertungen, Kommentare und Meinungen über einen Inhalt in klare Botschaften oder Warnungen

UGA: User-generierte Annotationen (Freitext)
UGM: User-generierte Metadaten (standardisiert)
MGM: Maschinen-generierte Metadaten (standardisiert)

Dezentrale Filter verarbeiten Inhalte automatisch und schicken Ergebnisse zum Browser (wichtig: Mehrsprachigkeit)

andere Nutzer

Website mit Inhalten

- Maschinelle Analyse von Freitext-Annotationen (NLP, IE, RE etc.).
- Extraktion von Meinungen, Argumenten, Behauptungen etc.

UGA

Web Annotations DB_1

UGM

UGM

Web Annotations DB_2

UGA

Web Annotations DB_3

Service$_2$

MGM

MGM

MGM

Dezentrale Repositorien speichern alle Annotationen

UGA

Annotationen anderer Nutzer

Web Annotations DB_4

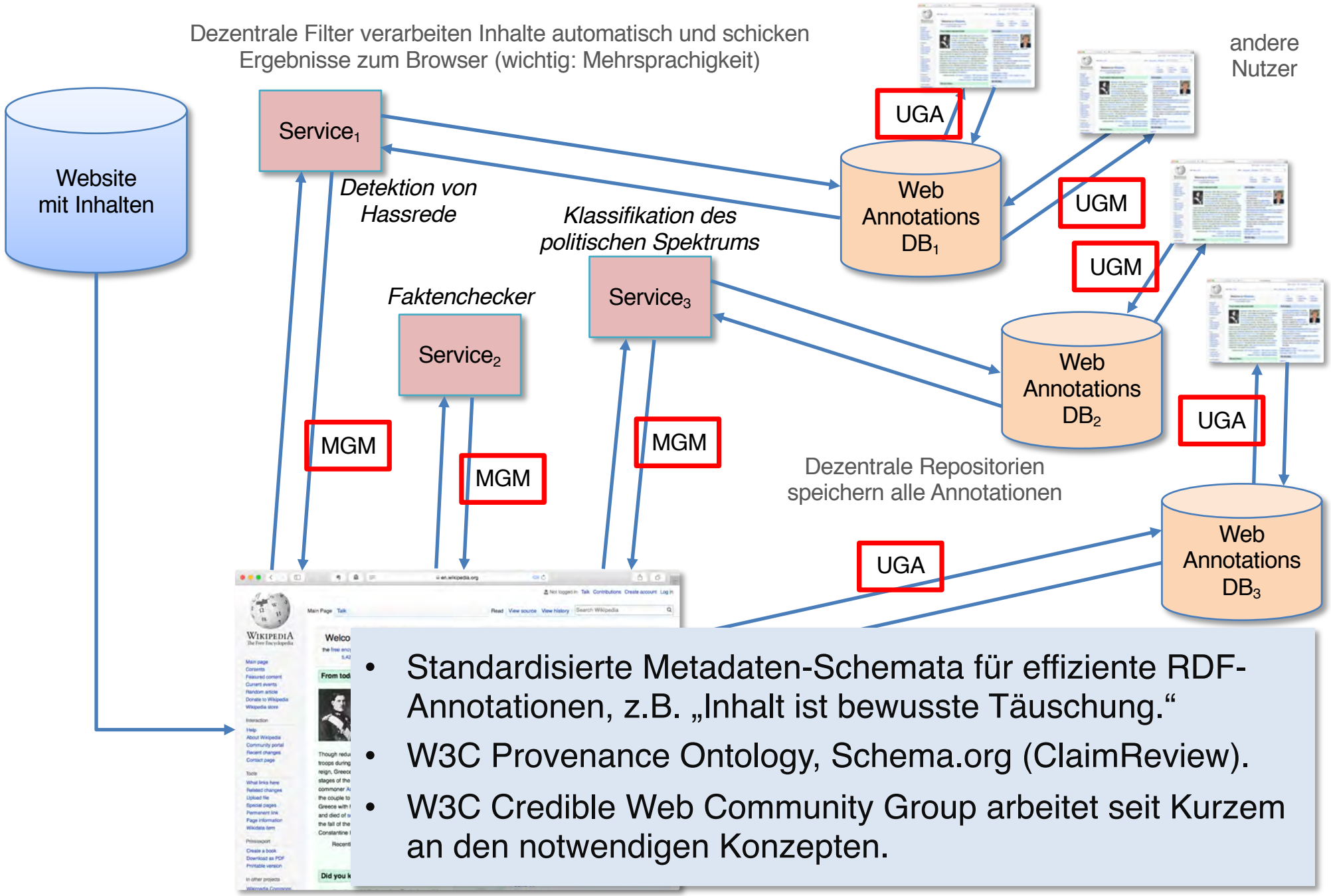
UGM

Beispiel: Nutzer bewertet Qualität eines Inhalts im Hinblick auf ein standardisiertes Schema

Browser unterstützt Infrastruktur nativ und aggregiert unterschiedlichen Bewertungen, Kommentare und Meinungen über einen Inhalt in klare Botschaften oder Warnungen

UGA: User-generierte Annotationen (Freitext)
UGM: User-generierte Metadaten (standardisiert)
MGM: Maschinen-generierte Metadaten (standardisiert)

Dezentrale Filter verarbeiten Inhalte automatisch und schicken Ergebnisse zum Browser (wichtig: Mehrsprachigkeit)



- Standardisierte Metadaten-Schemata für effiziente RDF-Annotationen, z.B. „Inhalt ist bewusste Täuschung.“
- W3C Provenance Ontology, Schema.org (ClaimReview).
- W3C Credible Web Community Group arbeitet seit Kurzem an den notwendigen Konzepten.

Browser unterstützt Infrastruktur nativ und aggregiert unterschiedlichen Bewertungen, Kommentare und Meinungen über einen Inhalt in klare Botschaften oder Warnungen

UGA: User-generierte Annotationen (Freitext)
UGM: User-generierte Metadaten (standardisiert)
MGM: Maschinen-generierte Metadaten (standardisiert)

Dezentrale Filter verarbeiten Inhalte automatisch und schicken Ergebnisse zum Browser (wichtig: Mehrsprachigkeit)

andere Nutzer

Website mit Inhalten

$Service_1$

UGA

Web Annotations DB_1

Detektion von Hassrede

Klassifikation des politischen Spektrums

UGM

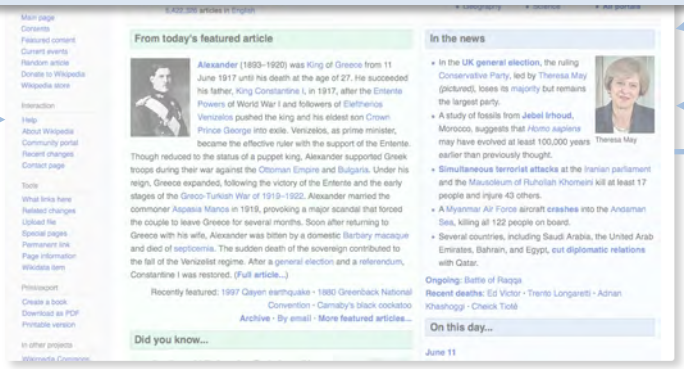
UGM

Faktenchecker

$Service_3$

Ziel: Nutzern Technologien an die Hand geben, mit denen sie digitale Medien besser rezipieren, analysieren, verifizieren und einschätzen können und die automatisch Hinweise auf problematische Inhalte liefern.

