Chapter 43
Deep Dive Data and Knowledge

Martin Kaltenböck, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg

Abstract This deep dive on data, knowledge graphs (KGs) and language resources (LRs) is the final of the four technology deep dives, as data as well as related models are the basis for technologies and solutions in the area of Language Technology (LT) for European digital language equality (DLE). This chapter focuses on the data and

Martin Kaltenböck · Artem Revenko
Semantic Web Company, Austria,
martin.kaltenboeck@semantic-web.com, artem.revenko@semantic-web.com

Khalid Choukri · Victoria Arranz
Evaluations and Language Resources Distribution Agency, France,
choukri@elda.org, arranz@elda.org

Svetla Boytcheva
Ontotext, Bulgaria, svetla.boytcheva@ontotext.com

Christian Lieske
SAP SE, Germany, christian.lieske@sap.com

Teresa Lynn
Dublin City University, ADAPT Centre, Ireland, teresa.lynn@adaptcentre.ie

German Rigau · Aritz Farwell · Itziar Aldabe · Ainara Estarrona
University of the Basque Country, Spain, german.rigau@ehu.eus, aritz.farwell@ehu.eus, itziar.aldabe@ehu.eus, ainara.estarrona@ehu.eus

Maria Heuschkel
Wikimedia Deutschland, Germany, maria.heuschkel@wikimedia.de

Gareth Jones
Bangor University, United Kingdom, g.jones@bangor.ac.uk

Katrin Marheinecke
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany,
katrin.marheinecke@dfki.de

Stelios Piperidis
R. C. "Athena", Greece, spip@athenarc.gr

Vincent Vandeghinste
Dutch Language Institute, The Netherlands, vincent.vandeghinste@ivdnt.org

Claudia Borg
University of Malta, Malta, claudia.borg@um.edu.mt

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_43

LRs required to achieve full DLE in Europe by 2030. The main components identified – data, KGs, LRs – are explained, and used to analyse the state-of-the-art as well as identify gaps. All of these components need to be tackled in the future, for the widest range of languages possible, from official EU languages to dialects to non-EU languages used in Europe. For all these languages, efficient data collection and sustainable data provision to be facilitated with fair conditions and costs. Specific technologies, methodologies and tools have been identified to enable the implementation of the vision of DLE by 2030. In addition, data-related business models and data-governance models are discussed, as they are considered a prerequisite for a working data economy that stimulates a vibrant LT landscape that can bring about European DLE.[1]

1 Introduction

Digital language equality (DLE) as well as the European data economy rely on the availability, the interoperability and the form of (unstructured, semi-structured, structured) data as a basis for further innovation and improved technological development, especially for trustworthy AI "made in Europe" and powerful language technology (LT) that respects and reflects European values. Data spaces,[2] data sharing and exchange platforms[3] and marketplaces are enablers, key to unleashing the potential of such data. However, data sharing and interoperability are still in their infancy. The diffusion of platforms for data sharing and availability of interoperable datasets is one of the key success factors which may help to drive the European data economy and industrial transformation.

The European Digital Single Market strategy that was adopted on 6 May 2015[4] has been built on three pillars: access, environment, and economy & society. The latter aims at maximising the growth potential of the digital economy, inspired by the 2018 Commission Communication "Towards a common European data space",[5] which provides guidance on B2B data sharing, bringing together data as a key source of innovation and growth from different sectors, countries and disciplines, into a common data space. Overall, the EU has specified its ambition[6] to become the world's most secure and trustable data hub.

This chapter provides insights into: 1. the main components of this deep dive, 2. the current state-of-the-art, 3. the main gaps identified in the field, 4. its contri-

[1] This chapter is an abridged version of Kaltenböck et al. (2022).

[2] Next-generation data acquisition and processing platforms as exemplified, among others, by the BDVA reference model: https://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf.

[3] Data sharing and exchange platforms, through which data is commercialised using open data, monetised data and trusted data sharing mechanisms.

[4] https://ec.europa.eu/commission/presscorner/detail/en/IP_15_4919

[5] https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-232-F1-EN-MAIN-PART-1.PDF

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

bution to DLE and the impact on society, 5. an analysis of the main breakthroughs needed in the area of data, language resources (LRs), and knowledge graphs (KGs), 6. the main technology visions and development goals identified to help achieve deep natural language understanding (NLU), all closed by 7. a summary and conclusions section.

1.1 Scope of this Deep Dive

This deep dive covers a relatively wide range of technologies in the area of LT, including machine translation (MT, Chapter 40), speech technologies (Chapter 41), text analytics and NLU (Chapter 42) as well as content management and knowledge management systems, text generation, and language learning systems, as data and LRs are the backbone for all these technologies as well as many more. In addition, the area of KGs plays an important role in this deep dive as KGs provide powerful mechanisms and principles to interlink and enrich data in a high-quality manner. KGs can build a powerful and relatively easy to maintain network of interlinked data – including and combining structured, semi-structured and unstructured data – that can be seen as a crucial element of the data infrastructure required to develop future LT solutions, which require not only a single underlying dataset but in addition a wide range of meaningful and contextualised data. Furthermore, the integrated data models inside of KGs (taxonomies, vocabularies and ontologies) allow the training of algorithms for LT solutions with higher precision requiring smaller amounts of training data.

The topic of metadata and data in this chapter is always related to LT, language understanding and DLE in Europe. Accordingly, metadata and data in this respect concern (mostly, but not exclusively) LRs, (annotated) corpora, translation memories, dictionaries and lexicographic resources, as well as other LRs and relevant data that is required for powerful multilingual LT. Such data and metadata constitute a strong enabler of AI and machine learning (ML), methodologies that have enabled innovative approaches and advances in the field of LT (Elliot et al. 2021).

In addition to these principal components, a number of related methodologies and tools are currently on the rise, and these form part of the technological vision for 2030 in this deep dive. The subject of data-related business models is tackled throughout the chapter, as functioning, sustainable data-related business models are a prerequisite for a thriving data economy and ecosystem that in turn stimulates and fosters those data-related components listed above, to enable a working LT landscape that can deliver European DLE.

1.2 Main Components

The main components of our analysis related to data, LRs, and KGs include: 1. availability of data and metadata, 2. accessibility of data, 3. quality of data, 4. data interoperability , 5. licensing and data-related regulations, 6. data and ethics, and 7. data literacy. At the same time, the following related concepts, methodologies and tools also need to be considered: 8. data infrastructures, data spaces and data markets; 9. data at scale; 10. KGs; 11. semantic AI (statistical and symbolic AI in combination); and 12. innovative data and metadata management tools.

