EUROPEAN LANGUAGE EQUALITY

# D2.16

# Technology Deep Dive – Data, Language Resources, Knowledge Graphs


| | |
|---|---|
| Authors | Martin Kaltenboeck, Artem Revenko with input from 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, Claudia Borg |



| | |
|---|---|
| Dissemination level | Public |
| Date | 28-02-2022 |

# About this document

| | |
|---|---|
| Project | European Language Equality (ELE) |
| Grant agreement no. | LC-01641480 – 101018166 ELE |
| Coordinator | Prof. Dr. Andy Way (DCU) |
| Co-coordinator | Prof. Dr. Georg Rehm (DFKI) |
| Start date, duration | 01-01-2021, 18 months |
| Deliverable number | D2.16 |
| Deliverable title | Technology Deep Dive – Data, Language Resources, Knowledge Graphs |
| Type | Report |
| Number of pages | 67 |
| Status and version | Final |
| Dissemination level | Public |
| Date of delivery | Contractual: 28-02-2022 – Actual: 28-02-2022 |
| Work package | WP2: European Language Equality – The Future Situation in 2030 |
| Task | Task 2.3 Science – Technology – Society: Language Technology in 2030 |
| Authors | Martin Kaltenboeck, Artem Revenko with input from 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, Claudia Borg |
| Reviewers | Teresa Lynn, German Rigau, Stelios Piperidis. |
| EC project officers | Susan Fraser, Miklos Druskoczi |
| Contact | European Language Equality (ELE)<br>ADAPT Centre, Dublin City University<br>Glasnevin, Dublin 9, Ireland<br>Prof. Dr. Andy Way – andy.way@adaptcentre.ie<br>European Language Equality (ELE)<br>DFKI GmbH<br>Alt-Moabit 91c, 10559 Berlin, Germany<br>Prof. Dr. Georg Rehm – georg.rehm@dfki.de<br>http://www.european-language-equality.eu |
| |  |

# Consortium

| | | | |
|---|---|---|---|
| 1 | Dublin City University (Coordinator) | DCU | IE |
| 2 | Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator) | DFKI | DE |
| 3 | Univerzita Karlova (Charles University) | CUNI | CZ |
| 4 | Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis | ILSP | GR |
| 5 | Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country) | UPV/EHU | ES |
| 6 | CROSSLANG NV | CRSLNG | BE |
| 7 | European Federation of National Institutes for Language | EFNIL | LU |
| 8 | Réseau européen pour l'égalité des langues (European Language Equality Network) | ELEN | FR |
| 9 | European Civil Society Platform for Multilingualism | ECSPM | DK |
| 10 | CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium | CLARIN | NL |
| 11 | Universiteit Leiden (University of Leiden) | ULEI | NL |
| 12 | Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH) | ERSCM | DE |
| 13 | Stichting LIBER (Association of European Research Libraries) | LIBER | NL |
| 14 | Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.) | WMD | DE |
| 15 | Tilde SIA | TILDE | LV |
| 16 | Evaluations and Language Resources Distribution Agency | ELDA | FR |
| 17 | Expert System Iberia SL | EXPSYS | ES |
| 18 | HENSOLDT Analytics GmbH | HENS | AT |
| 19 | Xcelerator Machine Translations Ltd. (KantanMT) | KNTN | IE |
| 20 | PANGEANIC-B. I. Europa SLU | PAN | ES |
| 21 | Semantic Web Company GmbH | SWC | AT |
| 22 | SIRMA AI EAD (Ontotext) | ONTO | BG |
| 23 | SAP SE | SAP | DE |
| 24 | Universität Wien (University of Vienna) | UVIE | AT |
| 25 | Universiteit Antwerpen (University of Antwerp) | UANTW | BE |
| 26 | Institute for Bulgarian Language "Prof. Lyubomir Andreychin" | IBL | BG |
| 27 | Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences) | FFZG | HR |
| 28 | Københavns Universitet (University of Copenhagen) | UCPH | DK |
| 29 | Tartu Ulikool (University of Tartu) | UTART | EE |
| 30 | Helsingin Yliopisto (University of Helsinki) | UHEL | FI |
| 31 | Centre National de la Recherche Scientifique | CNRS | FR |
| 32 | Nyelvtudományi Kutatóközpont (Research Institute for Linguistics) | NYTK | HU |
| 33 | Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies) | SAM | IS |
| 34 | Fondazione Bruno Kessler | FBK | IT |
| 35 | Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia) | IMCS | LV |
| 36 | Lietuvių Kalbos Institutas (Institute of the Lithuanian Language) | LKI | LT |
| 37 | Luxembourg Institute of Science and Technology | LIST | LU |
| 38 | Università ta Malta (University of Malta) | UM | MT |
| 39 | Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute) | INT | NL |
| 40 | Språkrådet (Language Council of Norway) | LCNOR | NO |
| 41 | Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences) | IPIPAN | PL |
| 42 | Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science) | FCULisbon | PT |
| 43 | Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy) | ICIA | RO |
| 44 | University of Cyprus, French and European Studies | UCY | CY |
| 45 | Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences) | JULS | SK |
| 46 | Institut Jožef Stefan (Jozef Stefan Institute) | JSI | SI |
| 47 | Centro Nacional de Supercomputación (Barcelona Supercomputing Center) | BSC | ES |
| 48 | Kungliga Tekniska högskolan (Royal Institute of Technology) | KTH | SE |
| 49 | Universität Zürich (University of Zurich) | UZH | CH |
| 50 | University of Sheffield | USFD | UK |
| 51 | Universidad de Vigo (University of Vigo) | UVIGO | ES |
| 52 | Bangor University | BNGR | UK |

# Contents

# List of Figures



# List of Tables



# List of Acronyms

| | |
|---|---|
| AI | Artificial Intelligence |
| AI4EU | AI4EU (EU project, 2019-2021) |
| API | Application Programming Interface |
| CALL | Computer Assisted Language Learning |
| CEF AT | Connecting Europe Facility, Automated Translation |
| DCAT | Data Catalogue Vocabulary |
| DGA | Data Governance Act |
| DL | Deep Learning |
| DMA | Data Market Austria |
| DSM | Digital Single Market |
| EC | European Commission |
| ECCMA | Electronic Commerce Code Management Association |
| EDM | Enterprise Data Management |
| EIM | Enterprise Information Management |
| ELE | European Language Equality *(this project)* |
| ELE Programme | European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)* |
| ELG | European Language Grid (EU project, 2019-2022) |
| ELRA | European Language Resource Association |
| ELRC | European Language Resource Coordination |

| | |
|---|---|
| EOSC | European Open Science Cloud |
| EMM | Enterprise metadata management |
| EP | European Parliament |
| EU | European Union |
| EUDAT | European Data (Infrastructure) |
| GDPR | General Data Protection Regulation |
| HPC | High-Performance Computing |
| IDSA | International Data Space Association |
| IT | Information Technology |
| LT | Language Technology/Technologies |
| MDM | Master Data Management |
| ML | Machine Learning |
| MT | Machine Translation |
| NER | Named Entity Recognition |
| NLG | Natural Language Generation |
| NLP | Natural Language Processing |
| NLT | Natural Language Technologies |
| NLU | Natural Language Understanding |
| PII | Personal Identifiable Information |
| PSI | Public Sector Information |
| RML | Regional and Minority Languages |
| SSH | Social Sciences and the Humanities |
| SSHOC | Social Sciences and the Humanities Open Cloud |
| W3C | World Wide Web Consortium |

## Abstract

This technology deep dive on: data, language resources, and Knowledge Graph provides a kind of an add-on to the other deep dives of the ELE project, as data as well as related models build the basis for technologies and solutions in the area of (Natural) Language Technology and thereby of Digital European Language Equality.

This report provides a technology deep dive with a clear focus on data and language resources required for a full language equality in Europe by 2030. The document on hand provides insights into: (i) the main components of this technology deep dive, (ii) the current state of the art, (iii) the main gaps identified in the field, (iv) a chapter about the contribution to digital language equality and the impact on society, (v) an analysis of the main breakthroughs needed in the area of data, language resources, and Knowledge Graphs, and (vi) the main technology visions and development goals identified, as$f$ well as (vii) a chapter in regard to deep natural language understanding and data, and is closed by (viii) a summary and conclusions section.

The methodology for the creation of this deliverable has taken into account (i) desktop research, (ii) two virtual workshops with ELE consortium members on the topics (a) state of the art and main gaps and (b) future use cases and requirements regarding data, as well as future technology visions, and (iii) discussions with industry representatives about the topics listed in (ii).

The main components identified for this technical deep dive: Data, Language Resources, Knowledge Graphs have been identified, and are being (i) explained and specified, (ii) used for a state of the art analysis, and (iii) a gap analysis.

All of these components need to be tackled in the future, and for the widest range of languages possible, from EU over dialects to non-EU languages used in Europe, to allow efficient data collection and sustainable data provision with fair conditions and costs to develop towards a digital European Language Equality.

In addition specific technology concepts, methodologies and tools have been identified to be part of the technology vision for 2030 for data and language resources.

Finally the topic of data-related business models, as well as data governance models are tackled, that build a prerequisite for a working data economy that stimulate and foster the above listed data related components, and thereby a working language technology landscape as a basis for a digital European language equality.


# 1 Introduction

Multilingualism is a key cultural cornerstone of Europe and signifies what it means to be and to feel European. Many studies and resolutions, as noted in the recent EP resolution "Language equality in the digital age", have found a striking imbalance in terms of support through language technologies and issue a call to action. The ELE project answers this call and lays the foundations for a strategic agenda and roadmap for making digital language equality a reality in Europe by 2030. The primary goal of ELE is to prepare the European Language Equality Programme, in the form of a strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030. This programme will be prepared jointly with the whole European Language Technology, Computational Linguistics and language centric AI community, as well as with representatives of relevant initiatives and associations, language communities and RML groups. The project consortium includes all relevant scientific and industrial stakeholders from all Member States and Associated Countries and engages them in the process. The whole community is included in the project through external consultation sessions. The project plan is

fully optimised towards this key goal of preparing the strategic agenda and roadmap and of involving the whole European LT community. Ensuring appropriate technology support for all European languages will create jobs, growth and opportunities in the digital single market. Equally crucial, overcoming language barriers in the digital environment is essential for an inclusive society and for providing unity in diversity for many years to come. The ELE project provides a roadmap and framework to achieve this.

Digital language equality as well as the European data economy rely on the availability, the interoperability and the structure of data (unstructured, semi-structured, structured data) as a basis for further innovation and exponential development of technologies, especially the development of trustworthy 'made in Europe' AI and powerful language technology that reflects European values. Data spaces[1], platforms[2] and marketplaces are enablers, key to unleash the potential of such data. However, data sharing and data interoperability are still at their infancy. The diffusion of platforms for data sharing and the 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 (DSM) strategy that was adopted on 6 May 2015[3] 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"[4] which outlines 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[5] to become the world's most secure and trustable data hub.

This report provides a technology deep dive with a clear focus on data and language resources required for a full language equality in Europe by 2030. The document on hand provides insights into: (i) the main components of this technology deep dive, (ii) the current state of the art, (iii) the main gaps identified in the field, (iv) a chapter about the contribution to digital language equality and the impact on society, (v) an analysis of the main breakthroughs needed, and (vi) the main technology visions and development goals identified, as well as finally (vii) a summary towards deep natural language understanding and data, and is closed by a summary and conclusions.

The methodology for the creation of this deliverable has taken into account (i) desktop research, (ii) two virtual workshops with ELE consortium members on the topics (a) state of the art and main gaps and (b) future use cases and requirements regarding data, as well as future technology visions, and (iii) discussions with industry representatives about the topics listed in (ii).

The two workshops with project members build the core of the findings presented in this deliverable and have been carried out in November 2021 (04.11.2021 and 18.11.2021) with participants from the organisations: Sirma AI – Ontotext, ELDA, UPV/EHU, DCU, SAP SE, DFKI, INT, CUNI, UM, Bangor University, Wikimedia, and ILSP/ARC. In addition project members of these organisations provided additional input in the form of collected ideas and statements.

The first workshop was organised in the form of statements by the participants and respective counter-statements, the second one was organised in the form of a world cafe in which the participants prepared statements for the following questions:

1. What are the favorite and most important future use cases in language technology for 2030?

[1] Next generation data acquisition and processing platforms as exemplified by the BDVA reference model (https://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf) and its reference implementations

[2] Data sharing and exchange platforms, where data is commercialized using Open Data, Monetized Data and Trusted Data sharing mechanisms.

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

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

[5] EU data strategy (COM(2020) 66)

2. Based on these use cases and scenarios: what is required in regard to data (language resources, knowledge graphs, etc.) to realise such future use cases in 2030?
3. What statements and important factors come to your mind regarding a "Vision towards: 2030 Deep Natural Language Understanding".