Web Annotations DB_3



Annotationen anderer Nutzer

Web Annotations DB_4

UGM

Beispiel: Nutzer bewertet Qualität eines Inhalts im Hinblick auf ein standardisiertes Schema

Browser unterstützt Infrastruktur nativ und aggregiert unterschiedlichen Bewertungen, Kommentare und Meinungen über einen Inhalt in klare Botschaften oder Warnungen

UGA: User-generierte Annotationen (Freitext)
UGM: User-generierte Metadaten (standardisiert)
MGM: Maschinen-generierte Metadaten (standardisiert)

1. Herausforderung

Das mehrsprachige Europa:
Sprachtechnologien für alle europäischen Sprachen?

2. Herausforderung

Online-Desinformationskampagnen:
Technische Lösungsansätze gegen „Fake News“?

3. Herausforderung

Digitaler Content:
Technologien für die effiziente Content-Kuratierung?

Relevanz von Content

- Content spielt wichtige Rolle in Gesellschaft und Industrie
- In vielen Branchen herrscht Druck, regelmäßig Content zu publizieren; man spricht auch von der Content-Industrie.
- Ziel: **Smarter, effizienter Umgang mit Content**

  Beschaffung, Konvertierung, Anreicherung, Analyse, Zusammenfassung, Übersetzung, Verknüpfung, Zusammenstellung und Publikation von Content.

  Ausspielen von Content auf div. Kanälen inkl. Social Media.

- Riesiges Potenzial für Disruptionen in allen Branchen durch **Technologien für die Kuratierung von Content.**

Was ist digitale Kuratierung?

Input **Prozesse** **Software** **Output**

Branchen

Input	Prozesse	Software	Output
Tweet	Analysieren	Textverarbeitung	Zeitungsartikel
Zeitungsartikel	Auswählen	Präsentationen	Multimedia-Website
Agenturmeldung	Fokussieren	Tabellenkalkulation	TV-Beitrag
Facebook-Meldung	Überarbeiten	Email	Ausstellungskatalog
Suchergebnis	Einlesen	Browser	Mobile Applikation
Email	Schreiben	Groupware	Mashup (z.B. Karte)
SMS	Gestalten	Branchenapplikationen	Textbeitrag
Konzept	**Recherchieren**	CMS	Konzept
Textdateien	**Bewerten**	ECMS	Zeitstrahl
Video	**Evaluieren**	CRM	Fachartikel
Karte	**Ordnen**	Unternehmens-Software	Studie
Stockfotos	**Sortieren**	Grafik-/Layout-Software	Präsentation
In-house Datenbank	**Strukturieren**	Telefonie	Faktensammlung
Kalendereintrag	**Zusammenfassen**	*etc.*	Exponatsartikel
Spreadsheets	**Kürzen**		Analysen
Archiv	**Übersetzen**		*etc.*
etc.	**Informieren**		
	Kombinieren		
	Abstrahieren		
	Einordnen		
	Visualisieren		
	Generieren		
	Annotieren		
	Referenzieren		
	etc.		

Merkmale und Dimensionen

- **Content:** Text, Ton, Bild, Video, Multimedia, AR/VR
- **Mehrsprachigkeit:** Text, Ton, ggf. mehrere Sprachen
- **Diverse Content-Typen:** U.a. Hunderte von Textsorten
- **Beteiligte:** Content wird oft in verteilten Teams kuratiert
- **Branchen:** Branchenspezifische Anforderungen
- **Workflows:** Flexible Komponierung und Konfigurierung
- **Services:** Spektrum – generisch bis branchenspezifisch
- **Geschwindigkeit und Effizienz:** Einsatz im Arbeitsalltag

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

ART+COM condat® Deutsches Forschungszentrum für Künstliche Intelligenz GmbH 3pc Neue Kommunikation kreuzwerker we touch running systems

DKT Digitale Kuratierungs-technologien

Digitale Kuratierungstechnologien

- Digitaler Kuratierung mit Sprach- und Wissenstechnologien
- Entwicklung innovativer Prototypen mit den KMU-Partnern
- Unterstützung der Experten – „human in the loop“!
- Weiterentwicklung der DFKI-Technologien und Transfer mittels **Plattform für digitale Kuratierungstechnologien**

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

WK · POTENZIAL UNTERNEHMEN REGION Die BMBF-Innovationsinitiative Neue Länder

DKT Auftaktveranstaltung – 25. September 2015

Georg Rehm und Felix Sasaki. “Digital Curation Technologies.” In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2016)*, Riga, Lettland, Mai 2016

Georg Rehm und Felix Sasaki. “Digitale Kuratierungstechnologien – Verfahren für die effiziente Verarbeitung, Erstellung und Verteilung qualitativ hochwertiger Medieninhalte.” In *Proceedings der Frühjahrstagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL 2015)*, S. 138-139, Duisburg, 2015

Prototypisch implementierte Plattform und Services

Output

Client nutzt das API

Client nutzt das API

Client nutzt das API

Client nutzt das API

Input

Plattform für digitale Kuratierungstechnologien

REME

Broker REST API

Kuratierungs-Workflow

Kuratierungsservice 1

Kuratierungsservice 2

Externer Service 1

Externer Service 2

DKT Digitale Kuratierungstechnologien

ART+COM

condat®

3pc Neue Kommunikation

kreuzwerker we touch running systems

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- Durch (Semi-)Automatisierung der Kuratierungsprozesse Reduktion zeitlicher und finanzieller Aufwände
- Flexible, robuste, skalierbare Services
- Interoperabilität durch generische APIs

Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring, Georg Rehm, Felix Sasaki, and Ankit Srivastava. “Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer.” In Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenić, Sören Auer, and Christoph Lange, editors, The Semantic Web, number 9989 in LNCS, pages 65-68. Springer, June 2016. ESWC 2016 Satellite Events. Heraklion, Crete, Greece, May 29 - June 2, 2016 Revised Selected Papers.