These main components always include structured data, semi-structured data and/or unstructured data, which can apply to different modalities, e. g., written, spoken, signs, etc. In addition, as for other technology areas, the data for LT may be available as raw data and/or curated data, at varying levels of quality.

With the rise of AI, the importance of large language models (such as, e. g., BERT[7] or GPT-3[8]), and comprehensive and multilingual KGs – all based on a broad range of domains and/or languages – is continuously increasing. For all LRs and data types there is the requirement for domain-specificity, so that domain- and industry-specific applications can be developed where specialised language and terminology are realised, e. g., in industries such as health, pharmaceuticals or finance. Let us now examine each of these aspects in detail:

Availability of data and metadata – As data and metadata form the backbone of any LT, the availability of data and metadata is the overall basis to enable such technologies and services. Availability therefore impinges on data collections, data types available, and how to find and explore such data.

Accessibility of data – The accessibility of data is crucial, it is also reflected in the FAIR principles (Wilkinson et al. 2016), initially advocated for research data management and stewardship in order to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. Since 2007, accessibility has also been one of the initial eight key principles of open (government) data.[9]

Quality of data – When data is available and accessible, users often consider additional attributes and components, one being quality of data. As the value of data is based on its fit for certain use-cases and business cases, data quality is a crucial issue reflecting and impacting the respective data value. Dimensions to measure data quality often include – but are not limited to – completeness, validity, timeliness, consistency, and integrity (Sebastian-Coleman 2012). Reliability is also an important factor of data quality, although it is hard to measure. When all things are considered, the quality of an LT application is often based largely on the quality of the underlying data used to train the system.

Data interoperability – Data interoperability is defined as[10] “addresses[ing] the ability of systems and services that create, exchange and consume data to have clear,

[7] https://en.wikipedia.org/wiki/BERT_(language_model)

[8] https://en.wikipedia.org/wiki/GPT-3

[9] https://opengovdata.org

[10] https://datainteroperability.org

shared expectations for the contents, context and meaning of that data." Interoperability ensures the seamless interplay of different LT systems regarding both APIs and data exchange. Not unexpectedly, it is often connected with and facilitated by the specification and adoption of related standards in the field.

Licensing and data-related regulations – Relevant data often comes from different owners and publishers, such as companies, public administrations or citizens, with different licences. Accordingly, proper licence clearing is a crucial task for all data-related activities in LT. The licences on data that are usually specified by data owners/publishers need to be taken into account as an important component, as well as the applicable laws and regulations around data, such as those concerning data privacy, security, processing and protection of personal identifiable information (PII), as laid out, for instance, in the General Data Protection Regulation (GDPR). National and regional as well as international regulations and policies around data use and re-use should also be taken into account.

Data and ethics – The rise of AI and ML has led to an increase in both data collection and processing, so the issue of data and ethics has become more and more important. It is closely connected to data-related regulations. Language, by its very nature, can be ambiguous and the associated interpretations can easily represent and expose bias. Accordingly, ethics plays a crucial role regarding the use of data in LTs and impacts equality in general, including language equality.

Data literacy – Gartner Research[11] defines data literacy as "the ability to read, write and communicate data in context, including an understanding of data sources and constructs, analytical methods and techniques applied, and the ability to describe the use-case, application and resulting value."

Data infrastructures, data spaces, data markets – The ideas behind data spaces and data markets follow the intentions underpinning data catalogues established in the course of the open data movement since the early 2000s to allow the sharing, exchange and trading of data. Data spaces and markets enable the availability of and allow accessibility to high-quality data, which follow standards (thus providing data interoperability) accompanied by clear licensing conditions. The Gaia-X[12] initiative defines a "data space" as "refer[ring] to a type of data relationship between trusted partners, each of whom apply the same high standards and rules to the storage and sharing of their data. However, of key importance to the concept of a data space is that data are not stored centrally but at source and are therefore only shared (via semantic interoperability) when necessary. A data space is the sum of all its participants – which may be data providers, users and intermediaries. Data spaces can be nested and overlapping, so that a data provider, for example, can participate in several data spaces all at once. Data sovereignty and trust are essential for the working of data spaces and the relationships between participants."

Data at scale – Practical LT solutions require high-quality data at scale and for a broad range of domains and available in various languages, with clear licences and fair conditions attached. Data infrastructures, data spaces and data markets provide

[11] https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy

[12] https://gaia-x.eu/what-is-gaia-x/

powerful means to discover, evaluate and access relevant data as well as related data-driven services, that are required for LT solutions.

Knowledge Graphs – A Knowledge Graph is a knowledge base that uses a graph-structured data model or topology to integrate data. KGs are used to store interlinked descriptions of entities – objects, events, situations or concepts – while also encoding the semantics underlying the terminology used.[13] Since the development of the Semantic Web, KGs have often been associated with Linked Open Data (LOD) projects, focusing on the connections between concepts and entities (Soylu et al. 2020; Auer et al. 2018). They are prominently associated with and used by search engines such as Google or Bing; knowledge-engines and personal assistants such as Wolfram Alpha, Apple's Siri, and Amazon Alexa; and social networks such as LinkedIn and Facebook. LT solutions require not only targeted datasets but also high-quality, interlinked, meaningful and contextualised data that can easily be used, quickly expanded and efficiently maintained with reasonable effort. KGs provide these characteristics and contribute to the data and knowledge backbone for LT.

Semantic AI – Modern approaches tend to combine statistical AI (ML) and symbolic AI (models like ontologies, knowledge bases for common sense knowledge, and cultural resources, among others). In October 2020, Agarwal defined semantic AI[14] as "provid[ing] a framework to perform end to end complex tasks automatically. It uses many different machine learning and logic-based approaches, and also utilizes the background knowledge often stored in knowledge graphs."

Innovative data and metadata management tools – Innovative data and metadata management tools enable the availability and accessibility of high-quality data and data interoperability (using relevant standards), that provide powerful data governance mechanisms (following relevant regulations), that enable mechanisms for the assessment of ethics in data, and that allow improvements in data literacy. In addition such tools should support (perhaps in combination with) secure data sharing mechanisms (data spaces), provide strong capability for interlinking data, support meaning and context (KGs) and provide semantic AI capability.