The results of this preparation exercise have been collected in the course of the virtual workshop, statements have been presented to the whole group, have been clustered and discussed together. Also here additional statements have been provided by ELE partners beside the workshop. These questions of workshop 2 listed above have also been raised to representatives from industry, namely language service providers, technology and software companies, language resource providers, and publishers.

## 2 Scope of this Deep Dive

The scope of this technical deep dive is on a relatively broad field of technologies in the area of language technology, including: machine translation, learning management systems, language learning systems, content management systems, knowledge management systems, text to speech and speech to text solutions, and/or natural text generation, as data and language resources build the basis and backbone for all of such listed technologies and beyond. In addition the area of Knowledge Graphs takes an important role in this deep dive as Knowledge Graphs provide powerful mechanisms and principles to interlink and enrich data in a high quality manner. Thereby Knowledge Graphs 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 data infrastructure element to develop future Language Technology Solutions. As such solutions require not only a single underlying dataset but a wide range of meaningful and contextualised data. In addition the integrated data models inside of Knowledge Graphs (taxonomies, vocabularies and/or ontologies) allow the training of algorithms for Language Technology solutions with higher precision and less training data.

This means that the subject of metadata and data in this technical deep dive is always related to language technology, natural language technology, language understanding and digital European language equality.

Thereby metadata and data in this respect means mainly – but is not limited to – data and metadata like / about: language resources, (annotated) corpora, translation memories, language pairs, dictionaries and lexicographic resources, as well as other language resources and relevant data that is required for powerful multilingual and natural language applications. Such data and metadata is a strong enabler of artificial intelligence and machine learning that both have enabled innovative approaches and advances in the field of natural language technologies (NLT) (Elliot et al., 2021).

The main components of this technical deep dive: Data, Language Resources, Knowledge Graphs have been identified and will be (i) explained and specified in the following section, as well as (ii) used for a state of the art analysis, and (iii) a gap analysis in the sections to come.

In addition to these components related technology concepts, methodologies and tools have been identified, that are currently on the rise and are part of the technology vision for 2030 in this deep dive document.

As an add-on component in this deep dive the topic of **data-related business models** is tackled throughout the document, as we have identified the importance of working and sustainable data-related business models as a prerequisite for a working data economy and ecosystem that thereby stimulates and fosters the above listed data related components, and

thereby a working language technology landscape as a basis for a digital European language equality.

Furthermore the main (broad) identified technology areas – also used for the technology visions in the field – that has been identified for this technical deep dive are:

**A) Conversational AI** – as defined by IBM Research as follows[6]:

> Conversational artificial intelligence (AI) refers to technologies, like chatbots or virtual agents, which users can talk to. They use large volumes of data, machine learning, and natural language processing to help imitate human interactions, recognizing speech and text inputs and translating their meanings across various languages.

Conversational AI provides its power often through Conversational Platforms – as defined by Gartner Research as follows:

> Conversational platforms can be used by developers to build conversational user interfaces, chatbots and virtual assistants for integration into messaging platforms, social media, SMS, website chat. The offerings include a development platform to build conversational interfaces with strong NLP engines, supporting voice and text input modalities. The platform provides capabilities like dialogue management, multiple chatbots orchestration, training data maintenance.
>
> Gartner, Market Guide for Conversational Platforms, published: 30 July 2019, ID: G00367775, Analyst(s): Magnus Revang, Van Baker, Brian Manusama, Anthony Mullen, Adrian Lee.

**B) Insight Engines** – as defined by Gartner Research as follows[7]:

> Insight engines apply relevancy methods to describe, discover, organize and analyze data. This allows existing or synthesized information to be delivered proactively or interactively, and in the context of digital workers, customers or constituents at timely business moments. Products in this market use connectors to crawl and index content from multiple sources. They index the full range of enterprise content, from unstructured content such as word processor and video files through to structured content, such as spreadsheet files and database records. Various "pipelines" are used to preprocess content according to type, and to derive from it data that can be indexed for query, extraction and use via a range of touchpoints. Insight engines differ from search engines in terms of capabilities that enable richer indexes, more complex queries, elaborated relevancy methods, and multiple touchpoints for the delivery of data (for machines) and information (for people).

Both technology areas have been identified as being highly relevant regarding language technology – thereby the focus of the document in regard to the Technology Vision is on multi- and cross-lingual Conversational AI and -Platforms and Insight Engines, as well as the additional requirement of deep natural language understanding.

But the overall focus of this technology deep dive is on the prerequisite for these Technology Visions that are: data, language resources, and Knowledge Graphs.

For the sake of clarity: this deep dive does not tackle any domains or industries in detail, as we see language technology as kind of a horizontal technology through nearly every domain, from health and pharmacy, over publishing, broadcasting and the legal domain, to

[6] https://www.ibm.com/cloud/learn/conversational-ai
[7] https://www.gartner.com/en/information-technology/glossary/insight-engines

construction, finance and insurance and the public administration – to just name a few core industries – language technology plays a crucial role and has very similar requirements regarding data, language resources and Knowledge Graphs. Thereby specific needs for certain industries and domains has been intentionally left out from this deep dive document.

A clear distinction has been made in the work for this deliverable between the technical deep dive 2.16: Data, Language Resources, Knowledge Graphs (data creation, annotation, curation, preservation, representation) and the other three technological deep dives: (i) Machine Translation, (ii) Speech technologies, (iii) Text Analytics, Text and Data Mining, Natural Language Understanding, but the deliverable on hand can be seen as a kind of an enabler or basis for the other 3 technical deep dive documents, as data builds the backbone of such technologies and solutions.

# 3 Data, Language Resources, Knowledge Graphs: Main Components

The main components of this technical deep dive: Data, Language Resources, Knowledge Graphs have been identified as follows:

- Availability of data and metadata
- Accessibility of data
- Quality of data
- Data Interoperability
- Licenses and data related regulations
- Data and ethics
- Data literacy

And related to these components the following related technology concepts, methodologies and tools in addition:

- Data infrastructures, data spaces and datamarkets
- Knowledge Graphs
- Semantic AI: statistical and symbolic AI in combination
- Innovative data and metadata management tools

These main components are described as follows in this section and are used for Section 4 as well as for Section 5, and include always: structured data, semi-structured data, and unstructured data. Furthermore it includes the different modalities of data: written, spoken, signs and others.

Before going into the single components listed above the following section provides (i) an overview explaining “the language data concept” as well as (ii) some definitions and basic information about the terms and concepts: data and metadata.

**The Language Data Concept** (Natural) Language Technology requires a range of specific language data resources that can be used to develop working monolingual, multilingual and cross-lingual applications. This starts from monolingual data over bilingual to multilingual data sources like, for example, monolingual corpora, bilingual/multilingual corpora (including parallel and/or comparable), monolingual/multilingual lexical and terminological resources. To allow a higher degree of automation in the field, as well as to enable machine learning mechanisms, data in the form of annotated data like annotated corpora is needed. Corpora are hereby enriched with additional annotations to provide machines with additional information, for instance about named entities, syntactic structures, or other application specific annotations (e. g., for summarisation, question-answering applications). Such annotated language data sets enable machines to identify patterns and thereby train models to continuously improve the underlying algorithms. In addition, an important characteristic of language resources is that data is available as written text (here in addition in different character-sets), as audio (speech, video), and signs (sign languages), and language is supported by gesture and facial expression.

Rayan Potter states in his medium article of 2020: "When various types of data like text, audio and video are labeled or annotated with additional metadata or added notes to make the entire sentence or document comprehensible through NLP, NLU or other language based AI model development. And in language annotation, data in different language can be annotated as per the machine learning training AI development requirements.".[8]

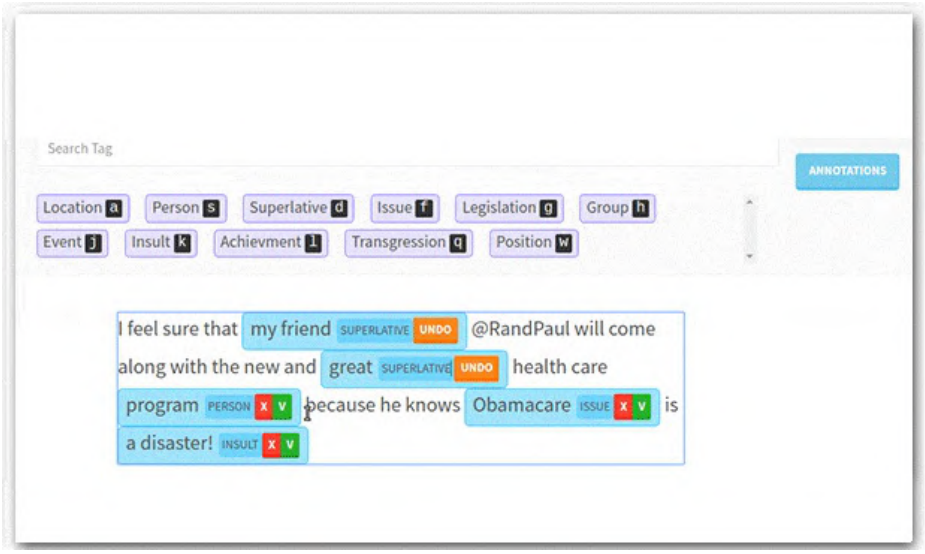


Figure 1: Language Annotation Example.

Like for other technology areas the data for Language Technology is available as raw data, curated data and cleaned data.

With the rise of artificial intelligence and beyond the importance of large language models, of foundation models (like Bidirectional Encoder Representations from Transformers (BERT)[9] or Generative Pre-trained Transformer 3 (GPT-3)[10]) AI models and comprehensive and multilingual Knowledge Graphs – all models that are based on a broad range of domains and/or languages – is continuously increasing. This supports also the creation and thereby

[8] https://medium.com/nerd-for-tech/what-is-language-data-annotation-and-how-it-is-useful-in-machine-learning-ai-dbd727a8207c
[9] https://en.wikipedia.org/wiki/BERT_(language_model)
[10] https://en.wikipedia.org/wiki/GPT-3

availability of annotated corpora and beyond.

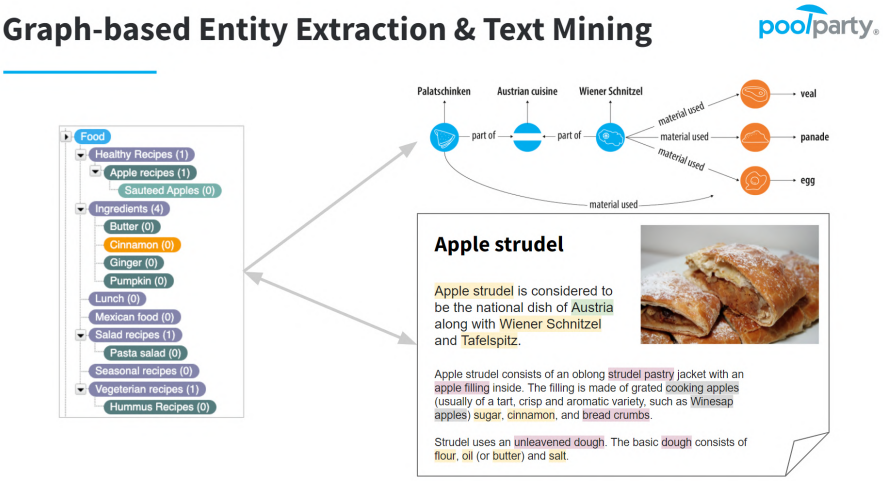


Figure 2: Graph-based Entity Extraction and Text Mining.

For all these language resources and data types there is the additional requirement for domain specific data in place, that allows to develop domain and industry specific applications in the respective field as of the very much specialised language and terminology in industries like for instance in health, pharmacy or finance.