NLP Interchange Format (NIF)

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix geo:    <http://www.w3.org/2003/01/geo/wgs84_pos/> .
@prefix nif:    <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://link.omitted/documents/document1#char=0,26>
  a                       nif:RFC5147String , nif:String , nif:Context ;
  nif:beginIndex          "0"^^xsd:nonNegativeInteger ;
  nif:endIndex            "26"^^xsd:nonNegativeInteger ;
  nif:isString            "Welcome to Berlin in 2016. "^^xsd:string ;
  dfkinif:averageLatitude "52.516666666666666"^^xsd:double ;
  dfkinif:averageLongitude "13.383333333333333"^^xsd:double ;
  dfkinif:stdDevLatitude  "0.0"^^xsd:double ;
  dfkinif:stdDevLongitude "0.0"^^xsd:double ;
  nif:meanDateRange       "20160101010000_20170101010000"^^xsd:string .

<http://link.omitted/documents/document1#char=21,25>
  a                       nif:RFC5147String , nif:String ;
  itsrdf:taldentRef       <http://link.omitted/ontologies/nif#date=20160101000000_20170101000000> ;
  nif:anchorOf            "2016"^^xsd:string ;
  nif:beginIndex          "21"^^xsd:nonNegativeInteger ;
  nif:endIndex            "25"^^xsd:nonNegativeInteger ;
  nif:entity              <http://link.omitted/ontologies/nif#date>.