2 State-of-the-Art and Main Gaps

2.1 State-of-the-Art

From the start of the open data movement in 2007 with its eight principles of open government data, the requirements of industry data as well as organisation-based data-sharing and collaboration have found their feet and culminated in the next era of data sharing: data catalogues and data portals, as well as, more recently, data spaces and data markets. In the area of LT, data availability, accessibility, aggregation, shar-

[13] https://ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph

[14] https://medium.com/@dr.puneet.a/what-is-semantic-ai-is-it-a-step-towards-strong-ai-5f0355be3597

ing and reuse have received attention since the early 1990s, with associations and organisations providing LR catalogues, like the European Language Resources Association[15] or the Linguistic Data Consortium.[16] Since the early 2010s, several research and innovation projects have contributed to the field including FLaReNet and META-NET with META-SHARE[17] (Piperidis 2012). They provided recommendations, specifications and implementations of platforms promoting and facilitating data discovery, sharing and reuse. At the same time, CLARIN[18] (Hinrichs and Krauwer 2014) has been established as a research infrastructure providing access to digital language data for scholars in the social sciences and humanities, and beyond. CLARIN is associated with the EUDAT Collaborative Data Infrastructure (EUDAT CDI),[19] and contributes to the European Open Science Cloud (EOSC)[20] with the EOSC-related project Social Sciences and Humanities Open Cloud (SSHOC)[21] and its data market for social sciences and humanities.[22]

Another example of research, development and infrastructure activities supported by the implementation of the Public Sector Information Directive[23] is the ELRC-SHARE repository[24] (Piperidis et al. 2018) that is used for documenting, storing, browsing and accessing LRs that are collected through the European Language Resource Coordination[25] initiative (Lösch et al. 2018) and considered useful for feeding the CEF Automated Translation (CEF.AT) platform.

In 2022, the European Language Grid (ELG)[26] (Rehm et al. 2020a; Rehm 2023) released the ELG platform providing access to LT resources and services from all over Europe, enabling users to try out the services or use the ELG APIs. ELG built bridges to a wide range of language data platforms including the European AI on Demand Platform (Labropoulou et al. 2023).

Turning to the LT industry, there are products like the TAUS Marketplace,[27] as well as APIs for lexicographical information or Natural Language Processing (NLP) APIs giving access to services from part-of-speech tagging and dependency parsing to MT, summarisation and question answering. Finally, there are active industry as-

[15] http://www.elra.info

[16] https://www.ldc.upenn.edu

[17] http://www.meta-share.org

[18] https://www.clarin.eu

[19] https://www.eudat.eu

[20] https://eosc-portal.eu

[21] https://sshopencloud.eu

[22] https://marketplace.sshopencloud.eu

[23] https://digital-strategy.ec.europa.eu/en/policies/public-sector-information-directive

[24] https://elrc-share.eu

[25] https://lr-coordination.eu

[26] https://www.european-language-grid.eu

[27] https://datamarketplace.taus.net

sociations and networks like LT-Innovate[28] or BDVA/DAIRO[29] that support the idea of data collection and provision and sharing to support better LT in the future.

Most if not all of the above platforms and initiatives have now endorsed the FAIR principles, adopting them as a de facto standard. In this context, data interoperability has been an important factor, related (mostly but not exclusively) to efficient data use and processing, as well as data exchange and sharing. There are dozens of standards regarding data in place worldwide, set up by several standardisation bodies in a range of industry domains. This diversity of data-related standards reinforces the problem as there is relatively little mapping between such standards and approaches. In the context of ELG and with regard to the wider area of AI/LT platform interoperability, initial attempts have been made at cross-platform search and discovery of resources and services, on the one hand, and composition of cross-platform service workflows, on the other (Rehm et al. 2020b).

Since the open data and data sharing movement began, every digital asset has needed to be accompanied by a clear and dedicated licence. While this issue has become more and more important, there are quite a lot of possible licences to choose from, inevitably reinforcing legal interoperability problems. While there are multiple commercial licensing options not centrally registered, a good source for open licences is the Open Definition of the Open Knowledge Foundation.[30]

Several data regulations and directives have been developed by the European Union over the last decade. They are an important foundation of the data economy, as well as the realisation of a working, sustainable data infrastructure across Europe. Some of the most important ones include, among others: GDPR,[31] European Strategy for Data,[32] European Data Governance (Data Governance Act),[33] EU Open Data Strategy and PSI Directive,[34] European Approach to Artificial Intelligence, including the EC AI Strategy,[35] Digital Single Market Strategy for Europe,[36] and Digital Action Education Plan.[37] As far as LT for DLE in Europe is concerned, all of these regulations have a clear impact. In terms of this deep dive, the Data Governance Act has a strong implication for data, LRs and KGs, as it lays the groundwork for the development of common data spaces in strategic sectors.

Setting technical issues to one side, data and ethics is a topic in which regulators and standards (such as those mentioned above) play a crucial role. After many years' discussion about data and ethics but also about AI and ethics, a standard has been published: *IEEE P7000 Engineering Methodologies for Ethical Life-cycle Concerns*

[28] https://www.lt-innovate.org

[29] https://www.bdva.eu

[30] https://opendefinition.org/guide/data/

[31] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[32] https://ec.europa.eu/info/sites/default/files/communication-european-strategy-data-19feb2020_en.pdf

[33] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767

[34] https://digital-strategy.ec.europa.eu/en/policies/open-data

[35] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

[36] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52015DC0192

[37] https://ec.europa.eu/education/education-in-the-eu/digital-education-action-plan_en

Working Group. It establishes a process model by which engineers and technologists can address ethical considerations throughout the various stages of system initiation, analysis and design. Expected process requirements include management and engineering views of new IT product development, computer ethics and IT system design, value-sensitive design, and stakeholder involvement in ethical IT system design.[38]

Data literacy is an underlying component of digital dexterity: an employee's ability and desire to use existing and emerging technology to drive better business outcomes. The European Union supports data literacy and beyond in the Digital Action Education Plan,[39] and globally programmes like the World Bank's Data Use and Literacy Programme[40] support the awareness, education and implementation of data literacy. Nevertheless, compared to data and data-related technologies available, the issue of data literacy lags far behind and needs more action and effort to be applied.

The idea of a KG follows the basic principles of the semantic web and linked data. For LTs, the KG principles have great potential for modelling common-sense knowledge and domain-specific knowledge, as well as provisioning rich context and meaning in monolingual, bilingual, multilingual and cross-lingual applications. KGs are often assembled from numerous sources, and as a result, can be highly diverse in terms of structure and granularity.

KGs aim to serve as an ever-evolving shared substrate of knowledge within an organisation or community (Noy and McGuinness 2001). We distinguish two types of KGs: open KGs and enterprise KGs. Open KGs are published online, making their content accessible for the public good. Enterprise KGs are internal to a company and applied to commercial use-cases. Applications based on KGs include search, recommender systems, personal agents, advertising, business analytics, risk assessment, and automation. Useful further reading includes Blumauer and Nagy (2020), Abu-Salih (2021), Colon-Hernandez et al. (2021), Ji et al. (2022), and Li et al. (2021).