### Data and Metadata Definitions

> *Data* (US: /ˈdæˈtə/; UK: /ˈdeɪˈtə/) are individual facts, statistics, or items of information, often numeric. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects, while a datum (singular of data) is a single value of a single variable.
>
> Although the terms “data” and “information” are often used interchangeably, this term has distinct meanings. In some popular publications, data are sometimes said to be transformed into information when they are viewed in context or in post-analysis.[11] However, in academic treatments of the subject data are simply units of information. Data are used in scientific research, businesses management (e. g., sales data, revenue, profits, stock price), finance, governance (e. g., crime rates, unemployment rates, literacy rates), and in virtually every other form of human organizational activity (e. g., censuses of the number of homeless people by non-profit organizations).
>
> Data are measured, collected, reported, and analyzed, and used to create data visualizations such as graphs, tables or images. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing. Raw data (“unprocessed data”) is a collection of numbers or characters before it has been “cleaned” and corrected by researchers. Raw data needs to be corrected to remove outliers or obvious instrument or data entry errors (e. g., a thermometer reading from an outdoor Arctic location recording a tropical temperature). Data processing commonly occurs by stages, and the “processed data” from one stage may be considered the “raw data”

[11] https://www.diffen.com/difference/Data_vs_Information

> of the next stage. Field data is raw data that is collected in an uncontrolled "in situ" environment. Experimental data is data that is generated within the context of a scientific investigation by observation and recording.
>
> Data has been described as the new oil of the digital economy[12] (Co-operation and Development, 2008).
>
> —https://en.wikipedia.org/wiki/Data

> *Metadata* is "data that provides information about other data", but not the content of the data, such as the text of a message or the image itself. There are many distinct types of metadata, including:
>
> **Descriptive metadata** — the descriptive information about a resource. It is used for discovery and identification. It includes elements such as title, abstract, author, and keywords.
>
> **Structural metadata** — metadata about containers of data and indicates how compound objects are put together, for example, how pages are ordered to form chapters. It describes the types, versions, relationships and other characteristics of digital materials.
>
> **Administrative metadata** — the information to help manage a resource, like resource type, permissions, and when and how it was created.
>
> **Reference metadata** — the information about the contents and quality of statistical data.
>
> **Statistical metadata** – also called process data, may describe processes that collect, process, or produce statistical data.
>
> **Legal metadata** — provides information about the creator, copyright holder, and public licensing, if provided.
>
> Metadata is not strictly bound to one of these categories, as it can describe a piece of data in many other ways.[13]

Finally it is important to make a distinction between *data*, *information*, and *knowledge* in the field of computer science to better understand the following sections of this document on hand. This is done by citing Dan McCreary, who specified the concept of the Knowledge Triangle in Medium.com as follows:

*The Knowledge Triangle*[14] specifies knowledge – for the area of computer science – as: Knowledge which is connected-information that is query ready, see Figure 3.

This triangle explains the journey that raw data (raw binary data as flat .csv files, numeric codes and strings) takes to (i) become information, by being analysed and processed in a way to identify and extract 'isolated business entities' (for instance persons, organisations or places), and (ii) how information is interlinked and put into context to become knowledge (as specified above), see Figure 3.

## 3.1 Components in Detail

With these definitions and basic concepts in place, the next step is to take a look at the relevant components in this field – that are strongly interconnected with each other – as being specified:

[12] https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language+-+what+are+data
[13] https://en.wikipedia.org/wiki/Metadata
[14] https://dmccreary.medium.com/the-knowledge-triangle-c5124637d54c

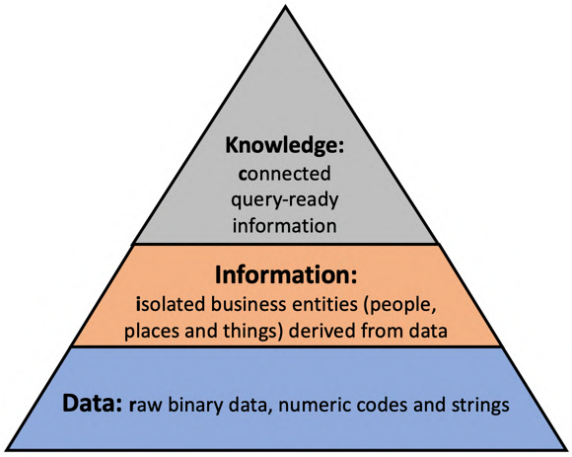


Figure 3: The Knowledge Triangle – a graph technologies metaphor where raw data is converted into information about people, places, and things and connected into a query-ready graph.

**Availability of data and metadata** As data and metadata build the backbone of any (natural) language technology, the availability of data and metadata is the overall basis to enable such technologies and services. Availability thereby is about data collections, data types available, and how to find and explore such data.

For the last few years a "big data hype" was established that slowly is replaced these days by the more important argument of "value of data". This means that often the volume of data was the most recognised attribute of the 4+1Vs in big data (volume, variety, velocity, veracity and value) that have been specified by IBM in 2019 (Leslie and Johnson-Leslie, 2020). The argument that availability of volume of data is still in use today, also as of the new movement in artificial intelligence, where big amounts of training data is required to provide precise and useful algorithms. Thereby availability of data for (natural) language technologies is key.

**Accessibility of data** Aside from the overall availability, the attribute of accessibility is the next most important component when speaking about data. This component of accessibility is also reflected by the FAIR principles that have been stated initially for scientific data:[15]

> In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in Scientific Data. The authors intended to provide guidelines to improve the **Findability**, **Accessibility**, **Interoperability**, and **Reuse** of digital assets. The principles emphasise machine-actionability (i. e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Accessibility has also been specified as one of the initial 8 key principles of open (government) data in 2007.[16]

[15] https://www.go-fair.org/fair-principles/
[16] https://opengovdata.org

**Quality of data** When data is available and accessible one can think about 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, the quality of the data is a crucial issue regarding value.

Dimensions to measure data quality often include – but are not limited to: completeness, validity, timeliness, consistency, and integrity (Sebastian-Coleman, 2012). Also reliability is an important factor of data quality, although it is hard to measure. All in all the quality of a language technology application is often based on the quality of the underlying / used data.

Data Governance: is an additional important concept centered around data quality and much more.

Data Governance is a collection of components – data, roles, processes, communications, metrics, and tools – that help organizations formally manage and gain better control over data assets. As a result, organizations can best balance security with accessibility and be compliant with standards and regulations while ensuring data assets go where the business needs them most.

Outcomes for better data control lead to efficient methods, technologies, and behaviors around the proper management of data, across all levels of the organization. From the senior leadership team to daily operations, governance ensures alignment by providing structure and services.

Data Governance often includes other concepts such as Data Stewardship and Data Quality. These bases help connect governance details with the data lifecycle, improving data integrity, usability, and integration. Both internal and external data flows, within an organization, fall under the jurisdiction of governance.[17]

### Data Interoperability

> Interoperability is a characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, at present or in the future, in either implementation or access, without any restrictions.[18]

Data interoperability is defined as:[19]

> Data interoperability addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data.

Interoperability thereby ensures the seamless interplay of different (natural) language technology systems regarding interfaces and regarding data. And it is often connected with the requirement of related standards in the field. Interoperability is not an objective by itself, it is always a vehicle to enable something in addition, as for instance easy data integration of heterogeneous data from different sources, what is a crucial task for working language technology systems, which ingest and make use of data from federated sources.

**Licenses and data related regulations** Following the already specified attributes of availability and accessibility of data one comes across licenses and data related policies and regulations. Since the open (government) data movement that started around 2007 there is an increasing awareness of the importance of assigning a clear license to any data(set) that clarifies the rights as well as the responsibilities of data publishers and data consumers what can be done with specific data.

[17] https://www.dataversity.net/what-is-data-governance/
[18] https://en.wikipedia.org/wiki/Interoperability
[19] https://datainteroperability.org

Relevant data often comes from different owners and publishers like companies, public administration or citizens, with different licenses. Thereby proper license clearing is a crucial task for all data related activities in language technology.

The licenses on data that are usually specified by the data owners / publishers as well as the applicable law and regulations around data, like the ones about data privacy, security, the data processing and protection of personal identifiable information (PII), as for instance the General Data Protection Regulation (GDPR), need to be taken into account as an important component for this deep dive: data, language resources, and knowledge graphs.

Finally regional, national and international regulations and policies around data should be taken into account, to ensure to be able to develop along the latest trends and regulations.

**Data and ethics** With the rise of artificial intelligence (AI) and machine learning, as well as the overall movement of data collection and processing, the component of: data and ethics becomes more and more important. It is strongly related with the already listed components of data interoperability as well as with licenses and data related regulations.

Language by itself is ambiguous in its meaning and thereby can easily create bias. Thereby ethics play a crucial role regarding the use of data in (natural) language technologies and regarding a digital European language equality.

**Data literacy** Gartner 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."

Furthermore: "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".[20]

Wikipedia defines it in a similar but slightly different language:[21]

> Data literacy is the ability to read, understand, create, and communicate data as information. Much like literacy as a general concept, data literacy focuses on the competencies involved in working with data. It is, however, not similar to the ability to read text since it requires certain skills involving reading and understanding data.

## 3.2 Related Technology Concepts, Methodologies and Tools

**Data infrastructures, data spaces and datamarkets** The overall idea behind data spaces and datamarkets follow the principle idea of data catalogues established in the course of the open data movement in the last 14 years to allow sharing and the exchange of / to trade data. And thereby to enable availability of and allow accessibility to high quality data, that is following certain standards (and thereby provides data interoperability), and have a clear license.

The GAIA-X initiative defines a data space as follows:[22]

> In general, the term "data space" refers 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

[20] https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy.
[21] https://en.wikipedia.org/wiki/Data_literacy
[22] https://www.gaia-x.eu/what-is-gaia-x/data-spaces

only shared (via semantic interoperability) when necessary. In the Gaia-X context, the data in the data spaces is held exclusively by the members of the Association.

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.

**Relevancy for Language Technology and Language Equality** Working Language Technology solutions require high quality data in scale and a broad range of domains and available in various languages, with clear licenses and fair conditions. Data infrastructures, data spaces and datamarkets provide a powerful infrastructure to find, evaluate and access relevant data as well as often related data-driven services, that are required for Language Technology solutions. Data spaces often follow the principles of findability, availability, interoperability and re-usability (the FAIR-principles) and thereby support Language Equality.

**Knowledge Graphs** Dan McCreary, provides the following definition of Knowledge in the context of AI and graph databases,[23] bringing us back to the concept of the Knowledge Triangle, that is of high importance to understand the idea of a Knowledge Graph:

"Knowledge is connected-information that is query-ready."

This definition is much shorter than the Wikipedia Knowledge page, which is: ...a familiarity, awareness, or understanding of someone or something, such as facts, information, descriptions, or skills, is acquired through experience or education by perceiving, discovering, or learning.

The Wikipedia definition is longer, more general, and applicable to many domains like philosophy, learning, and cognitive science. Our definition is shorter and only intended for the context of computing. Our definition is also dependent on how we define "information," "connected," and "query ready." To understand these terms, let's reference the Knowledge Triangle figure above. In the knowledge triangle diagram, let's start at the bottom Data Layer. The data layer contains unprocessed raw information in binary codes, numeric codes, dates, strings, and full-text descriptions that we find in documents. The data layer can also include images (just as a jpeg file), speech (in the form of a sound file), and video data. You can imagine raw data as a stream of ones and zeros. It is a raw dump of data from your hard drive. Some raw data types, such as an image — can be directly understood by a person just by viewing it. Usually, raw data is not typically useful without additional processing. We call this processing of raw data enrichment."

**Definition of a Knowledge Graph** In knowledge representation and reasoning, knowledge graph is a knowledge base that uses a graph-structured data model or topology to integrate data.[24] Knowledge graphs are often used to store interlinked descriptions of entities – objects, events, situations or abstract concepts – while also encoding the semantics underlying the used terminology.[25]

Since the development of the Semantic Web, knowledge graphs are often associated with linked open data projects, focusing on the connections between concepts and entities (Soylu et al., 2020; Auer et al., 2018). They are also prominently associated with and used by search

[23] https://dmccreary.medium.com/the-knowledge-triangle-c5124637d54c
[24] https://en.wikipedia.org/wiki/Knowledge_graph
[25] https://ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph

engines such as Google, Bing, and Yahoo; knowledge-engines and question-answering services such as WolframAlpha, Apple's Siri, and Amazon Alexa; and social networks such as LinkedIn and Facebook.

A more business oriented definition has been provided by Semantic Web Company:[26]

> A Knowledge Graph is a model of a knowledge domain created by subject-matter experts with the help of intelligent machine learning algorithms. It provides a structure and common interface for all of your data and enables the creation of smart multilateral relations throughout your databases. Structured as an additional virtual data layer, the Knowledge Graph lies on top of your existing databases or data sets to link all your data together at scale – be it structured or unstructured.

**Relevancy for Language Technology and Language Equality** Language Technology solutions require not only single encoupled datasets but high quality, interlinked, meaningful and contextualised data that can be easily used, can be quickly expanded and can be efficiently maintained over time with reasonable efforts. Knowledge Graphs provide these characteristics and thereby are an ideal data-backbone for Language Technology and thereby Language Equality.

**Semantic AI: statistical and symbolic AI in combination** Artificial Intelligence (AI) sees a second spring or even summer these days, thanks to the computing power available, as well as the possibilities of state of the art machine learning algorithms. But this AI is often not precise enough in regard to simple challenges that require common sense knowledge or context and meaning (semantics) and thereby "deep language understanding" regarding language technologies.

The approach here is to combine the 2 main fields of AI, namely: statistical AI (machine learning & algorithms) and symbolic AI (models like ontologies, knowledge bases for common sense knowledge, cultural resources) and the term Semantic AI has been rising in the last couple of months.

Puneet Agarwal defined Semantic AI in his medium.com article in October 2020 as follows:[27]

> Semantic AI provides 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.