<http://link.omitted/documents/#char=11,17>
  a                       nif:RFC5147String , nif:String ;
  nif:anchorOf            "Berlin"^^xsd:string ;
  nif:beginIndex          "11"^^xsd:nonNegativeInteger ;
  nif:endIndex            "17"^^xsd:nonNegativeInteger ;
  itsrdf:taClassRef       <http://dbpedia.org/ontology/Location> ;
  nif:referenceContext    <http://link.omitted/documents/#char=0,26> ;
  geo:lat                 "52.516666666666666"^^xsd:double ;
  geo:long                "13.383333333333333"^^xsd:double ;
  itsrdf:taldentRef       <http://dbpedia.org/resource/Berlin> .
```

“Welcome to Berlin in 2016.”

- RDF/OWL-basiertes Format für NLP-Anwendungen
- Ermöglicht Interoperabilität
- Durch pures RDF „natürliche“ Integration von Linked-Data-Daten
- Entwickelt von der Universität Leipzig
- Plattform unterstützt neben NIF auch Web Annotations

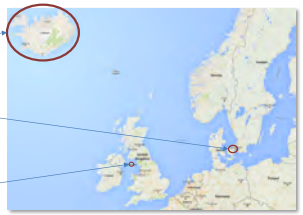
Exemplarische Basisdienste

NER, Linking, Geolokalisierung

- **Modus 1:** Modell-basiert (für Domänen, für die annotierte Trainingsdaten verfügbar sind)
- **Modus 2:** Wörterbuch-basiert (für Domänen, für die lediglich Namenslisten verfügbar sind)
- Basiert auf OpenNLP (mit NIF-Integration)

- Entity-Linking durch SPARQL-Querys auf DBPedia.
- Für Lokationen werden GPS-Koordinaten bezogen.
- Es werden Durchschnittsangaben berechnet auf Dokumentebene (über alle Lokationen), um diese auf einer Karte visualisieren zu können.

...
In the Viking colony of Iceland, an extraordinary vernacular literature blossomed in the 12th through 14th centuries
...
...
The ships were scuttled there in the 11th century, to block a navigation channel and thus protect Roskilde, then Copenhagen from seaborne assault
...
...
Viking Age inscriptions have also been discovered on the Manx runestones on the Isle of Man.
...



Geolokalisierung als visuelles Zusammenfassen!

Plain Text **NIF-Anreicherung** **Visualisierung**

http://api.digitale-kuratierung.de/api/e-nlp/namedEntityRecognition?analysis=ner

http://http://dev.digitale-kuratierung.de/admin/pages/geolocalization.php

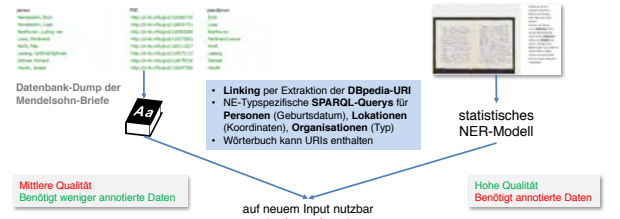
NER und Linking

A

- http://api.digitale-kuratierung.de/api/e-nlp/trainModel?analysis=dict
- Falls lediglich Listen von Namen oder Termen und deren URIs in einer Ontologie zur Verfügung stehen.

B

- http://api.digitale-kuratierung.de/api/e-nlp/trainModel?analysis=ner
- Falls annotierte Trainingsdaten zur Verfügung stehen



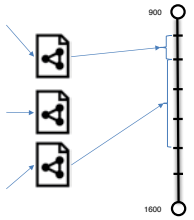
C

- Falls – z.B. bei Spezialdomänen – weder das eine (A) noch das andere (B) vorliegt, können potenzielle Entitäten in Kollektionen berechnet werden.
- Diese Liste kann vom Wissensarbeiter überprüft und anschließend als Wörterbuch (A) eingesetzt werden.

Mittlere Qualität
Menschliche Intervention notwendig
Benötigt keine annotierten Daten

Zeitausdrücke

...
The ships were scuttled there in the 11th century, to block a navigation channel and thus protect Roskilde, then Copenhagen from seaborne assault
...
...
Viking Age inscriptions have also been discovered on the Manx runestones on the Isle of Man.
...
...
In the Viking colony of Iceland, an extraordinary vernacular literature blossomed in the 12th through 14th centuries
...



Plain-Text **NIF-Anreicherung** **Visualisierung**