The technological leaps in LT and AI in the past few years and the widely recognised importance of data and knowledge resources for their accomplishment have called for new concepts and instruments in the area of data technologies and naturally so also in AI and LTs. In Europe, data spaces are a (relatively) new concept and solution to stimulate the data economy by providing secure and trustworthy mechanisms and platforms for data sharing and data trading. The European Commission lists a number of data spaces in its Data Strategy as of February 2020[41] that is strongly interconnected with the EU Data Governance Act.[42] EU Member States have supported research on data spaces in recent years, as for example Gaia-X[43] and the International Data Spaces initiative (Germany) that channeled into the establishment of the International Data Spaces Association (IDSA) and the publication of several standards and recommendations in the field (IDS Information Model or the

[38] https://sagroups.ieee.org/7000/

[39] https://ec.europa.eu/education/education-in-the-eu/digital-education-action-plan_en

[40] https://www.worldbank.org/en/programs/data-use-and-literacy-program

[41] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy

[42] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[43] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

Reference Architecture Model),[44] or the Data Market Austria (DMA)[45] prototype for a public marketplace for data trading. In January 2023, the European Commission launched the Common European Language Data Space which aims to focus on language data and models discoverability, sharing and trading covering all EU languages and aiming to support a wide range of LT applications in different modalities, domains and contexts.

2.2 Main Gaps

The following observations have been formulated, collected and further analysed together with researchers and practitioners in the field and reflect our joint understanding of the current gaps in the components of this deep dive.

There is untapped potential when it comes to data available in archives as well as old data files. There is a real need for open AI models in LT that are provided to interested parties with open licences. Not only ready-to-use models are required, but also the raw data needs to be made available in order for developers to create their own models. Annotated corpora are often available mainly in English, and it is often the case that they are not available in other languages, let alone all those required for different technologies and applications. The ELG dashboard[46] offers a visual overview of the current standing of Europe's languages (and beyond) with respect to available language data, tools and services. Through such availability counts the dashboard approximates the technological readiness of each language (see Chapter 3). There is an urgent need for monolingual, bilingual and multilingual domain-specific corpora. Such data can only rarely be found via available resources, mostly because it simply is not there, but also because of incorrect or missing documentation of data and metadata. Manually annotated data is lacking; although the quality of automated and semi-automated annotations is increasing, manual annotation by human experts in a certain field is still the best means of acquiring high-quality data.

Overall, there are missing open LRs. Domain-specific LRs are required to be available for scientific purposes with open licences. If the FAIR principles were systematically applied, this would be a huge benefit where data and metadata is concerned, but they are not really being rolled out properly. Although Europe has benefited from a strong open data movement for about 15 years now, there is still a gap in the provision of clearly specified licences for data. At the moment, benchmark approaches are not harmonised or standardised, and benchmarks on domain-specific vocabularies and annotated data and corpora are often missing. Metadata provides only very limited data provenance. Overall data quality is weak and so it often happens that use-cases cannot be realised as specified as labeled data is not available for the use-case at hand. Non-existing policies around data and metadata management

[44] https://internationaldataspaces.org/use/reference-architecture/

[45] https://datamarket.at

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

that should be part of a data governance model often result in low data and metadata quality. There are increasingly many data silos in place that are neither connected nor interoperable, and there are more and more data infrastructures available that are simply not interoperable either, as the harmonisation of relevant standards in the field is missing. This is a clear problem and gap in the combination of research data (e. g., via EOSC)[47] and industry data (e. g., via industry data markets) as well as data from public administration or government data catalogues and portals (e. g., the European Open Data Portal).[48] More and clearer directives and regulations in the field should be developed to overcome these gaps in relation to data, LRs, and KGs. The effect of regulations on data-related topics should be evaluated continuously and regulations and directives adapted for identified gaps and changing environments. For example, GDPR has a strong effect on data collection.

Guidelines and policies are not available for each language in order to achieve DLE in Europe. Data for non-EU languages and beyond are not sufficiently in place, and so services for such languages cannot be developed with sufficient quality for them to be useful. National crowd-sourcing platforms that facilitate data collection for low-resource languages are not available hampering DLE in Europe.

There is a strong need for education that can deliver improved understanding of better data management processes in science, academia, as well as in business and industry. This should lead to better understanding of the value of data, and so improve data management principles and techniques. There is a need to inform educational bodies of the importance of sharing data; for example, if more learner corpora were made asvailable, this would lead to improved computer-assisted language learning and adaptive educational technologies. More senior staff and experts in AI need to work on data-related topics and deliver AI and deep learning mechanisms.

As an overall gap, there is a strong difference with regard to the level of digitisation in Europe. Data catalogues and portals often provide metadata only with links to the listed data that is provided by the data publishers and data owners themselves, with only a small amount also providing the data itself. The resulting issues and gaps relate to 1. *the availability of and access to the data itself*, as information in catalogues as to whether such data continues to be provided by publishers and owners is insufficient; 2. *lack of interoperability in metadata but mainly in the data itself*. The metadata often provides data interoperability (e. g., by using the same catalogue software CKAN),[49] and at least in Europe (but also beyond) we are making use of the de facto metadata standard for open data and data portals DCAT-AP (Data Catalogue Vocabulary (DCAT) expanded for Application Profiles);[50] and 3. *a fragmentation of data catalogues and data portals*.

[47] https://eosc-portal.eu

[48] https://data.europa.eu

[49] https://ckan.org

[50] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

Regarding data spaces and data markets, the TRUSTS project (Trusted Secure Data Sharing Space)[51] has carried out a study[52] on the definition and analysis of the EU and worldwide data market trends and industrial needs for growth, that includes a section on data market challenges, which includes a good summary of the gaps and challenges in this area (Figure 1). All these gaps and issues can only be addressed by working business models in the area of data sharing and trading in a working and successful data economy. IDSA published a relevant report in May 2021.[53]