**Relevancy for Language Technology and Language Equality** Language is a lot about meaning and context, thereby Language Technology Solutions require data that is rich in meaning and context. Furthermore machine learning is an important attribute as well to ensure continuous improvement of such solutions. Semantic AI – as the combination of statistical AI and symbolic AI provides both characteristics mentioned and is thereby a pre-requisite for Language Technology that enables Language Equality.

**Innovative data and metadata management tools** The final component of importance in this area is the availability of innovative data and metadata management tools, that allow availability, accessibility, high quality of data, that enable data interoperability (als by making use of relevant standards), that provide powerful data governance mechanisms (thereby also follow the relevant regulations), and that finally enable mechanisms and features for

[26] https://www.poolparty.biz/what-is-a-knowledge-graph
[27] https://medium.com/@dr.puneet.a/what-is-semantic-ai-is-it-a-step-towards-strong-ai-5f0355be3597

ethics in data and finally allow the improvement of data literacy. In addition such tools should support or can be used in combination with secure data sharing mechanisms (data spaces), provide strong capability for interlinking data and providing meaning and context (knowledge graphs) and provide Semantic AI capability.

Innovative and effective metadata and data management tools can be described along the concept of the data life cycle.

The data life cycle, also called the information life cycle, refers to the entire period of time that data exists in your system. This life cycle encompasses all the stages that your data goes through, from first capture onward.[28]

**Relevancy for Language Technology and Language Equality** Thereby this last component makes use and combines all the other listed components of this technical deep dive: data, language resources and Knowledge Graphs. And enables a way of data- and metadata management that allows to create and provide future (natural) Language Technologies that work and support a digital European Language Equality by 2030.

# 4 Data, Language Resources, Knowledge Graphs: Current State of the Art

## 4.1 Availability of Data and Metadata

The strong movement of data collection and data provision started with the Open Data movement in 2007. On December 7-8, 2007, thirty open government advocates gathered in Sebastopol, California and wrote a set of eight principles of open government data.[29]

The Sunlight Foundation expanded the principles to 10 principles 3 years later.[30] In 2010, the Sunlight Foundation updated and expanded upon the Sebastopol list, and identified ten principles that provide a lens to evaluate the extent to which government data is open and accessible to the public. In 2013, the Sunlight Foundation began maintaining a list of open data policy guidelines that built upon these principles.

These principles build somehow the foundation of data collection and sharing, as follows:

1. Completeness
2. Primacy
3. Timeliness
4. Ease of Physical and Electronic Access
5. Machine readability
6. Non-discrimination
7. Commonly owned or open Standards
8. Licensing
9. Permanence
10. Usage Costs

[28] https://www.talend.com/resources/data-lifecycle-management/
[29] https://opengovdata.org
[30] https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

Lots of research has been carried out in the field since then and several governments but also private and non-government organisations have carried out open data initiatives to collect and provide open data to a broad public.

From this movement of open data the requirements of industry data as well as personal data sharing and collaborating has found its way and culminated in the next area of data sharing: data catalogues and data portals, as well as data spaces and datamarkets (see more details below).

A good example for the outcome of such research activities, supported by the Implementation of the Public Sector Information Directive[31] – is the ELRC-SHARE repository,[32] that is used for documenting, storing, browsing and accessing Language Resources that are collected through the European Language Resource Coordination[33] and considered useful for feeding the CEF Automated Translation (CEF.AT) platform.

## 4.2 Accessibility of Data

In regard to accessibility of data the movement of FAIR Data and Principles, that has been established in the scientific community and now spreads also to other data communities, has become a de facto standard idea.

The FAIR principles are: Findability, Accessibility, Interoperability, and Reuse of digital assets.

ACCESSIBILITY in FAIR:[34] The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

## 4.3 Quality of Data

Quality of data is one of the main pillars for the development of successful data-driven applications, not only in the field of language technology.

Dimensions to measure data quality often include – but are not limited to – completeness, validity, timeliness, consistency, and integrity (Sebastian-Coleman, 2012).

Gartner defines Data Quality Solutions as follows:[35]

> The discipline of data quality assurance ensures that data is “fit for purpose” in the context of existing business operations, analytics and emerging digital business scenarios. It covers much more than just technology. It includes program management, roles, organizational structures, use cases and processes (such as those for monitoring, reporting and remediating data quality issues).

It is also linked to broader initiatives in the field of enterprise information management (EIM), including information governance and master data management (MDM).

[31] https://digital-strategy.ec.europa.eu/en/policies/public-sector-information-directive
[32] https://elrc-share.eu
[33] https://lr-coordination.eu
[34] https://www.go-fair.org/fair-principles/
[35] https://www.gartner.com/reviews/market/data-quality-solutions

As digital business requires innovations in data quality tools, vendors are competing fiercely by enhancing existing capabilities and creating new capabilities in eight key areas: audience, governance, data diversity, latency, analytics, intelligence, deployment and pricing."

For data quality related to data on the web a W3C standard exists already since 2016: Data on the Web Best Practices: Data Quality Vocabulary.[36]

The Electronic Commerce Code Management Association (ECCMA)[37] is a member-based, international not-for-profit association committed to improving data quality through the implementation of international standards. ECCMA is the current project leader for the development of ISO 8000 and ISO 22745, which are the international standards for data quality and the exchange of material and service master data, respectively.

ECCMA provides a platform for collaboration amongst subject experts on data quality and data governance around the world to build and maintain global, open standard dictionaries that are used to unambiguously label information. The existence of these dictionaries of labels allows information to be passed from one computer system to another without losing meaning, and thereby provides interoperability.

## 4.4 Data Interoperability

Data interoperability is another important factor in regard to data acquisition, sharing and efficient use. There are dozens if not even hundreds of standards regarding data interoperability in place worldwide by several standardisation bodies and in several industry domains.

This diversity of data interoperability standards is exactly the problem as there is only little mapping between such standards and approaches. Standards often are developed in research projects without the involvement of the end users and thereby might lack implementation experience.

The latest developments – also mentioned below – in the areas of data infrastructures, data spaces and datamarkets have shown an interesting movement towards and harmonisation of standards in data interoperability by IDSA and GAIA-X and beyond.

## 4.5 Licenses and Data related Regulations

Since the open data movement the paradigm has been defined as follows for data (on the web): provide every digital asset with a clear and dedicated license (Open Data principle #8 of the Sunlight Foundation).

Since then the importance of clearly defined licenses for data has become more and more important, and there are quite a lot of them, too many to provide a comprehensive list as part of this deep dive.

For Open Licenses a good source is the Open Definition of the Open Knowledge Foundation.[38] Another helpful source of information is the Open Data Support training programme, with its module: Training Module 2.5. Data & metadata licensing.[39]

Regulations and Directives have been developed by the European Commission over the last decade in regard to data and thereby build an important groundwork for the data economy and the realisation of a working and sustainable data infrastructure across Europe that as well as can support a digital European Language Equality.

The most important (not complete list) are:

[36] https://www.w3.org/TR/vocab-dqv/
[37] https://eccma.org
[38] https://opendefinition.org/guide/data/
[39] https://data.europa.eu/sites/default/files/d2.1.2_training_module_2.5_data_and_metadata_licensing_en_edp.pdf

- The EU General Data Protection Regulation (GDPR) went into effect on May 25, 2018[40]
- A European Strategy for Data[41]
- European data governance (Data Governance Act)[42]
- EU Open Data Strategy and PSI Directive[43]
- A European Approach to Artificial Intelligence, including the EC AI Strategy[44]
- Digital Single Market Strategy for Europe[45]
- Digital Action Education Plan[46]

Regarding Language Technology and a European Language Equality all of the listed regulations have a clear impact. For this deep dive document we point out, that the Data Governance Act has strong a implication for data, language resources and Knowledge Graphs, as the Digital Governance Act is laying the groundwork for the development of common data spaces in strategic sectors.

The bitkom Position Paper on the Data Governance Act, provides a related statement to be taken into account as follows: "While we believe that data intermediaries and data altruism can expand data sharing, we are also convinced that different frameworks, setups and governance structures need to be the future basis for these data spaces (vertical approach). By narrowing the scope, the DGA will avoid restricting other possible setups of data spaces. The DGA should also take initiatives such as GAIA-X / the upcoming European Alliance for Industrial Data and Cloud and the European Cloud Federation into account (e. g., with regard to the conditions set out for data sharing intermediaries who use/ offer underlying Cloud infrastructure) to actively build a coherent framework on the basis of existing initiatives – avoiding building parallel rules."[47] Data Spaces for Language Technology should be seen as such a strategic sector, as language technology is kind of a horizontal throughout many industries and domains.

## 4.6 Data and Ethics

The topic of data and ethics is – beside technical issues – a topic in which regulators and standards (see above) play a crucial role. After long years discussion about data and ethics but also about AI and ethics, a standard has been published in the field: *IEEE P7000 – Engineering Methodologies for Ethical Life-cycle Concerns Working Group* with the following scope: The standard 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 view of new IT product development, computer ethics and IT system design, value-sensitive design, and, stakeholder involvement in ethical IT system design.[48]

[40] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[41] https://ec.europa.eu/info/sites/default/files/communication-european-strategy-data-19feb2020_en.pdf
[42] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767
[43] https://digital-strategy.ec.europa.eu/en/policies/open-data
[44] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
[45] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52015DC0192
[46] https://ec.europa.eu/education/education-in-the-eu/digital-education-action-plan_en
[47] Position Paper, Bitkom Comments on the Data Governance Act, February 2021, https://www.bitkom.org/sites/default/files/2021-02/20210209_bitkom-position-data-governance-act.pdf
[48] https://sagroups.ieee.org/7000/

Prof. Sarah Spiekermann, who is a leading researcher in the field, stated in 2019 as follows:[49]

> AI is a highly technical matter and when it comes to technical standardization, ethics is a relatively new field. Technology standardization is traditionally dealing with protocols, with hardware specifications, etc. The fuzzy domain of ethics as well as the context-sensitivity of any ethical matter seems almost contrary to the straight and homogeneous logic of the engineering world.

Prof. Spiekermann has been the vice-chair of the brand new standard: IEEE 7000 Ethical System Engineering Standard, since its initiation and has philosophically grounded it in material value ethics.[50]

## 4.7 Data Literacy

In regard to data literacy there are big differences regarding digitalisation between the European Member States, as well as inside the EU member states, as well as between domains and industries.

Gartner defines data literacy as:[51]

> 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.

Further, 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,[52] and globally programmes like the World Bank's Data Use and Literacy Programme[53] supports the awareness, education and implementation of data literacy.

Nevertheless compared to the data and the data related technologies available the data literacy is far behind its development and needs more action and efforts.

## 4.8 Data infrastructures, data spaces, and data markets

Data Spaces are a relatively young concept and solution approach to stimulate the economy and business in Europe by the provision of secure and trustworthy mechanisms and platforms for data sharing and data trading. The European Commission lists 9 Data Spaces in its Data Strategy as of February 2020[54] that is strongly interconnected with the EC Data Governance Act[55]. EU Members States supported the research on data spaces in the last years, as for example the International Data Space project (Germany) that channeled into the establishment of IDSA (the International Data Space Association) and the publication of several standards and recommendations in the field (IDS Information Model or Reference Architecture Model[56]), or the Data Market Austria (DMA)[57] that developed a prototype for a

[49] https://www.sustainablecomputing.eu/blog/4103/sarah-spiekermann-who-looks-after-the-ethics-of-ai-on-the-role-of-the-regulators-and-standards/
[50] https://www.wu.ac.at/ec/projects/ieee-p7000-standard
[51] https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy
[52] https://ec.europa.eu/education/education-in-the-eu/digital-education-action-plan_en
[53] https://www.worldbank.org/en/programs/data-use-and-literacy-program
[54] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy
[55] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act
[56] https://internationaldataspaces.org/use/reference-architecture/
[57] https://datamarket.at

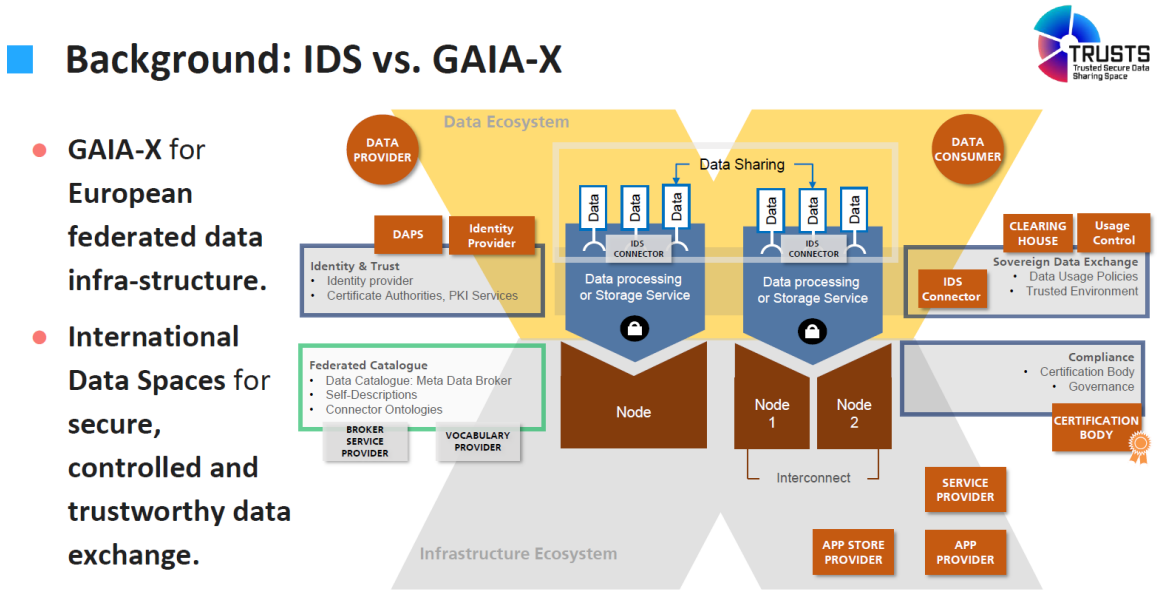


Figure 4: Background: IDS vs. GAIA-X. From Prof. Dr Jan Juerjens, Presentation: TRUSTS, EBDVF2021.

public marketplace for data trading. Several Research & Innovation projects in the field are in progress, for instance the TRUSTS project,[58] the ELG[59] and AI4EU project,[60] or the EOSC related project SSHOC[61] for research data.