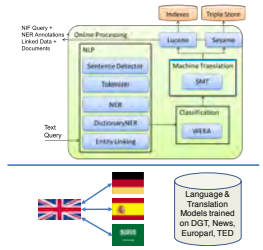
http://api.digitale-kuratierung.de/api/e-nlp/namedEntityRecognition?analysis=temp

http://dev.digitale-kuratierung.de/admin/pages/timelining.php

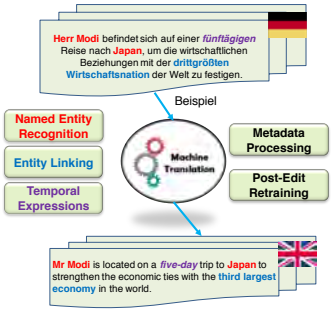
- Sortiert Dokumente auf einer chronologischen Skala.
- Regelbasiertes System, um unsere Zielsprachen bestmöglich bedienen zu können (EN, DE).
- Analyse von Zeitausdrücken in einem Dokument.
- Berechnet Durchschnittswerte und Intervalle.
- Plan: Mechanismus für nutzerbasierte Regeln.
- Verwandte Arbeiten: SUTime, HeidelTime, Tango, Tarsgi.

Maschinelle Übersetzung

- Robuste, adaptierbare MT-Modelle (nutzen Moses, Cdec, Giza++, SRILM etc.)
- Parallele und monolinguale Korpora: Europarl, DGT-TM, TED, UN, Newscrawl u.a.
- Kombination mit anderen DKT-Services (Summariser, NER, Temporal Analyser); ITS 2.0, NIF
- Diverse Linked-Data-Datenquellen unterstützen MT (z.B. Dbpedia, BabelNet, WordNet)



Ankit Srivastava, Georg Rehm, and Felix Sasaki. Improving Machine Translation through Linked Data. The Prague Bulletin of Mathematical Linguistics, 108(1):255-366, June 2017. Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT 2017).



Textzusammenfassen

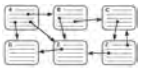
- Kuratierungsservice rankt Sätze – basierend auf div. Features – hinsichtlich ihrer Wichtigkeit.
- Modul ist in der Entwicklung.
- Beispiel: Artikel über den fallenden Aktienkurs von RWE (Daten stammen von Condat).
- Ausblick: Integration der Analyseergebnisse anderer DKT-Services in den Algorithmus.

Im letzten Monat und den letzten 3 Monaten verlor die RWE-Aktie 3,79% bzw. 18,95% und in den letzten 3 Tagen 3,55%.

Die Aktie der RWE AG fiel am Donnerstag um 0,21% auf 19,16 EUR und schwankte am Handelstag zwischen 19,08 und 19,32 EUR. Das Handelsvolumen der Aktie lag bei 1,79 Millionen Aktien und so unter dem 52-Wochen und 150-Tagesvolumen von 3,40 Millionen bzw. 3,96 Millionen Aktien. Im letzten Monat und den letzten 3 Monaten verlor die RWE-Aktie 3,79% bzw. 18,95% und in den letzten 3 Tagen 3,55%. Das PE und PB-Verhältnis der Unternehmensaktie liegt aktuell bei 11,44 bzw. 1,29, während die historischen PE und PB-Werte jeweils bei 11,77 bzw. 2,13 liegen.

Semantic Storytelling

- **Eingabe:** Kohärente, in sich geschlossene Textkollektion
- **Ausgabe:** Semantisch angereicherte Kollektion
- **Idee:** Aufgabenspezifisch multiple Rezeptionspfade generieren, vorschlagen, präsentieren
- **Lösung:** **Identifizierung, Ranking und Empfehlung sinnvoller, *überraschender* Hypertextpfade**
- Es gibt noch zahlreiche Herausforderungen.



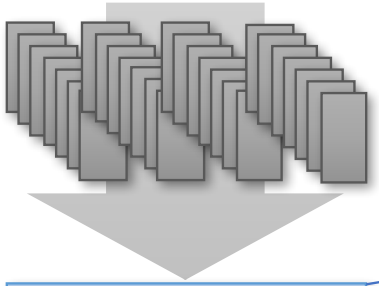
Peter Bourgonje, Julian Moreno Schneider, Georg Rehm und Felix Sasaki. Processing Document Collections to Automatically Extract Linked Data: Semantic Storytelling Technologies for Smart Curation Workflows. In Aldo Gangemi und Claire Gardent, Hrsg., *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, S. 13-16, Edinburgh, UK, Sept. 2016. Association for Comp.Linguistics.

Julian Moreno Schneider, Peter Bourgonje, Jan Nehring, Georg Rehm, Felix Sasaki, and Ankit Srivastava. Towards Semantic Story Telling with Digital Curation Technologies. In Larry Birnbaum, Octavian Popescuk und Carlo Strapparava, Hrsg., *Proceedings of Natural Language Processing meets Journalism - IJCAI-16 Workshop (NLPMU 2016)*, New York, Juli 2016.

Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring, Georg Rehm, Felix Sasaki und Ankit Srivastava. "Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer." In Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenić, Sören Auer und Christoph Lange, Hrsg., *The Semantic Web: ESWC 2016 Satellite Events*, Juni 2016.

Beispiel: Die Mendelsohn-Briefe

Experiment:
Überführung einer Sammlung von Briefen in einen Reisebericht



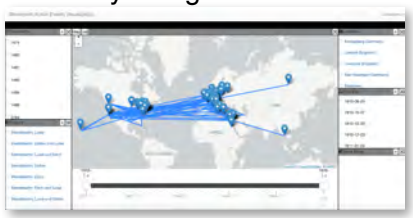
Semantic Storytelling Backend

- Input: Self-contained document collection
- Example: Mendelsohn letters, 2796 documents, written in German, English, French

Semantic Storytelling: Analysis and Annotation Steps

- Language identification (for cross-lingual processing)
- Temporal expression analysis (TimeX)
- Geographic location analysis (GeoX)
- Participants and actors analysis (Person X)
- Coreference analysis
- Event detection (cross-lingual, including German and French, through machine translation)
- Mode of transportation analysis
- Identification of Movement Action Events out of the set of identified events (filtering)

Experimental Storytelling Dashboard



RDF DB

Authoring Environment

- Assists the editor in putting together stories based on the semantic analyses
- Enables the construction of new stories, for example, by (1) focussing on the specific requirements of different text types such as *biography* or *travelogue* or (2) through highlighting and recommending to the human expert specific relationships between entities

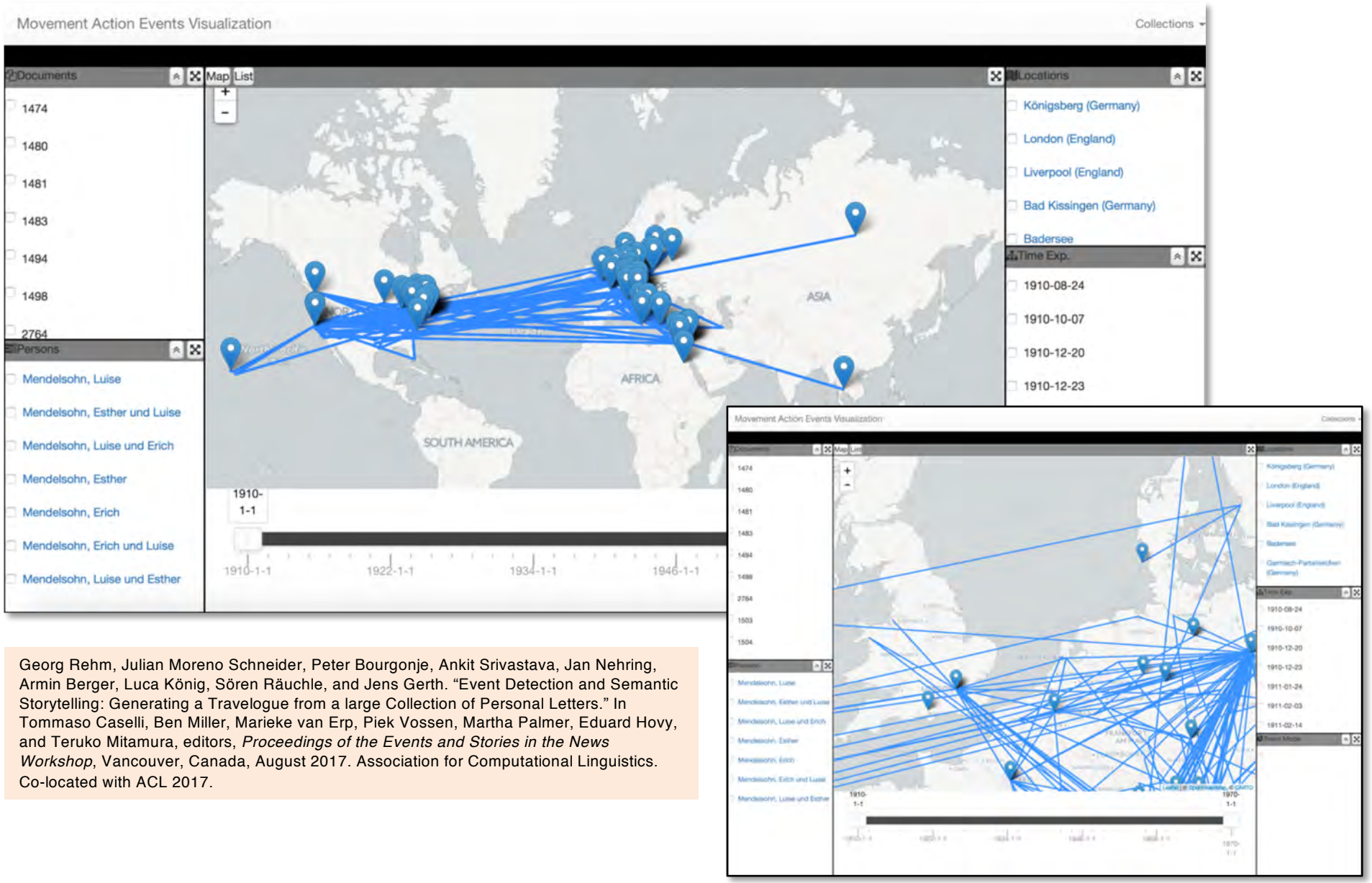
RDF DB

- Automatic transformation of RDF database contents into play-out formats for different channels and media

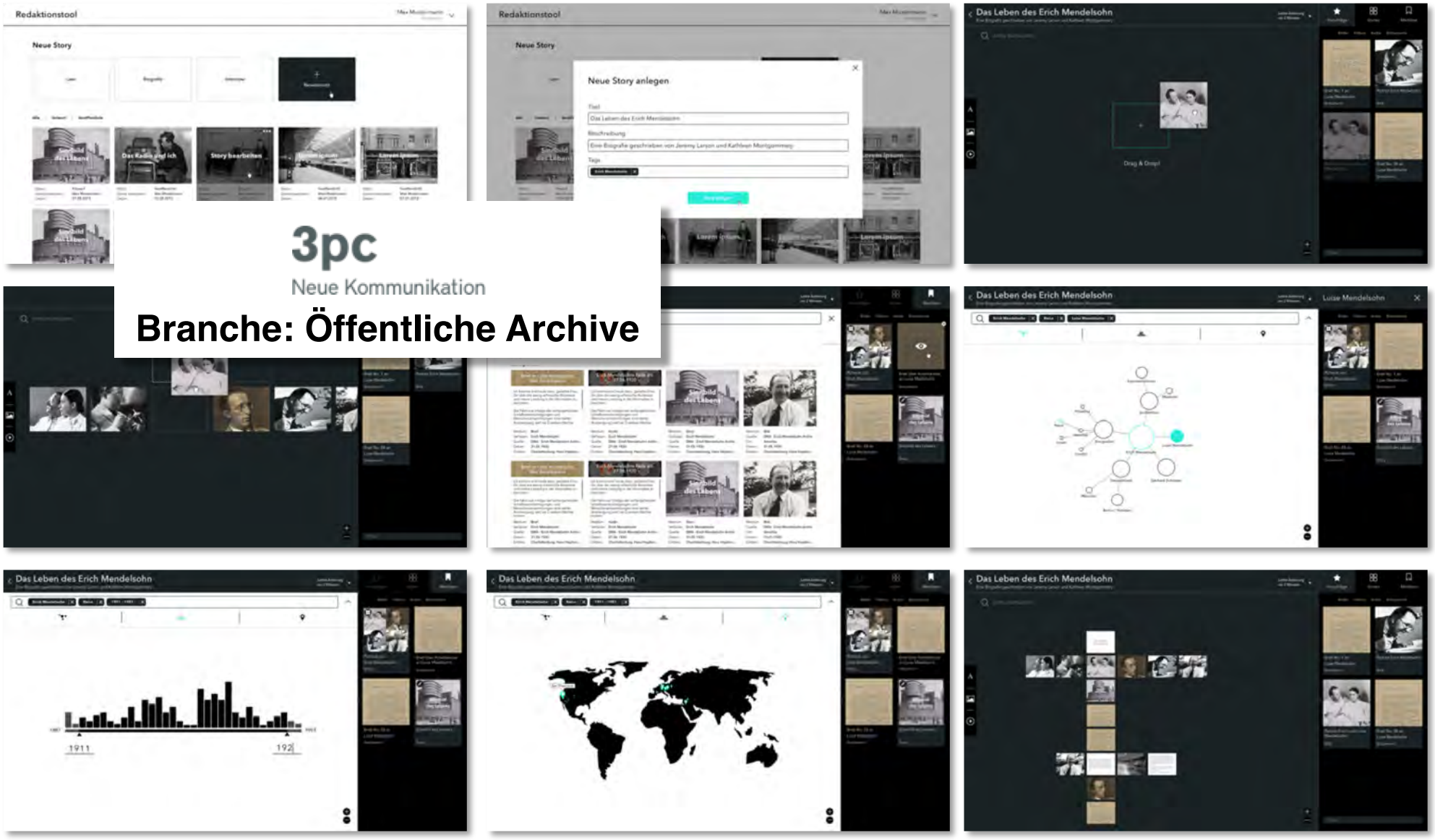
iOS App ePub HTML5 Android App …

Georg Rehm, Julian Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Jan Nehring, Armin Berger, Luca König, Sören Räuchle, and Jens Gerth. "Event Detection and Semantic Storytelling: Generating a Travelogue from a large Collection of Personal Letters." In Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard Hovy, and Teruko Mitamura, editors, *Proceedings of the Events and Stories in the News Workshop*, Vancouver, Canada, August 2017. Association for Computational Linguistics. Co-located with ACL 2017.

Beispiel: Die Mendelsohn-Briefe



Georg Rehm, Julian Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Jan Nehring, Armin Berger, Luca König, Sören Räuchle, and Jens Gerth. “Event Detection and Semantic Storytelling: Generating a Travelogue from a large Collection of Personal Letters.” In Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard Hovy, and Teruko Mitamura, editors, *Proceedings of the Events and Stories in the News Workshop*, Vancouver, Canada, August 2017. Association for Computational Linguistics. Co-located with ACL 2017.



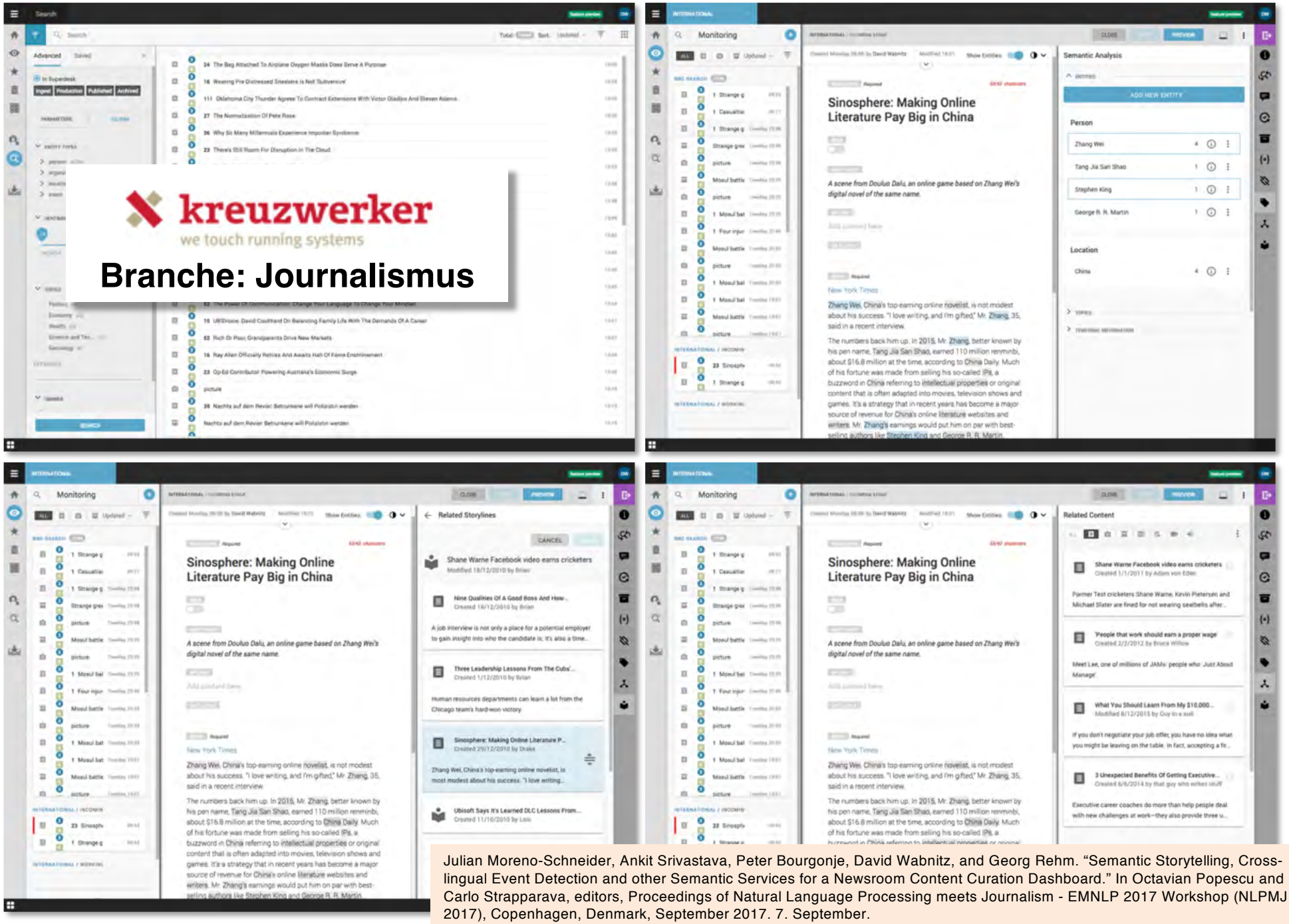
3pc

Neue Kommunikation

Branche: Öffentliche Archive

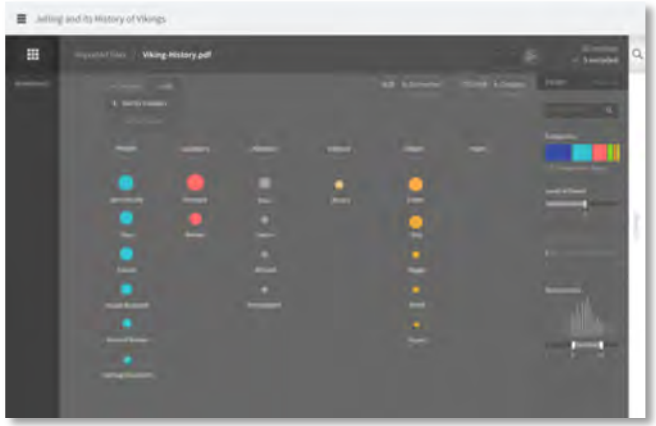