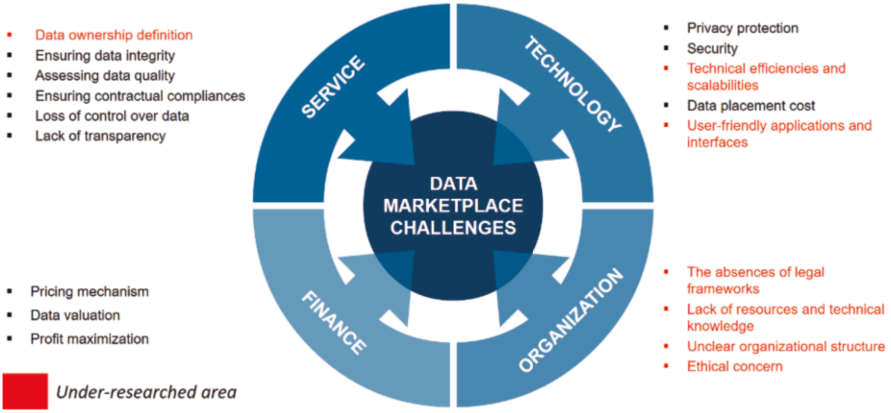
Data ownership definition
Ensuring data integrity
Assessing data quality
Ensuring contractual compliances
Loss of control over data
Lack of transparency
SERVICE
TECHNOLOGY
Privacy protection
Security
Technical efficiencies and scalabilities
Data placement cost
User-friendly applications and interfaces
DATA MARKETPLACE CHALLENGES
Pricing mechanism
Data valuation
Profit maximization
FINANCE
ORGANIZATION
The absences of legal frameworks
Lack of resources and technical knowledge
Unclear organizational structure
Ethical concern
🟥 Under-researched area

Fig. 1 Challenges of data marketplaces

For a KG to become useful for a downstream application there is a need for it to contain a certain amount of application- and domain-specific knowledge. Often, openly available resources are not suitable for a particular task, so to reduce the entry barriers there is a need to be able to generate a suitable ontology or schema for said task and then to populate the schema with instances.

Currently, a KG is mainly developed based on textual and numerical data as an input format, with other formats like video and audio only very rarely taken into account. Working LT in the required languages using mechanisms like speech-to-text could support the creation of KGs.

Finally, there is a gap in the availability of comprehensive KGs. While there are some common-knowledge KGs even freely available (DBpedia, or Yago being just two examples), there is a clear lack of bigger KGs in specific domains and industries, that can act as a kind of foundation model, but also as training input for AI algorithms. Even if such specific graphs were available, there is a clear gap in the availability of multilingual domain-specific KGs that can be used for LT applications.

[51] https://www.trusts-data.eu

[52] https://www.trusts-data.eu/wp-content/uploads/2021/07/D2.1-Definition-and-analysis-of-the-EU-and-worldwide-data-market-trends-....pdf

[53] https://internationaldataspaces.org/the-ecosystem-effect-of-business-models-driven-by-data-sovereignty/

Regarding the gaps in semantic AI, we see that the fields of statistical and symbolic AI are still not fully combined; the two fields often exist in isolation beside one another and so cannot provide their full potential to the solution of a problem. This is largely the case overall in the machine learning and semantic web communities, but it is also the case in areas like LT, or domains like health or energy.

Finally, the following gaps regarding innovative data and metadata management tools have been identified: 1. *The need for user-friendly, flexible, open-source corpus annotation tools* that can easily be used by linguists as well as domain experts in-house and with fair costs and conditions. 2. *The need for user-friendly visualisation tools* in order to be able to understand the content of datasets at hand quickly and properly without the need for significant efforts in data integration and data wrangling. 3. *Better detection techniques for harmful content* are required to avoid bias, and identify and filter toxic content, or fake news and fake data, etc. In a time where AI and ML are being used more and more, even small portions of toxic data and content can influence an algorithm during training and so needs to be identified and filtered out. 4. *Better techniques for corpus filtering* are required regarding domain filtering, noise cleaning (see above, also) as well as the filtering and removal of bias. 5. *A clear lack of preservation technologies and tools* have been identified that are required to ensure that lesser-spoken languages can be archived for the long term (e. g., that are available on tape only) and made available as data that is easy to use, including the provision of proper data documentation in the form of rich metadata. 6. *Intelligent data analytics of small content nuggets* is needed, as, at the moment, often only huge corpora are being analysed by the available technologies and tools, but there is an increasing trend towards smarter data analytics that can be applied on ever smaller datasets, including for instance to just one paragraph or section, rather than the whole text. 7. *Add-on business models* are needed as gaps have been identified in the area of business models around data creation and provision, and so the development of tools and technologies is often limited to small experiments in funded research projects. Having clearly defined and successfully working business models in place would improve the industrial development in the field and stimulate the availability of the required innovative data and metadata management tools.

3 The Future of the Area

3.1 Contribution to Digital Language Equality

The major issue is the lack of available relevant and required data and LRs, as well as KGs in all European languages, official or not. At the moment sufficient data is available mainly in English, and to a lesser degree in German, French and Spanish. However, even in these languages data gaps exist that hinder LT development.

Looking further into this area it is easy to identify an even greater gap in the availability of data regarding dialects of European languages as well as regional languages. Dialects and regional languages exist, they are actively used and form part of a country's or a region's identity and culture. Language diversity is so strong that sometimes in a small region several different languages or dialects are used.

In addition, there is very little data available for sign languages which is a clear issue for the inclusion of those with disabilities, as well as there being little to no respective data available for non-EU languages that are widely spoken in the EU, like Turkish or Arabic, for example.

DLE is a fundamental aspect of a functioning European society, in which diversity and inclusion are valued in every single EU Member State and across Europe with its colourful regional cultures and identities. The lack of DLE in Europe carries the risk of dividing society as it fosters misunderstanding, and may even support the promotion of toxic content, fake news, or lead to wrong interpretations of regional policies and regulations or the misinterpretation of research results in times of crisis.

We have identified the following three approaches: 1. *Digital Language Equality Strategy*: more funding and support by regional and national governments and the European Union to support the development of DLE in Europe for years to come for EU languages as well as regional languages and dialects (and for non-EU languages, too), aided by a data and LR matrix that shows which data and LRs should be available when and for which languages (see Chapter 45); 2. *Crowdsourcing and citizen science*: the creation of the required data needs the support of native speakers as well as linguists with the respective language experience and skills, and the support by data experts providing guidance with regard to the creation of useful and high-quality data; and 3. *Data-related business models*: these are required in the field to foster data creation and acquisition for minor languages and dialects by industry and the private sector.

In addition, and to allow DLE for certain domains like health, for instance, there is a strong need for the continuous development and maintenance of monolingual, bilingual and multilingual domain-specific vocabularies and KGs, to enable the multilingual and cross-lingual development of innovative domain-specific applications that provide value to the economy and society as a whole.

3.2 Breakthroughs Needed

Based on the identified components, the state-of-the-art analysis and the gap analysis, the areas of data infrastructures, data spaces and data markets are major issues where future technology visions and breakthroughs are needed in the field, as this area provides the overall umbrella for the availability and accessibility of the required data for powerful LTs that can help bring about DLE in Europe.