One of the latest initiatives established in the field of data infrastructures is **GAIA-X** with the following objectives:[62]

> With Gaia-X, representatives from business, science and politics on an international level create a proposal for the next generation of data infrastructure: an open, transparent and secure digital ecosystem, where data and services can be made available, collated and shared in an environment of trust.
>
> Gaia-X is a project initiated by Europe for Europe and beyond. Representatives from business, politics, and science from Europe and around the globe are working together, hand in hand, to create a federated and secure data infrastructure. Companies and citizens will collate and share data – in such a way that they keep control over them. They should decide what happens to their data, where it is stored, and always retain data sovereignty.
>
> The architecture of Gaia-X is based on the principle of decentralisation. Gaia-X is the result of a multitude of individual platforms that all follow a common standard – the Gaia-X standard. Together, we are developing a data infrastructure based on the values of openness, transparency, and trust. So, what emerges is not a cloud, but a networked system that links many cloud services providers together.

[58] https://www.trusts-data.eu
[59] https://www.european-language-grid.eu
[60] https://www.ai4eu.eu
[61] https://sshopencloud.eu
[62] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

A continuously growing number of datamarkets and data spaces also in the field of language technology can be seen at the moment, with more to come following the funding schemes in Horizon Europe of the European Commission.

With publicly funded projects like the European Language Grid – ELG,[63] the EUDAT Collaborative Data Infrastructure (EUDAT CDI),[64] the full EOSC infrastructure[65] including datamarkets like the one for social sciences and humanities, the SSHOC Marketplace,[66] as well as its initiatives like CLARIN[67] that these days bring the established research infrastructures into the area of datamarkets and data spaces. But also first successful industrial products like the Taus Marketplace,[68] or the Lexicala data API[69] for lexicographical information for more than 50 languages, or the NLP API of Experts AI.[70] Furthermore there are well-known associations and organisations in place that provide language resource catalogues, like LDC – The Linguistic Data Consortium,[71] or ELDA – The Evaluations and Language resources Distribution Agency,[72] or ELRA – The European Language Resources Association.[73] Finally there are active industry associations and networks like LT-Innovate[74] or the BDVA/DAIRO[75] in place that support the idea of data collection and provision / sharing in their networks to support – beside other objectives – better Language Technology in the future.

Following the announcement of January 2021 by the Support Centre for Data Sharing the Digital Europe Programme includes a dedicated action for a Data Space for Language, that shall include and integrate several of the above listed existing resources.[76]

## 4.9 Knowledge Graphs

The idea of a knowledge graph is a relatively young concept as well, that comes from the semantic web community following the basic principles of semantic web and linked data. There are several definitions that all in all provide a good idea of the meaning of a Knowledge Graph, although being a little different from one to one.

For Language Technologies the Knowledge Graph principles provide a huge potential in regard to the possibilities in modeling of common-sense knowledge and domain specific knowledge, as well as the provision of rich context and meaning in mono-lingual, bi-lingual, multilingual and cross-lingual applications.

**Definition of a Knowledge Graph** A knowledge graph is a directed labeled graph in which the labels have well-defined meanings. A directed labeled graph consists of nodes, edges, and labels. Anything can act as a node, for example, people, company, computer, etc. An edge connects a pair of nodes and captures the relationship of interest between them, for example, friendship relationship between two people, customer relationship between a company and person, or a network connection between two computers. The labels capture the meaning of the relationship, for example, the friendship relationship between two people.[77]

[63] https://www.european-language-grid.eu
[64] https://www.eudat.eu
[65] https://eosc-portal.eu
[66] https://marketplace.sshopencloud.eu
[67] https://www.clarin.eu
[68] https://datamarketplace.taus.net
[69] https://lexicala.com
[70] https://try.expert.ai
[71] https://www.ldc.upenn.edu
[72] http://www.elra.info/en/about/elda/
[73] http://www.elra.info
[74] https://www.lt-innovate.org
[75] https://www.bdva.eu
[76] https://eudatasharing.eu/de/news/digital-europe-programme-explained-language-data-space
[77] https://web.stanford.edu/class/cs520/2020/notes/What_is_a_Knowledge_Graph.html


Figure 5: Graph with nodes A and C and a directed edge B from A to C.

More formally, given a set of nodes N, and a set of labels L, a knowledge graph is a subset of the cross product N × L × N. Each member of this set is referred to as a triple, and can be visualized as shown in Figure 5.

The directed graph representation is used in a variety of ways depending on the needs of an application. A directed graph such as the one in which the nodes are people, and the edges capture friendship relationship is also known as a data graph. A directed graph in which the nodes are classes of objects (e. g., Book, Textbook, etc.), and the edges capture the subclass relationship, is also known as a taxonomy. In some data models, A is referred to as *subject*, B is referred to as *predicate*, and C is referred to as *object*.

Many interesting computations over graphs can be reduced to navigation. For example, in a friendship knowledge graph, to calculate the friends of a friend of a person A, we can navigate the knowledge graph from A to all nodes B connected to it by a relation labeled as friend, and then recursively to all nodes C connected by the friend relation to each B.

A path in a graph G is a series of nodes $(v_1, v_2, \ldots, v_n)$ where for any $i \in N$ with $1 \square i < n$, there is an edge from $v_i$ to $v_{i+1}$. A *simple path* is a path with no repeated nodes. A *cycle* is a path in which the first and the last nodes are the same. Usually, we are interested in only those paths in which the edge label is the same for every pair of nodes. It is possible to define numerous additional properties over the graphs (e. g., connected components, strongly connected components), and provide different ways to traverse the graphs (e. g., shortest path, Hamiltonian path, etc.).

**Definition of a Knowledge Graph. Dataversity** A knowledge graph, which can be considered a type of ontology, depicts "knowledge in terms of entities and their relationships," according to GitHub. Knowledge graphs developed from the need to do something with or act upon information based on context. For example, knowledge graphs help identify fraud, keep track of inventories, and write novels. Knowledge graphs have been gaining more traction with machine learning so that AI processes can use the same information, as needed, for multiple situations. Knowledge graphs simplify complex concepts at one glance and promise good training data for AI to learn new tasks.

Key abilities of knowledge graphs, according to Ralph Hodgson, CTO of TopQuadrant, include:

**Extensibility:** The ability to accommodate diverse data and metadata that evolve over time.

**Introspection/Query Ability:** Models that can be inspected to find what things are knowable and findable.

**Semantic:** The meaning of the data is stored within the graph alongside the data to understand connections.

**Intelligence Enabling:** The ability to infer dependencies and other relationships between objects.

Other definitions of a Knowledge Graph includes:

- An interconnected set of information, able to meaningfully bridge enterprise data silos and provide a holistic view of the organization through relationships. (Amber Lee Dennis)
- Layers atop the existing data infrastructure that reveal the relationships within the data, regardless of the source or format. (Keith D. Foote)
- An enhanced graph database enriched with business rules that allow for inference to be performed upon the connected data. (Keith D. Foote)
- A means of storing and using data, which allows people and machines to better tap into the connections in their datasets. (Datanami)
- A database which stores information in a graphical format – and, importantly, can be used to generate a graphical representation of the relationships between any of its data points. (Forbes)
- Encyclopedias of the Semantic World. (Forbes)

Knowledge graphs are often assembled from numerous sources, and as a result, can be highly diverse in terms of structure and granularity. To address this diversity, representations of *schema, identity*, and *context* often play a key role, where a schema defines a high-level structure for the knowledge graph, *identity* denotes which nodes in the graph (or in external sources) refer to the same real-world entity, while *context* may indicate a specific setting in which some unit of knowledge is held true. As aforementioned, effective methods for *extraction, enrichment, quality assessment,* and *refinement* are required for a knowledge graph to grow and improve over time.

Knowledge graphs aim to serve as an ever-evolving shared substrate of knowledge within an organisation or community (Noy et al., 2001). We distinguish two types of knowledge graphs in practice: *open knowledge graphs* and *enterprise knowledge graphs*. Open knowledge graphs are published online, making their content accessible for the public good. The most prominent examples – DBpedia (Lehmann et al., 2015), Freebase (Bollacker et al., 2007), Wikidata (Vrandecic and Krötzsch, 2014), etc. – cover many domains and are either extracted from Wikipedia, or built by communities of volunteers. Open knowledge graphs have also been published within specific domains.

Enterprise knowledge graphs are typically internal to a company and applied for commercial use-cases. Prominent industries using enterprise knowledge graphs include Web search (e. g., Bing,[78] Google,[79]) commerce (e. g., Airbnb,[80] Amazon[81]), social networks (e. g., Facebook (Noy et al., 2019)), finance, among others. Applications based on Knowledge Graphs include search, recommendations, personal agents, advertising, business analytics, risk assessment, automation.

[78] https://blogs.bing.com/search-quality-insights/2017-07/bring-rich-knowledge-of-people-places-things-and-local-businesses-to-your-apps
[79] https://blog.google/products/search/introducing-knowledge-graph-things-not/
[80] https://medium.com/airbnb-engineering/scaling-knowledge-access-and-retrieval-at-airbnb-665b6ba21e95
[81] https://www.amazon.science/blog/building-product-graphs-automatically

Useful resources for further reading are for instance: Ji et al. (2021), Abu-Salih (2021), Li et al. (2021), or Colon-Hernandez et al. (2021). A nice resource for further reading about Knowledge Graphs, as well as for how to build (Enterprise) Knowledge Graphs is Blumauer and Nagy.

## 4.10 Semantic AI: Statistical and Symbolic AI in Combination

The field of Semantic Artificial Intelligence investigates the integration of symbolic representations and statistical inductive models usually based on machine learning. A closely related field with a longer history is neuro-symbolic systems (d'Avila Garcez and Lamb, 2020; d'Avila Garcez et al., 2002). A more narrow understanding of Semantic AI is the integration of machine learning with data from the semantic web represented using RDF or related formalisms (SHaCL, ShEx, OWL, etc.).

Recent surveys demonstrate a growing interest to Semantic AI (Sarker et al., 2021; d'Amato, 2020; d'Avila Garcez et al., 2002).

Modern Semantic AI systems typically include either a reasoner or a classical machine learning system. The reasoner only relies on structured formalized data, therefore it is more often used with expressive ontologies. The usage of machine learning allows efficient processing of unstructured data, therefore, such systems are successfully applied on tasks that include NLP or video/image processing. Recent advances in Deep Learning classifiers enabled a new wave of research in this field. In particular, the usage of symbolic knowledge in combination with DL models set state of the art in such NLP tasks as NER and relation extraction (Wang et al., 2020), aid in image processing (Zhang et al., 2019). Embedding techniques and models are often used to solve purely graph tasks, for example, link prediction or new node classification.

These approaches are widely used in different domains including natural sciences, especially biology, culture, economics, news & social media analysis.Typical tasks include text analysis and annotations, entity linking and disambiguation, information extraction and retrieval, image classification, object recognition, link prediction, knowledge graph extension. Over the years these systems get more complex both in the interaction patterns between different subsystems as well as the flow of data.

## 4.11 Innovative Data and Metadata Management Tools

Transitioning from big data to small and wide data is one of the Gartner top data and analytics trends for 2021. These trends represent business, market and technology dynamics that data and analytics leaders cannot afford to ignore.

> These data and analytics trends can help organizations and society deal with disruptive change, radical uncertainty and the opportunities they bring over the next three years. Data and analytics leaders must proactively examine how to leverage these trends into mission-critical investments that accelerate their capabilities to anticipate, shift and respond.
>
> Rita Sallam, Distinguished VP Analyst, Gartner.