Georg Rehm, Julian Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Jan Nehring, Armin Berger, Luca König, Sören Räuchle, and Jens Gerth. “Event Detection and Semantic Storytelling: Generating a Travelogue from a large Collection of Personal Letters”. In Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard Hovy, and Teruko Mitamura, editors, *Proceedings of the Events and Stories in the News Workshop*, Vancouver, Canada, August 2017. Association for Computational Linguistics. Co-located with ACL 2017.

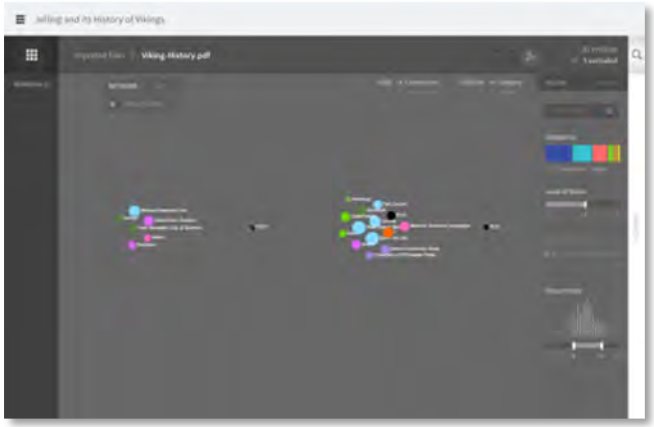
Georg Rehm, Julián Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Rolf Fricke, Jan Thomsen, Jing He, Joachim Quantz, Armin Berger, Luca König, Sören Räuchle, Jens Gerth, and David Wabnitz. “Different Types of Automated and Semi-Automated Semantic Storytelling: Curation Technologies for Different Sectors”. In Georg Rehm and Thierry Declerck, editors, Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 232-247, Cham, Switzerland, January 2018. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.



Julian Moreno-Schneider, Ankit Srivastava, Peter Bourgonje, David Wabnitz, and Georg Rehm. “Semantic Storytelling, Cross-lingual Event Detection and other Semantic Services for a Newsroom Content Curation Dashboard.” In Octavian Popescu and Carlo Strapparava, editors, Proceedings of Natural Language Processing meets Journalism - EMNLP 2017 Workshop (NLPMJ 2017), Copenhagen, Denmark, September 2017. 7. September.





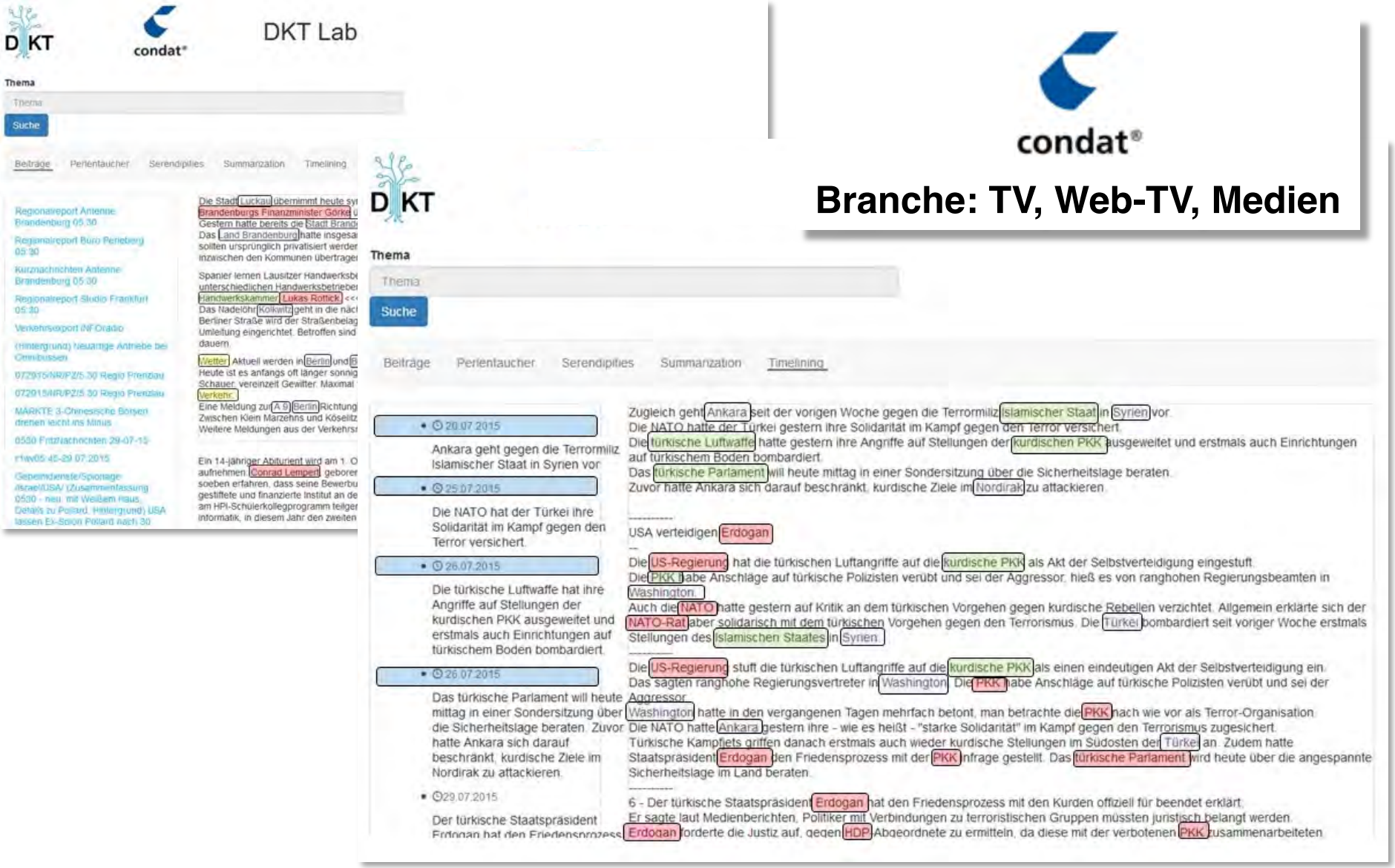



ART+COM

Branche: Museen, Showrooms, Ausstellungen

Georg Rehm, Jing He, Julian Moreno Schneider, Jan Nehring, and Joachim Quantz. *Designing User Interfaces for Curation Technologies.* In Sakae Yamamoto, editor, *Human Interface and the Management of Information: Information, Knowledge and Interaction Design, 19th International Conference, HCI International 2017*, number 10273 in Lecture Notes in Computer Science (LNCS), pages 388-406, Vancouver, Canada, July 2017. Springer.

Georg Rehm, Julián Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Rolf Fricke, Jan Thomsen, Jing He, Joachim Quantz, Armin Berger, Luca König, Sören Räuchle, Jens Gerth, and David Wabnitz. “Different Types of Automated and Semi-Automated Semantic Storytelling: Curation Technologies for Different Sectors”. In Georg Rehm and Thierry Declerck, editors, Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings, number 10713 in Lecture Notes in Artificial Intelligence (LNAI), pages 232-247, Cham, Switzerland, January 2018. Gesellschaft für Sprachtechnologie und Computerlinguistik e.V., Springer. 13/14 September 2017.



- **BMBF-Projekt Digitale Kuratierungstechnologien:**
  - Museen, Showrooms, Ausstellungen
  - TV, Web-TV, Medien
  - Öffentliche Archive
  - Journalismus

- **Außerdem Kuratierungstechnologien konzipiert für:**
  - Bibliothekswissenschaft und Digital Libraries
  - Customer-Relationship-Management
  - Film- und Kinobranche
  - Digital Humanities

Georg Rehm. "KI für die Kundenkommunikation: Der Markt der Zukunft." Rethink! Connected Customer 360°. Hamburg, June 22/23, 2017.

Georg Rehm. "Artificial Intelligence for the Film Industry." FilmTech Meetup Berlin, July 25, 2017.

Clemens Neudecker und Georg Rehm. "Digitale Kuratierungstechnologien für Bibliotheken". *Zeitschrift für Bibliothekskultur 027.7*, Open Access. Nov. 2016.