The main breakthroughs needed in terms of *data infrastructures, data spaces and data markets* include: 1. designing working architectures and ensuring effective workflows for compliant data provision and consumption; 2. developing specifications and building blocks to enable data and metadata interoperability; 3. developing and deploying technologies that embed data sovereignty and build trust among data providers and consumers; 4. developing specifications and building blocks that enable data value creation including data publishing and discovery mechanisms as well as accounting and billing; and 5. specifying and developing data governance models with clear roles, rules and policies for all stakeholders.

A recent study by the European Commission (Cattaneo et al. 2020) examines trends in data markets. The study measures *the value of a data market*, i. e., "the marketplace where digital data is exchanged as products or services as a result of the elaboration of raw data", and the *value of the data economy*, i. e., "[by] measur[ing] the overall impacts of the Data Market on the economy as a whole". The study compares the value of the data market and data economy from 2018 to 2019. It also projects the facts and figures for the year 2025 based on three scenarios.

Growth in data markets and the data economy brings with it several implications. According to the European Commission,[54] the total number of data professionals (i. e., those who deal with data endeavours as their primary task) will continue to rise consistently. Many opportunities will open in data-related jobs, and more knowledge workers are needed. Despite these positive trends, there is still a potential lack of supply of data professionals in high-growth scenarios. Companies taking a role as data providers and data buyers will also grow in number and market share.

KGs and semantic AI combined and provided as part of a data infrastructure can bring clear value, and should be part of any data infrastructure in the future. Gartner Research states that from 2021 onwards, graphs will form the foundation of modern data analytics with the capabilities to enhance and improve user collaboration, ML models and explainable AI. Although graph technologies are not new to data analytics, there has been a shift in thinking about them as organisations identify an increasing number of use-cases where they could play an important role. In fact, as many as 50% of Gartner client inquiries around the topic of AI involve a discussion around the use of graph technology.[55] In 2020, it was estimated that by 2023, graph technologies would facilitate rapid contextualisation for decision making in 30% of organisations worldwide.[56]

The main breakthroughs needed in the area of *KGs and semantic AI* include: 1. developing KG principles and technology from the current status of a "rising star" to a natural part of any data infrastructure and any data-related organisational infrastructure; 2. fostering the development of multilingual KGs under fair conditions and costs for use and re-use; 3. fostering the development of domain-specific KGs under fair conditions and costs for use and re-use; 4. KGs need a higher level of automation in their creation and maintenance, and more consideration needs to be given to the format of data beside textual data, such as audio and video; 5. a high level of deep and continuous learning will enable KGs to maintain themselves over time regarding new domain-specific and language-specific terminology. This means that new terms will be identified, analysed and inserted into the graph in the correct position, as well as being applied to the applications used by the KG; 6. bringing together the two main AI communities of statistical AI and symbolic AI to work together on

[54] https://op.europa.eu/s/vbSA

[55] https://www.gartner.com/smarterwithgartner/gartner-top-10-data-and-analytics-trends-for-2021

[56] https://info.tigergraph.com/gartner-graph-steps-onto-the-main-stage-of-data-and-analytics

future semantic AI approaches; and 7. developing the areas of responsible AI and explainable AI by making use of semantic AI in multilingual environments to provide AI-based applications that deliver the correct results with benefits for research, industry and society.

The global enterprise metadata management market is forecast to grow at a rate of 20.3% from USD 7.45 Billion in 2019 to USD 27.24 Billion by 2027. Enterprise metadata management (EMM) provides the control and clarity needed to manage the change that often accompanies a complex enterprise data ecosystem. EMM and the various pieces of management software created for it provide administration for data integration, and allow users to inspect the metadata's links and roles.[57]

The main breakthroughs needed in the area of *innovative data and metadata management tools* include: 1. the development of tools that can be easily integrated with data infrastructures, data spaces and data markets; 2. the development of technologies and tools that can identify and remove bias, toxic content and fake data from data and content; 3. the provision of tools in the field of semantic AI, thus combining statistical and symbolic AI, that provide out-of-the-box responsible and explainable AI capability; 4. the development of a landscape where models and algorithms based on semantic AI can be created, ultimately with smaller amounts of data; 5. tools for data and metadata management that work not only in major languages like English but which can be easily adapted with low cost to smaller languages or dialects; 6. tools that allow deeper modelling of cultural aspects, gender aspects, etc. to avoid bias in data; 7. tools that are able to combine input from various types of data like text, images, audio and video but also gestures; and 8. tools along the whole data life cycle for all languages and all relevant use-cases are required to ensure powerful LT which can help enable DLE.

3.3 Technology Visions and Development Goals

We identified several technology visions and development goals for the area of data, LRs and KGs regarding DLE as a result of a comprehensive list of use-cases in the field, highlighting the related requirements. The majority of use-cases for LT involve human-to-machine and human-to-human communication and interaction via digital tools. To a large extent, these can be categorised using the concepts of *conversational AI* and *platforms and insight engines* that are covered by the other deep dive chapters in this book. In summary, the following excerpts represent identified data and technology development goals:

- LRs (speech, text) for official EU languages as well as for other European and non-European languages, for languages of minorities and dialects;
- pre-trained and fine-tuned language models for general and vertical domains for at least all EU-24 languages;
- speech models addressing at least the EU-24 languages;

[57] https://www.reportsanddata.com/report-detail/enterprise-metadata-management-market

- NLP pipelines of tokenisers, taggers, parsers etc., which require labelled linguistic datasets (e. g., treebanks) and evaluation sets;
- interfaces and content should be available in *all* languages via the web, i. e., the information available on a specific object, person or event provides the same amount of information in all languages;
- knowledge and content available in the form of audio files should be available in all languages so that it can be easily consumed;
- appropriate data required to train and develop monolingual, bilingual and multilingual models that cover the type of knowledge (domain-specific) and the type of language required for MT, (multi)document summarisation and speech-to-text technologies;
- efficient APIs required to integrate organisation-specific data and systems with social media platforms;
- pseudonymised or anonymised data for all EU languages, as well as domain-specific annotated corpora;
- data and models which address gender bias or minority bias etc.;
- data and technologies for identifying and ideally also removing toxic content, hate speech, fake news;
- comprehensive multilingual ontologies in vertical domains;
- KGs for common concepts, event descriptions for daily activities, and patterns for frequent questions;
- text-to-speech resources for common vocabularies and terminologies, as well as computer vision technologies for sign languages;
- data and technologies for modelling culture specific phenomena;
- better designed crowdsourcing platforms to enable more citizen science efforts towards building speech and language systems.