Regarding the current state of the art we follow Gartner's Top 10 Data and Analytics Trends for 2021, that are as follows.[82] Each of the trends fit under one of these three main themes:

1. Accelerating change in data and analytics: Leveraging innovations in AI, improved composability, and more agile and efficient integration of more diverse data sources.

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

2. Operationalizing business value through more effective XOps: Enables better decision making and turning data and analytics into an integral part of business.
3. Distributed everything: Requires the flexible relating of data and insights to empower an even wider audience of people and objects.

Trends:

1. Smarter, more responsible, scalable AI
2. Composable data and analytics
3. Data fabric as the foundation
4. From big to small and wide data
5. XOps
6. Engineered decision intelligence
7. Data and analytics as a core business function
8. Graph relates everything
9. The rise of the augmented consumer
10. Data and analytics at the edge

Combined with Gartner Hype Cycle for Natural Language Technologies (see Figure 6), these trends provide a good state of the art analysis of innovative metadata and data management tools.[83] It shows that AI is everywhere (Trend #1) and becomes more responsible and explainable, but also scalable which supports the topic of Deep Natural Language Understanding. It mentions the data fabric as a foundation (Trend #3), that we have identified in this deep dive as a future Technology Trend if it works in its full potential as a Semantic Data Fabric (see section below in this document), and its states the requirements of small and wide data (Trend #4), as well as the need of graphs (thereby Knowledge Graphs, Trend #8) and finally the data and analytics at the edge (Trend #10) that supports conversational AI solutions and Insight Engines

Metadata and data management tools often are classified along the concept of a data lifecycle, that refers to the entire period of time that data exists in your system. This life cycle encompasses all the stages that your data goes through, from first capture onward.

The steps and stages of a data life cycle usually looks as follows, although there are plenty of data life cycle approaches, differing in its complexity and often adapted by specific domains or industries[84]:

**Data creation, ingestion, or capture** Whether you generate data from data entry, acquire existing data from other sources, or receive signals from devices, you get information somehow. This stage describes when data values enter the firewalls of your system.

**Data processing** There are many processes involved in cleaning and preparing raw data for later analysis. While the order of operations may vary, data preparation typically includes integrating data from multiple sources, validating data, and applying the transformation. Data is often reformatted, summarized, subset, standardized, and enriched as part of the data processing workflow.

[83] https://www.uniphore.com/research/2021-gartner-hype-cycle-for-natural-language-technologies/
[84] https://www.talend.com/resources/data-lifecycle-management/

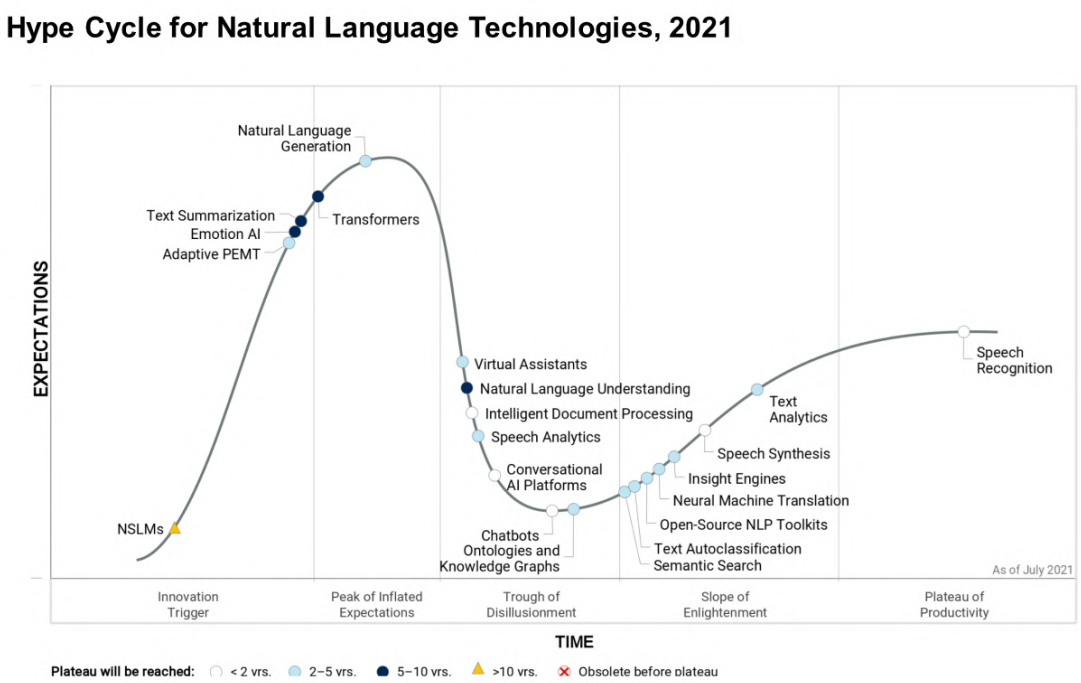


Figure 6: Gartner Hype Cycle for Natural Language Technologies

**Data analysis** However you analyze and interpret your data, this is where the magic happens. Exploring and interpreting your data may require a variety of analyses. This could mean statistical analysis and visualization. It can also mean using traditional data modeling or applying artificial intelligence (AI).

**Data sharing or publication** This stage is where forecasts and insights turn into decisions and direction. When you disseminate the information gained from data analysis, your data delivers its full business value.

**Archiving** Once data has been collected, processed, analyzed, and shared, it is typically stored for future reference. For archives to have any future value, it's important to keep metadata about each item in your records, particularly about data provenance.

The data life cycle proceeds from the last step back to the first in a never-ending circle. Of course, in the twenty-first century, one factor has seriously complicated the way we work with data.

# 5 Data, Language Resources, Knowledge Graphs: Main Gaps

This section provides a gap analysis in regard to the identified components for this deep dive in relation to data, as well as in regard to the related and identified technology concepts, methodologies and tools. The structure for this section follows the identified components, related technology concepts, methodologies and tools of section 3. Main Components of this document.

## 5.1 Main Gaps: Components

**Availability of data and metadata** The following statements have been identified, collected and specified together with researchers and practitioners in the field and reflect the current gaps in the components:

Legacy data: there is a fully untapped potential of relevant data available in archives as well as old data files, that should be made accessible and thereby available. This gap is also related to data accessibility as well as to data management tools, and thereby also mentioned there.

Open AI models: there is a strong need for open AI models in language technology that are provided to the interested parties with an open license, and thereby can be easily and freely be re-used.

Raw data to build own models: not only ready-to-use (open AI) models are required but also the raw data used to develop and calculate / train such models, to allow the creation of own models. And thereby the ability to avoid bias (by re-using existing models only) and also to enabling responsible AI[85] and explainable AI[86] by developing and publishing a proper documentation of the data used as well as the algorithms used in such models.

Multi model data: data often is available in a single model as well as in a single serialisation. The provision of multi model data as well as data in different serialisations and formats can foster use and re-use and the faster development of useful applications and solutions in the field.

Hand crafted models (also along business cases and models) versus machine learning built models: similar to the identified gap above of the need of the raw data to build AI and other

[85] https://ec.europa.eu/jrc/communities/sites/default/files/03_dignum_v.pdf
[86] https://www.darpa.mil/program/explainable-artificial-intelligence

language technology related models this gap shows that there is a strong requirement for hand crafted models along a specific use case and/or business model, in addition to purely machine learning built models. As there is a clear lack of the availability of both directions (handcrafted and machine learning built) the hand-crafted models are even a rare species of needed models as being very specific but as of this also of a much higher value for such a specific use case and/or business case.

Lack of annotated corpora in all languages: annotated corpora are often available in English language only (if they are) but very rarely available in other / all languages. This means in EU 24 languages, in dialects of EU languages, in sign languages, as well as in Non-EU languages that are intensively used inside Europe (e.g Arabic, or Turkish to provide examples).

Lack of domain specific, multilingual corpora: in addition to the gap specified above there is also a strong need in domain specific corpora, mono-lingual, bi-lingual and multilingual. Domain specific annotated corpora allow the development of domain specific solutions, as domains (as for instance health, medicine, pharma or finance) use a very specific language that is also developing quickly over time as of the specification of new terminology.

Too little domain specific data available: beside domain specific annotated corpora, there is an overall need for more domain specific data available to use. For example the fast development of terminology in the course of the covid-19 pandemic worldwide, a specific "language" around the topic has developed, that also differs from language to language as of cultural and scientific differences and diversity.

Not enough volume of data for certain areas and use cases, not enough application related data: very much related to domain specific data but also going beyond is the lack of volume data for certain use cases, where the complexity of this gap lies in the clear specification of a use case and thereby the clear specification of the required data for such a use case. Proper documentation in the related metadata, classification schemas as well as Knowledge Graphs can support to overcome this issue.

Weak findability of data via available resources, also because of missing proper documentation of data and provision of sufficient metadata: the current data repositories – that are rarely connected to each other and often make use of different metadata schemas – provide weak search- and recommendation mechanisms often because of the weak underlying documentation by means of rich metadata available.

Manually annotated data is missing: although the quality of automated and semi-automated annotations is increasing, the manual annotation by humans (by experts in a certain field ) is still the best quality to receive in regard to annotation. But there is a clear gap in the availability of manually annotated data as of the workload and literacy for such an activity. Nevertheless such manually annotated data in each language are crucial to be used for benchmarks or as gold standards to fine grain and improve automated annotations in languages.

Overall missing open language resources: even basic dictionaries, but also specific language resources are required to be available for scientific purposes with an open license. Currently – if available anyway – such data is available only behind a paywall.

Availability of data and language resources for "smaller / more languages": language (technology) related data is available often for English language only or in Europe for the 5 major languages only (English, German, French, Spanish and Italian), but to enable language equality and working language technology solutions in all languages data needs to be available in at least the EU 24 languages, but for real language equality even beyond these 24 languages.

**Accessibility of data** The following statements have been identified, collected and specified together with researchers and practitioners in the field and reflect the current gaps in the components:

The promotion of the importance of a large/ high quality Wikipedia-alike resources for a

language: such resources are crucial for the development of powerful language technology solutions, and as such resources are available under open licenses the access is provided. Thereby the continuous Wikipedia development (or the development of similar resources, but the advantage of Wikipedia is the respective already available infrastructure) and expansion should be funded by governments as for example per the approach taken by the Welsh[87]). The usefulness of such availability and the accessibility of such data can be easily seen by the experiments carried out and papers published around BERT, M-BERT and RoBERTa.

Legacy data: same gap as listed in Availability above, there is a fully untapped potential of relevant data available in archives as well as old data files, that should be made accessible..

Implemented FAIR principles: the overall idea of FAIR principles are a clear advantage regarding the use of data and metadata (and include the principle of accessibility) but the FAIR principles are not really rolled out properly and in a broad way in the scientific community and in industry the picture of FAIR implementation is even worse. Thereby FAIR should be more promoted in all communities and domains in relation to proper data and metadata management.

Too little open access language resources: required language data is often provided behind a paywall and if provided in the form of open access then for major languages only. This clearly hinders the idea of a digital language equality in Europe and thereby needs to be addressed.

Unclear Licenses: although Europe follows a strong open data movement for approximately 15 years, there is still a gap in the provision of clearly specified licenses for data and information available. Thereby huge amounts of data available are simply not accessible as of the legal uncertainty in its use.

**Quality of data** The following statements have been identified, collected and specified together with researchers and practitioners in the field and reflect the current gaps in the components:

Better benchmarking approaches: at the moment benchmark approaches are not harmonised or standardised. But this is a strong requirement for a thorough discussion on the issues of data-sourcing and benchmarking (Parra Escartín et al., 2021).

Benchmarks on domain specific vocabularies and on annotated data and corpora: are often missing but are heavily required to provide high quality language resources and thereby working solutions for language related applications and beyond.

Metadata provides only very limited data lineage / provenance: although there is more and more metadata available in several repositories, there is a lack in the metadata in regard to the provision of data lineage information, as well as data provenance (provenance: the source / origin of data and data lineage: the full history of data, thereby also tracking any change in a dataset, see also Wikipedia[88]) that allows us to evaluate the origin of data and the changes in data and thereby take such attributes into account for making valuable use of data.

Overall data quality is weak: and thereby often use cases cannot be realised as specified. Labeled data is not available.[89]

> Labeled data is a group of samples that have been tagged with one or more labels. Labeling typically takes a set of unlabeled data and augments each piece of it with informative tags. For example, a data label might indicate whether a photo

[87] https://www.bbc.com/news/uk-wales-north-west-wales-23402054 and https://www-2018.swansea.ac.uk/press-office/news-archive/2015/firsteverwikipediaedit-a-thoninwalestakesplaceatswanseauniversity.php

[88] https://en.wikipedia.org/wiki/Data_lineage

[89] https://en.wikipedia.org/wiki/Labeled_data

> contains a horse or a cow, which words were uttered in an audio recording, what type of action is being performed in a video, what the topic of a news article is, what the overall sentiment of a tweet is, or whether a dot in an X-ray is a tumor. Labels can be obtained by asking humans to make judgments about a given piece of unlabeled data. Labeled data is significantly more expensive to obtain than the raw unlabeled data.