Georg Rehm. *Der Mensch bleibt im Mittelpunkt – Smarte Technologien für alle Branchen*. Vitako Aktuell. Zeitschrift der Bundes-Arbeitsgemeinschaft der Kommunalen IT-Dienstleister e.V., 2-2016:26-27, 2016.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. "Domain-specific Entity Spotting: Curation Technologies for Digital Humanities and Text Analytics." In Nils Reiter and Gerhard Kremer, editors, CUTE Workshop 2017 – CRETA Unshared Task zu Entitätenreferenzen. Workshop bei DHd2017, Berne, Feb. 2017.

Generierte Folgeprojekte

- **Digitale Kuratierung von Archivmaterialien zur deutschen Wiedervereinigung** (Kooperation mit der FU Berlin, 2018/2019)
- **Kuratierungstechnologien für Forschungsdaten aus den Bereichen Sprache, Text und Digital Humanities** (Kooperation mit der TU Berlin, 2018/2019)
- **QURATOR – Curation Technologies** (BMBF, 2018-2021)

QURATOR
CURATION TECHNOLOGIES

- **Interfaces to Data for Historical Social Network Analysis and Research (SoNAR)** (DFG, Start ca. März/April 2019)

QURATOR

Start am 1. November 2018
36 Monate Laufzeit

Zielsetzung

- Schaffung eines **Ökosystems für Kuratierungstechnologien** – wir haben dieses neue Schlagwort seit 2015 maßgeblich geprägt
- Metropolregion **Berlin-Brandenburg** als **Exzellenzstandort** für deren Entwicklung und industrielle Anwendung etablieren

Branchenlösungen

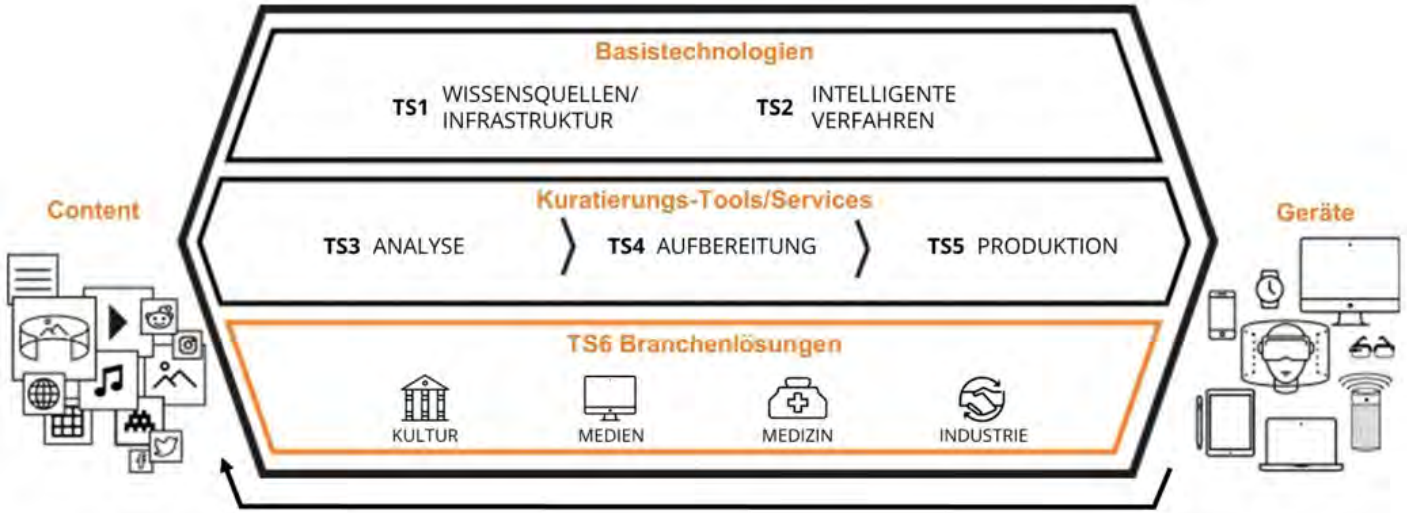
- Medical Content Curator
- Smarte Exponate
- Medien-Kurator
- Intelligente Navigation
- Risiko-Monitoring
- Next Reality Storytelling

Weitere Showcases

- Content Curation Engine
- Corporate Smart Insights
- Speech-to-Text API
  Video Action API
- Semantische Anreicherung
- Basistechnologien

QURATOR im Überblick

- **12 Teilprojekte – ein Verbundprojekt (Koordinator: DFKI)**
- **Sechs Themenschwerpunkte** in **drei Ebenen**
- **Technologieplattform** folgt dem **Baukastenprinzip**
- In Teilprojekten entwickelte Verfahren sind individuell nutzbar – sie bilden die **flexibel kombinierbaren Bestandteile** und **Services** der **QURATOR-Plattform**



1. Herausforderung

Das mehrsprachige Europa:
Sprachtechnologien für alle europäischen Sprachen?

- Sehr großer Bedarf! Von einer langfristigen Investition sind zahlreiche wissenschaftliche, wirtschaftliche & gesellschaftliche Effekte zu erwarten.
- ELG für erstes Deployment. Ab 2021: Human Language Project?

2. Herausforderung

Online-Desinformationskampagnen:
Technische Lösungsansätze gegen „Fake News“?

- Automatische Klassifikation nur für drastische Inhalte/Flagging geeignet.
- Technologien können Nutzern kritischere Medienrezeption ermöglichen.

3. Herausforderung

Digitaler Content:
Technologien für die effiziente Content-Kuratierung?

- Wissensarbeiter können durch branchenspezifische Technologien bei der Kuratierung von Content sinnvoll und effektiv unterstützt werden.
- Vorteil: Die US-Tech-Riesen bedienen weder Branchen noch Nischen.

Herzlichen Dank für die Aufmerksamkeit!

Dr. Georg Rehm
georg.rehm@dfki.de

Vielen herzlichen Dank an:
Peter Bourgonje, Aljoscha Burchardt, Thierry Declerck, Kathrin Eichler, Antske Fokkens, Josef van Genabith, Stefanie Hegele, Brigitte Jörg, Tina Klüwer, Sebastian Krause, Arle Lommel, Sebastian Möller, Julian Moreno Schneider, Jan Nehring, Roland Roller, Nieves Sande, Felix Sasaki, Ankit Srivastava, Hans Uszkoreit, Wolfgang Wahlster, Sarah Weichert.