Some of these points are already available and in use in different data infrastructures. Beyond investing in the design and development of the missing parts, it is the integrated combination of all of them that could, from a technology perspective, be the main breakthrough and technology vision for the future management of metadata and data, as well as of LRs, that can act as the backbone for powerful LTs to realise DLE in Europe. Existing LT data infrastructure providers, such as ELG, ELRC-SHARE, CLARIN, META-SHARE, and ELRA as well as industrial and national initiatives can provide the seeds for a kind of federated data infrastructure, i. e., a data space that enables seamless and trusted interactions between data providers and data consumers, and enables cross-fertilisation by means of interoperability, aided among others by semantic KG technologies. Interoperability challenges can be broadly classified in four different layers:

- *technical interoperability*, enabling technical components (i. e., data space connectors) to communicate with each other;
- *semantic interoperability*, ensuring that attributes and policies have the same meaning;
- *organisational interoperability*, ensuring that the different (business) procedures and operations are compatible;

- *legal interoperability*, ensuring that contractual statements are legally equivalent.

Different federation architectures can be designed for building data spaces ranging from architectures with some central components (e. g., a data space catalogue) to fully decentralised ones. Whatever the architectural choice, data spaces will promote data sovereignty, enhance data exchange and trading, and enable the creation of value from data. The Language Data Space, coupled with the data space-inherent data integration capabilities, and developments in machine learning, deep learning, transfer learning and federated learning is expected to help fill in the gaps. Of key importance in the development of language data spaces is the compliance of data and operations with the rules, regulations and values of the European Union. LTs themselves are expected to play a crucial role in ensuring such compliance. Privacy preservation technologies, such as data anonymisation technologies and ethics compliance (through bias detection technologies, say), will be important tools in the hands of data providers, data consumers and data space operators. By its nature, the Language Data Space is conceived of as one of the horizontal data spaces in the data space ecosystem designed by the European Commission. In addition to the intra-data space interoperability, the Language Data Space will have to ensure interoperability with vertical data spaces (e. g., health, manufacturing, skills, mobility, etc), enabling cross-fertilisation, data discovery, exchange and trading at the inter-data space level.

Zooming out of the data spaces discourse and moving to technology visions regarding data access and sharing in general, one of the top-10 data and analytics technology trends identified by Gartner Research is the notion of a *Semantic Data Fabric*.[58] Although the notion was already identified in 2019, they predicted that the first real-world implementations would not be available before 2023. According to Gartner Research,[59] a data fabric enables frictionless access and sharing of data in a distributed data environment. It enables a single consistent data management framework, which allows seamless data access and processing by design across otherwise siloed storage. In the coming years, bespoke data fabric designs will be deployed primarily as a static infrastructure, forcing organisations into a new wave of costs to completely redesign their infrastructures for more dynamic data-mesh approaches. A data fabric must have the ability to collect and analyse all forms of metadata, and analyse and convert passive metadata to active metadata. It must have the ability to create a KG that can operationalise the data fabric design, and enable users to enrich data models with semantics. Extreme levels of distribution, scale and diversity of data assets add complexity to data integration rendering necessary a strong data integration backbone to enable versatile data sharing.

[58] https://www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo

[59] https://www.gartner.com/en/documents/3978267/data-fabrics-add-augmented-intelligence-to-modernize-you

3.4 Towards Deep Natural Language Understanding

Several areas of this deep dive on data, LRs, and KGs have already provided an overview of the state-of-the-art, a gap analysis and an outlook towards deep NLU. The way to help achieve deep NLU is once again by enabling the previously listed components for data, i. e., availability and accessibility of data and metadata; quality of data; interoperability; licences and data-related regulations; data and ethics; and data literacy. Related to these components, where data and metadata are concerned, data infrastructures, data spaces and data markets, integrating KGs, semantic AI and innovative data and metadata management tools need to be built. Furthermore, the following areas are of great importance: the ability to model emotions and culture-specific phenomena to facilitate cross-cultural understanding; the availability of world- and situation-specific knowledge in as many languages as possible; and of course tools that allow the modelling as well as the continuous learning of such attributes need to be built.

Continuous adaptation of LRs in all languages via automated and handcrafted mechanisms is key for deep NLU, to ensure new concepts and terminology are immediately taken into account and provided in monolingual, bilingual, and multilingual formats to ensure that new topics (like the COVID-19 pandemic) can be handled properly, but also so that the impact can be fully understood by a broad population to avoid bias, for example. Issues in digital language inequality will clearly support the division of societies, which needs to be avoided at all cost given the precarious times we live in, and the global nature of the problems we all face.

4 Summary and Conclusions

Data, LRs, and KGs form the basis and backbone for LTs. We identified the following main components: availability and accessibility of data and metadata; quality of data; data interoperability; licensing and data-related regulations; data and ethics; and data literacy. All of these need to be tackled in the future to allow data collection and provision with fair conditions and costs for all relevant stakeholders to help bring about DLE in Europe. Related to these components, where data and metadata are concerned, we identified the following technology concepts, methodologies and tools, that are currently on the rise and that are also part of our technology vision for 2030: data infrastructures, data spaces and data markets; KGs; semantic AI; and innovative data and metadata management tools.

As an add-on component, we tackled the topic of data-related business models, as we identified the importance of sustainable data-related business models as a prerequisite for a working data economy and ecosystem that stimulate and foster the above-listed data-related components in a well-functioning LT landscape.

Besides technology, interoperability and data-related aspects, there must be a strong focus on applying all these mechanisms and methodologies to the widest range of languages possible, at least to the EU-24 languages but also regional and minority languages and also local dialects, as well as to non-European languages that are widespread across Europe. Without such data and LRs in place, DLE simply cannot be achieved.

To fill the identified gaps in data, LRs and KGs, we recommend a future path for Europe towards comprehensive and interlinked data infrastructures, which provide interoperability out-of-the-box by following harmonised and well-tested standards, regarding 1. (semantic) data interoperability as well as 2. services and 3. innovative data and metadata management tools available in all phases of the data lifecycle.

Metadata, data, data-driven tools and services need to be easily integratable into these data infrastructures, without today's huge efforts in data cleaning and integration, or service and tool integration. This future technology vision of integrated and interoperable data spaces follows the approach of federated architectures interlinking data providers and consumer spaces in a trusted framework. Existing data platforms and infrastructures as well as newly developed ones should be integrated where appropriate and possible.

In such a federated ecosystem, data regarding a domain or language can easily be identified, used, re-used, and evaluated for specific use-cases. Data-driven services can be delivered to meet an end-user's requirements. Crowdsourcing and citizen science mechanisms will allow human-machine interaction to foster data acquisition, cleaning and enrichment (e. g., annotation, classification, quality validation and repair, domain-specific model creation, etc.). Raw data can be loaded into available tools to build models for specific use-cases, but also existing algorithms, models or vocabularies will be available for easy loading and re-use to avoid unnecessary energy consumption/computing power to deliver energy-efficient data management.