Data for all languages in required quality: there is too little quality data for dialects in place as well as for non-EU languages, same for older languages and for sign languages. Example German language: German vs Austrian but also dialects in Germany of German language (for instance: speech recognition in German does not work properly when tested by an Austrian or a German native speaker).

Weak data governance and thereby weak data quality: non existing policies around data and metadata management that should be part of a data governance often result in low data quality.

**Data Interoperability** The following statements have been identified, collected and specified together with researchers and practitioners in the field and reflect the current gaps in the components:

Data Silos: there are many, as well as more and more data silos in place that are not connected and not interoperable, as of the use of different metadata schemas, technologies, data models etc. Thereby data acquired from different sources need to be tested, evaluated, improved in quality and finally integrated with a huge effort, that comes on top of the costs for the data itself, and this makes working solutions based on such data costly.

Data infrastructures: similar to the gap described above regarding data silos: there are more and more data infrastructures available that are simply not interoperable and the data inside such infrastructures does not provide proper data interoperability.

Standards: the harmonisation of relevant standards in the field is missing. As an example: when 2 or more relevant standards are combined, one develops a new standard. Sometimes the processes are too strict and too slow adaptation of standards. The problem often is that adaptation of standards is complex and takes a long time, thereby more flexibility in standards adaptation and harmonisation is required to overcome this gap.

Cross community data and interoperability: there is a clear problem and gap in the combination of scientific data (e. g., via European Open Science Cloud, EOSC[90] and industry data (e. g., via an industry data market) as well as data from public administration / government data catalogs and portals (e. g., the European Open Data Portal[91]). This is caused by the development and availability of more and more data islands (see the gap already identified above in the section: Availability) in different communities and domains that are using different standards and technologies and thereby different means of data interoperability.

*Remark*: in Europe often the following problem is stressed, that we are lacking sufficient amounts of data for powerful AI, language technology and other applications and that e. g. North America or Asia have a clear competitive advantage thereby in this field. But taking into account the above described issue of the heterogeneity of the available data repositories in place this seems not an issue of volume but an issue of variety and thereby of data and metadata interoperability.

**Licenses and data related regulations** The following statements have been identified, collected and specified together with researchers and practitioners in the field and reflect the current gaps in the components:

[90] https://eosc-portal.eu
[91] https://data.europa.eu

Open Access, open licenses: such licenses are often not provided for data available and there is a lack of the knowledge about open licenses, that could be promoted like via the Open Definition of the OKFN.[92]

Gaps in available data related directives: for example, the EU Open Data Directive makes it easier to collect public data (seen through the ELRC[93]). Thereby more directives and regulations in the field should be developed to solve and overcome the gaps identified in this deep dive document in relation to data, language resources, and Knowledge Graphs.

Effect of regulations and directives: the effect of regulations on data related topics should be evaluated continuously and regulations and directives adapted along identified gaps and changed environments. For example the General Data Protection Regulation (GDPR[94]) has a strong effect on data collection.

Lack of machine-readable license information: as stated above there is an overall lack in the assignment of proper license information to data, but in addition there is a strong need and an existing gap in the provision of machine-readable license information that can be harvested and analysed easily in the course of a data collection/acquisition process.

Personal Identifiable Information (PII) and anonymisation: there is often too little anonymisation in place for available data what is clearly creating bias.This means there is a strong need for powerful anonymisation in some areas (e. g., justice, health) to (i) follow the respective regulations and (ii) avoid bias via the data.

Guidelines and policies: are not available for each (!) language in place to achieve digital equality. This means such guidelines and policies should be developed and implemented by regional-national-international public administrations. The same approach should apply for new languages appearing.

**Data and ethics** The following statements have been identified, collected and specified together with researchers and practitioners in the field and reflect the current gaps in the components:

Data for Non-EU languages and beyond: are not in place sufficiently – neither as open access nor behind a paywall – and thereby services for such languages cannot be developed properly, meaning in a proper quality to be used. Same issue for dialects and sign-languages. This means also that bias is created by wrong language data, and thereby ethical issues come into play.

**Data literacy** The following statements have been identified, collected and specified together with researchers and practitioners in the field and reflect the current gaps in the components:

Crowdsourcing gaps: national crowd-sourcing platforms that facilitate data collection for low-resource languages are not available but would support the digital language equality in Europe (for example, it is not possible to get crowd workers for Irish on Mechanical Turk).

Citizen Science: on that same note, making citizen science more widely known and understood (e. g., through schools etc.) would help to involve more citizens that are interested in the support of e g local languages / dialects, e. g., for transcription, see as an example: https://www.duchas.ie/en/meitheal/.

Data management and data literacy: there is a strong need for education and understanding for better data management processes in science, academia, as well as in business and industry, also to better understand the value of data, and thereby improve data management principles, and techniques.

[92] https://opendefinition.org/guide/
[93] https://lr-coordination.eu
[94] https://gdpr.eu

Educational Bodies: there is a need to informing educational bodies on the importance of sharing data as for example learner corpora, to for example: improve CALL or adaptive learning systems.

Senior staff and experts in AI: are clearly missing to work in the field regarding data related topics as well as applications. The same situation is in place for artificial intelligence (AI) and for deep learning mechanisms and beyond.

Diversity in Digitization: as an overall gap there is a strong difference with regard to the level of digitization in Europe. On the one hand there is a high level of data literacy available and on the other hand there is simply no data literacy or even awareness in place.

Gartner defines poor data literacy as a clear risk and gap[95] as follows:

> Poor data literacy is ranked as the second-biggest internal roadblock to the success of the CDO's office, according to the Gartner Annual Chief Data Officer Survey. By 2023, data literacy will become essential in driving business value, demonstrated by its formal inclusion in over 80% of data and analytics strategies and change management programs.

## 5.2 Main Gaps: Data Infrastructures, Data Spaces and Datamarkets

The following gaps have been identified and specified:

The topic of data infrastructures, data spaces and data markets has been promoted and pushed by national government as well as by the European Commission in the last years along the European Strategy for Data[96]. But the idea of data catalogues in the form of open (metadata) repositories has already been in place and promoted worldwide by the open data movement for around 15years now.

Hundreds of open data catalogues have been established worldwide and the DataPortals.org website developed and maintained by the Open Knowledge Foundation, is a good entry point to get an overview of available catalogues, listing 592 data catalogues and portals worldwide at the moment.[97]

But those 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, and only a little number of them provide also the data itself.

The resulting issues and gaps are regarding: (i) availability of and access to the data itself, as such catalogues are not aware if such data is still provided by the publishers and owners, (ii) lack of interoperability in metadata but mainly in data. Although the metadata is often providing data interoperability (as of using the same catalogue software CKAN,[98] and at least in Europe but also beyond are making use of the de-facto metadata standard for open data and data portals in Europe: DCAT-AP (Data Catalogue Vocabulary (DCAT) expanded for Application Profiles)[99] and (iii) a fragmentation of data catalogues and data portals.

Regarding data spaces and datamarkets the TRUSTS project (Trusted Secure Data Sharing Space[100]) has carried out an interesting study[101] on: 'Definition and analysis of the EU and worldwide data market trends and industrial needs for growth' that includes a section: Datamarket Challenges that provides a good summary of the gaps and challenges in this area – as follows a short summary:

[95] https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy
[96] https://digital-strategy.ec.europa.eu/en/policies/strategy-data
[97] https://datacatalogs.org
[98] https://ckan.org
[99] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe
[100] https://www.trusts-data.eu
[101] https://www.trusts-data.eu/wp-content/uploads/2021/07/D2.1-Definition-and-analysis-of-the-EU-and-worldwide-data-market-trends-....pdf

The challenges regarding datamarkets were categorised using the STOF model (see Table 1) which is a framework that provides the logic of business and its ecosystem (Bouwman et al., 2008). The STOF model consists of the service domain (S), technology domain (T), organization domain (O), and finance domain (F).

| Category | Challenge | Short description | Source | Perspective |
|---|---|---|---|---|
| Service | Data ownership definition | Who should own the data? The idea of data ownership is still debatable. Some scholars suggest that data (especially personal data) should be owned by individuals, while others argue it is not feasible or conceptually flawed. | (Koutroumpis et al., 2020) | Data providers |
| | Ensuring data integrity | Ensure data integrity, i. e., data is not altered during the lifecycle of data trading processes. | (Lawrenz et al., 2019) | Data providers, data buyers |
| | Assessing data quality | Data buyers do not know how to assess the quality of the data or evaluate the data before purchasing it. | (Koutroumpis et al., 2020) | Data buyers |
| | Ensuring contractual compli-ances | Data buyers may violate data access and usage restrictions. | (Koutroumpis et al., 2020) | Data providers |
| | Loss of control over data | Data providers are unable to track down data usage and ensure compliances towards data sharing agreements. In consequence, they are afraid that competitors may benefit from their data in unanticipated ways. It also brings potential privacy risks. | (Spiekermann, 2019) | Data providers |
| | Lack of trans-parency | Little transparency between data providers and data brokers. In some cases, data providers do know how data brokers assess their data value and how the assessment occurred in fair processes. | (Oh et al., 2019) | Data providers |

| | | | | |
|---|---|---|---|---|
| Technology | Privacy protection | Personal data trade can lead to unintended disclosure of personal information, especially when data providers are individuals and data buyers are huge corporations or governments, causing an imbalance of power. | (Virkar et al., 2019) | Data providers |
| | Security | In general, data marketplaces must provide security technologies to protect data trading processes from hacking, counterfeiting, and other unwanted behaviours. | (Virkar et al., 2019; Lawrenz et al., 2019) | Data providers, data buyers |
| | Technical efficiencies and scalabilities | Technical efficiencies and scalabilities, especially in data marketplaces employing a distributed-ledger technology or decentralized architecture, are known as a general problem. In general, data marketplaces consume high computation and communication cost. | (Liu et al., 2019; Ishmaev, 2020) | Data providers, data buyers |
| | Data placement cost | Data placement and replication cost (i. e., after the purchase) is considerably high and consumes both bandwidth and latency. | (Ren et al., 2018) | Data providers, data buyers |
| | User-friendly applica-tions and interfaces | User-friendly applications and interfaces are required for advancing data marketplaces. | (Ramachandran et al., 2018) | Data providers, data buyers |
| Organization | The absences of legal frame-works | No IPR (e. g., digital right management models) is attached to data. In addition, no clear liability rules can be asserted to contracts or violations thereof. | (Sørlie and Altmann, 2019; Spiekermann, 2019) | Data providers |

| | | | | |
|---|---|---|---|---|
| | Lack of resources and technical knowledge | Lack of resources (e. g., operating cost) and technical knowledge to manage technical complexities. | (Oliveira et al., 2019) | Data marketplace owners and operators |
| | Unclear organizational structure | The absence of well-defined models regarding actor definitions, their roles, and interactions between actors. | (Oliveira et al., 2019) | Data marketplace owners and operators |
| | Ethical concern | Giving monetary incentives to individuals to share their sensitive personal data e. g., health data raises ethical concerns such as undue influence. | (Ishmaev, 2020) | Data providers |
| Finance | Pricing mechanism | Data providers have no clear and standardized mechanisms to price data assets. | (Niu et al., 2020) | Data providers |
| | Data valuation | Data providers and data sellers do not recognize the value of data because of the limitations in calculating potential benefits/revenues. | (Spiekermann, 2019) | Data buyers, data sellers |
| | Profit maximization | Data providers struggle to find a strategy to optimize profit by finding the balance between revenue maximization strategy and cost structure for data acquisition. | (Zhang et al., 2019) | Data providers |

Table 1: Challenges of data marketplaces

In the *service* category, the majority of challenges have been discussed adequately. Scholars attempt to find solutions to *ensure contractual completions, retain control over data, and provide transaction transparency* via technological enforcements. For instance, initiatives to implement blockchain, smart contracts, and access controls have flourished. The attempt to discuss *data quality and data protection* have also been conducted. So far, however, there has been little discussion about defining *data ownership* in data marketplace research.

In the *technology* domain, some challenges such as *privacy protection and security technology* (e. g., cryptography) have been major themes in data marketplace research. Moreover, efficient *data placement* via cloud computing or digital storage has attracted much interest. What we still lack in the literature is the discussion that focuses on technical efficiency and scalability. Previous studies also have not dealt with user-friendly aspects of data marketplace applications and interfaces sufficiently.