A high level of importance needs to be placed on privacy protection (related to personal identifiable information, PII, and beyond) and the avoidance of bias (e. g., on gender), and the respective privacy preservation and ethics compliance technologies should be available to all stakeholders.

Data infrastructures require working and sustainable business models that promote data sovereignty, enable data trading, sharing and collaboration. Policies and sustainable data governance models around data creation, data provision and data sharing will be needed. Targeted publicly funded programmes and activities in the area of data literacy are needed from early education onward, to ensure that sufficient human resources in the field are available in the future.

In addition, we need to invest in the collection and development of data and LRs that are relevant for LT to ensure the availability of sufficient data in all EU languages. We make recommendations in three areas: 1. targeted national and European funding along a matrix of relevant resources and languages, combined with 2. more measures in the fields of crowdsourcing and citizen science, and 3. the development of functioning data-related business models, all of which are of critical importance (see Chapter 45).

Europe has a number of difficult problems to solve if DLE is to be achieved, including 1. the specifics of the European language space with EU official languages, a broad range of dialects and regional languages, as well as a high number of non-EU languages in use by a growing number of citizens across the continent, 2. the

European societal characteristics with a rich variety and diversity in culture and society, and 3. the overall challenging requirements of the continuous digitisation in a more and more globalised world, and the related critical need for an efficient, working (language) data infrastructure, that provides a rich, easy-to-use and sustainable backbone for European LT. Despite these challenges, there is a huge potential to become a world leader in LT and a role model for DLE if they can be overcome.

The availability of high-quality data, LRs and KGs in as many languages as possible, that are easily accessible with fair conditions and costs in a clearly specified legal environment providing transparent rules and regulations, has clear benefits and brings with it a competitive advantage for all stakeholders. For the European research community to foster innovations in the field, for the European industry to successfully compete in a global market, and for the benefit of European citizens and society, data, LRs, and KGs are crucial if European DLE is to be achieved.

References

Abu-Salih, Bilal (2021). “Domain-Specific Knowledge Graphs: A Survey”. In: *Journal of Network and Computer Applications* 185, p. 103076.

Auer, Sören, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal (2018). “Towards a Knowledge Graph for Science”. In: *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. WIMS ’18. Novi Sad, Serbia: Association for Computing Machinery. https://doi.org/10.1145/3227609.3227689.

Blumauer, Andreas and Helmut Nagy (2020). *Knowledge Graph Cookbook: Recipes for Knowledge Graphs that work*. https://www.poolparty.biz/the-knowledge-graph-cookbook/.

Cattaneo, Gabriella, Giorgio Micheletti, Mike Glennon, Carla La Croce, and Chrysoula Mitta (2020). *The European Data Market Monitoring Tool: Key Facts & Figures, First Policy Conclusions, Data Landscape and Quantified Stories: D2.9 Final Study Report*. Publications Office. DOI: 10.2759/72084.

Colon-Hernandez, Pedro, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal (2021). “Combining Pre-Trained Language Models and Structured Knowledge”. In: *arXiv preprint arXiv:2101.12294*.

Elliot, Bern, Anthony Mullen, Adrian Lee, and Stephen Emmott (2021). *Gartner Research: Hype Cycle for Natural Language Technologies*.

Hinrichs, Erhard and Steven Krauwer (2014). “The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland: ELRA, pp. 1525–1531. http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf.

Ji, Shaoxiong, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu (2022). “A Survey on Knowledge Graphs: Representation, Acquisition, and Applications”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2, pp. 494–514. DOI: 10.1109/TNNLS.2021.3070843.

Kaltenböck, Martin, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg (2022). *Deliverable D2.16 Technology Deep Dive – Data, Language Resources, Knowledge Graphs*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/data-knowledge-deep-dive.pdf.

Labropoulou, Penny, Stelios Piperidis, Miltos Deligiannis, Leon Voukoutis, Maria Giagkou, Ondřej Košarko, Jan Hajič, and Georg Rehm (2023). "Interoperable Metadata Bridges to the wider Language Technology Ecosystem". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 107–127.

Li, Xinyu, Mengtao Lyu, Zuoxu Wang, Chun-Hsien Chen, and Pai Zheng (2021). "Exploiting Knowledge Graphs in Industrial Products and Services: A Survey of Key Aspects, Challenges, and Future Perspectives". In: *Computers in Industry* 129, p. 103449.

Lösch, Andrea, Valerie Mapelli, Stelios Piperidis, Andrejs Vasiļjevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri, and Josef van Genabith (2018). "European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management". In: *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), pp. 1339–1343.

Noy, Natalya and Deborah L. McGuinness (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. Stanford Knowledge Systems Laboratory. http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html.

Piperidis, Stelios (2012). "The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: ELRA.

Piperidis, Stelios, Penny Labropoulou, Miltos Deligiannis, and Maria Giagkou (2018). "Managing Public Sector Data for Multilingual Applications Development". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by 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. Miyazaki, Japan: ELRA. http://www.lrec-conf.org/proceedings/lrec2018/pdf/648.pdf.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch (2020a). "European Language Grid: An Overview". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3359–3373. https://www.aclweb.org/anthology/2020.lrec-1.413/.

Rehm, Georg, Dimitrios Galanis, Penny Labropoulou, Stelios Piperidis, Martin Welß, Ricardo Usbeck, Joachim Köhler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julián Moreno-Schneider, Florian Kintzel, Elena Montiel, Víctor Rodríguez Doncel, John P. McCrae, David Laqua, Irina Patricia Theile, Christian Dittmar, Kalina Bontcheva, Ian Roberts, Andrejs Vasiljevs, and Andis Lagzdiņš (2020b). "Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability". In: *Proc. of the 1st Int. Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020)*. Ed. by Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs. Marseille, France, pp. 96–107. https://www.aclweb.org/anthology/2020.iwltp-1.15.pdf.

Sebastian-Coleman, Laura (2012). *Measuring Data Quality for Ongoing Improvement: a Data Quality Assessment Framework*. Newnes.

Soylu, Ahmet, Oscar Corcho, Brian Elvesæter, Carlos Badenes-Olmedo, Francisco Yedro Martínez, Matej Kovacic, Matej Posinkovic, Ian Makgill, Chris Taggart, Elena Simperl, Till C. Lech, and Dumitru Roman (2020). "Enhancing Public Procurement in the European Union Through Constructing and Exploiting an Integrated Knowledge Graph". In: *The Semantic Web – ISWC 2020 – 19th International Semantic Web Conference, 2020, Proceedings*. LNCS. Germany: Springer, pp. 430–446.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* 3. DOI: 10.1038/sdata.2016.18. http://www.nature.com/articles/sdata201618.