The keyword and term analysis revealed only a few studies attempting to solve the organization domain's challenges, such as *the absence of legal frameworks*, *lack of resources and*

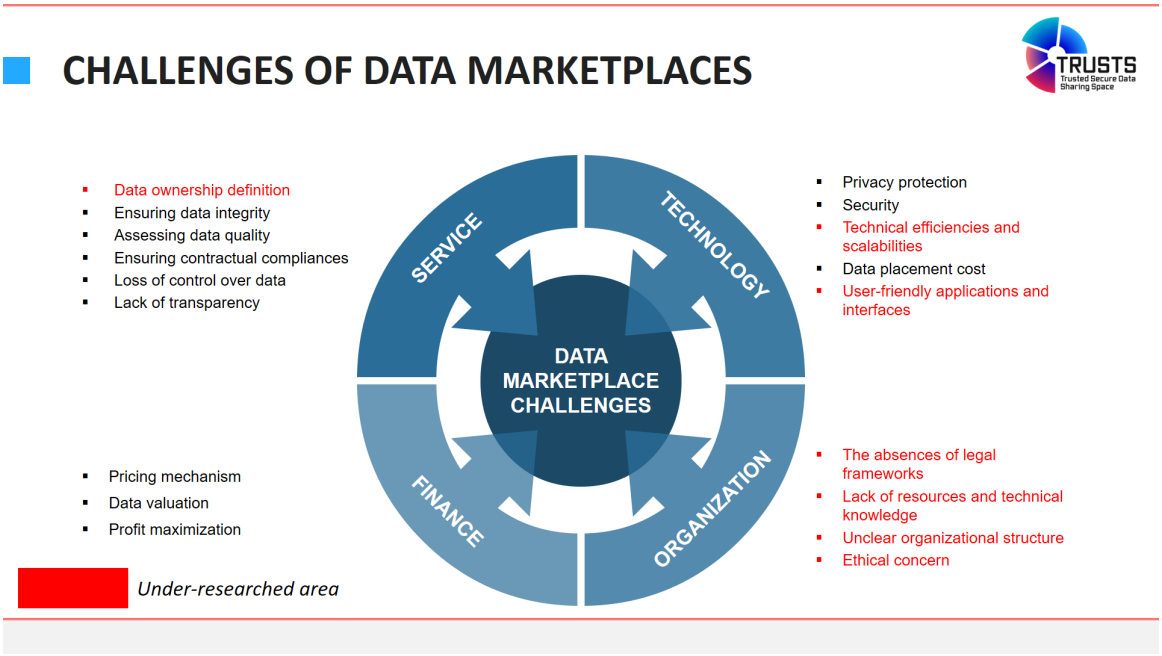


Figure 7: Challenges of data marketplaces

*technical knowledge, unclear organization, and ethical concern.* Further research in this area will be done to foster the development of data marketplaces.

Lastly, the challenges in the *financial* cluster have gained a lot of attention. The financial aspect of data marketplaces has always been a major topic in data marketplaces. Scholars use advanced technology (e. g., machine learning, query processing) and mathematical concepts (e. g., polynomial approximation) to determine price or budget for data.

Figure 7 summarizes the above-mentioned challenges of data marketplaces.

**In addition the following gaps have been identified** *More corpus query and analysis platforms are required*: like GREW Match for the UD treebanks[102] and others that could become part of services provision of data spaces and data markets.

*Secure data sharing*: it is still too risky – e g for industry players – to share data with others as of the risk that data is available publicly afterwards. At the moment there are no clearly working mechanisms to avoid this, like watermarking, or secure data sharing mechanisms.

*Missing computing infrastructure*: beside the data itself there is a lack of availability of affordable computing infrastructure that allows to run experiments, do model creation, or execute other NLP tasks that need high computing power like for instance making use of the BERT language model (BERT, Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). There are offerings in place by cloud providers like Amazon Web Services or Microsoft Azure, but the models of use are often costly and complicated.

*Lack of energy efficiency mechanisms in computing and data management*: another topic raised and becoming more and more important as of the increasing costs for energy as well as the climate change developments is the lack of energy efficient data management by making use of e g compression, slice and dice and filtering of data et al. This gap is also related to mentioned issues in the areas of (i) volume of available data and (ii) the creation of data models by using raw data (as such model creation consumes a lot of energy), and (iii) related

[102] http://match.grew.fr

data management approaches and tools. Data infrastructures, data spaces and datamarkets do not provide such mechanisms out of the box at the moment.

*Working business models*: finally all these gaps and issues in the field can only be solved and tackled with working business models in the field of data sharing and trading in a working and successful data economy. The International Data Space Association (IDSA) has published a relevant report in May 2021.[103]

## 5.3 Main Gaps: Knowledge Graphs

The following gaps have been identified and specified:

**Schema induction** For a knowledge graph to become useful for down-stream application there is a need to contain a certain amount of application and domain-specific knowledge. Often openly available resources are not suitable for a task or a domain of the task. Therefore, in order to reduce the entry barrier there is a need to be able to generate a suitable ontology/schema for the task and then to populate the schema with instances.

**Graph/ontology/taxonomy alignment** With the popularity of expert-based and/or crowdsourced knowledge graphs there exists a number of well curated external knowledge resources. However, in practice it is observed that different resources are rarely combined in a single solution. This fact contradicts the original ideas of semantic web and, in particular, linked data. We assume that the full power of knowledge based structures could be leveraged if the information about entities from one resource could be enriched with different types of information about the same entity from other resources. We deem that the difficulty consists in aligning the knowledge from different resources.

**Mechanisms to generate Knowledge Graphs in a more automated way and with high quality** Knowledge Graphs are a powerful idea to provide an interlinked, rich, high quality network of data providing context and meaning and thereby a great technology and methodology being a part of a working data infrastructure. But the work of knowledge graph creation is still costly as it is a combination of (i) manual work, as for instance: ontology modeling, and (ii) automated work as for instance automated data linking and machine learning mechanisms. The lack of more automation is still a criteria for the decision for or against the Knowledge Graph technology and thereby needs to be improved.

**Deep text analytics and smart text analytics** As Knowledge Graphs analyse metadata, as well as structured, semi-structured and unstructured data, there is a clear need for powerful deep text analytics and smart text analytics mechanisms and technologies for all languages to ensure multilingual comprehensive knowledge graphs. Smart text analytics takes into account rules and constraints in text analytics and text extractions to extract with more precision but also to avoid bias on extraction level, where possible.

**Cross-format analysis for graph enrichment** A Knowledge Graph at the moment is mainly built based on textual and numerical data as an input format (beside models). Thereby at the moment other formats like video, and audio is rarely taken into account. With a working (natural language technology) in the required languages mechanisms like speech to text could support the creation of Knowledge Graphs.

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

**Link prediction, entity resolution and constraints & rules modeling** Furthermore there is a gap in the available mechanisms and technologies regarding link prediction and entity resolution, where link prediction and entity resolution are two ways to identify missing information in networks. Link prediction helps identify edges that are likely to appear in the future, if they do not exist already. Entity resolution uses attributes and network structure data to link nodes that represent the same individual (Golbeck, 2013). In addition constraints and rules modeling are important areas that need to develop towards its full potential, for example in the field of SHACL: Shapes Constraints Language,[104] that is a W3C recommendation since 20th of July 2017, still has not unfolded its full potential nor is broadly used by the communities.

**Availability of comprehensive multilingual Knowledge Graphs** Finally there is a gap in the availability of comprehensive Knowledge Graphs at all. While there are some common knowledge Knowledge Graphs even freely available (DBpedia, or Yago to provide examples), there is a clear lack of bigger Knowledge Graphs in specific domains and industries, that can act as kind of foundation models but also as training input for AI algorithms and model. If such specific graphs are available there is a clear gap in availability of multilingual domain-specific Knowledge Graphs that can be used for Language Technology applications.

## 5.4 Main Gaps: Semantic AI

The following gaps have been identified and specified:

Regarding the gaps in the field of Semantic AI we see that the two relevant fields in AI: statistical and symbolic are still not fully combined and often the two fields exist beside each other and thereby cannot provide its full potential. This is the case in the overall machine learning community or the semantic web community but it is also the case in more specific technology domains like language technology or industry domains like health or energy.

Alan Morrison of PWC explained this issue of current AI (in comparison: what it needs and what it has) in detail at his keynote talk in 2018 at SEMANTiCS conference in Vienna titled: “Collapsing the IT Stack and Clearing a Path for AI Adoption”,[105] and taking place in 09/2018 in Vienna, Austria as follows:

**What it needs:** Contextualized, disambiguated, highly relevant and specific integrated data, flowing to the point of need

**What it has:** Single batch datasets cleaned up to be good enough by data scientists, who spend 80% of their time on cleanup

**What it needs:** Knowledge engineers, and many bold Data Visionaries in addition to big D Data Scientists, data-centric architects, pipeline engineers, specialists in many new data niches

**What it has:** A growing group of tool users versed only in probability theory, neural networks, python and R, including small D data scientists, engineers and architects, plus scads of entrenched application-centric developers.

And also Gartner Research as well as the Harvard Business Review stated,[106] already some time ago similar problems of current AI approaches as follows, that also fit very well with the gap analysis provided for other components of this deep dive:

[104] https://www.w3.org/TR/shacl/

[105] https://2018.semantics.cc/collapsing-it-stack-and-clearing-path-ai-adoption

[106] https://www.gartner.com/en/documents/3872125/cios-can-manage-the-risks-of-ai-investments https://www.gartner.com/en/documents/3868483/clarify-strategy-and-tactics-for-artificial-intelligence0, https://hbr.org/2016/12/breaking-down-data-silos

**Lack of AI Governance** Companies / organisations have concerns about validity, explainability and unintended bias of AI.

**Lack of AI Strategy** Many organisations are currently undertaking POCs from a large pool of AI vendors only for tactical benefits.

**Low Data Quality & Data Silos** 80% of the work of data scientists is acquiring and preparing data. A demon that can drive up that 80% and often makes initiatives impossible are data silos.

**Danger of Vendor Lock-in** Use of black-boxes instead of hybrid middleware approaches connecting internal training assets to third-party machine-learning solutions.

**Lack of Knowledge** Only 1 in 10 enterprises feel they have a competent approach to mining data, which ultimately hampers AI efforts. A shortage of AI skills and risk managers' lack of familiarity with the technology increase the risk.

*Knowledge-based Transfer Learning* Currently there is a growing interest in transfer learning, i. e., reusing the model trained on one task/domain on a different task/domain. Special interest for Semantic AI is zero-learning systems that do not require additional examples for transferring their learnings. Large knowledge graphs with a large number of facts and definitions could enable more efficient transfer learning of ML and especially DL models.

*Integration of general intelligence into ML* Often ML systems learning purely from data miss to capture some general intelligence that is available to most people. Therefore, often results might be inconsistent with some very general concepts. Current systems that aim at tackling this problem often use constraints induced from ontology to filter the solutions produced by ML. An integration of such constraints as soft constraints into ML would enable the next generation of models..

## 5.5 Main Gaps: Innovative Data and Metadata Management Tools

This area is strongly interconnected with the above listed gaps and issues in the fields of data infrastructures, data spaces and datamarkets, as well as with the Knowledge Graph topic, in addition, the following gaps have been identified and specified:

*The need for user-friendly flexible / open-source corpus annotation tools:* that can be used by linguistic experts as well as domain experts in-house easily and with fair costs and conditions.

*The need for user-friendly visualisation tools:* in order to be able to understand the content of the datasets at hand quickly and properly without the need of high efforts in data integration and data wrangling.

*Better detection techniques for harmful content:* are required to avoid bias, and identify and filter toxic content, or fake news and fake data, etc. This is not an easy task but in a time where artificial intelligence and machine learning is more and more used also small portions of toxic data and content can already influence a trained model or algorithm and thereby needs to be identified and filtered out.

*Better techniques for corpus filtering:* are required regarding domain filtering, noise cleaning (see also above) as well as the filtering and removal of bias.

*A clear lack of preservation technologies and tools:* have been identified that are required to ensure that small languages as well as old languages can be archived for 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.

*Intelligent data analytics of small content nuggets:* at the moment often only huge corpora are being analysed and even can be analysed by the available technologies and tools, but

the trend goes more and more towards a smart data analytics that can be applied on small datasets as well as on small content nuggets, for instance to one paragraph or section in a legal text only and not to the whole text.

*Add-on on business models:* finally gaps have been identified in the area of working business models around data creation and provision, and thereby the developments of respective tools and technologies is often limited to funded research or small experiments. Clearly defined and successfully working business models in place would stimulate the industrial development in the field and thereby stimulate the availability of the required innovative data and metadata management tools.

*Missing tools along the data life cycle:* finally innovative tools along the data life cycle for language technology are missing, not in place sufficiently, and need to be developed along the needs and use cases in the field.

# 6 Data, Language Resources, Knowledge Graphs: Contribution to Digital Language Equality and Impact on Society

This section provides insights how the area of data, language resources and knowledge graphs can contribute to digital language equality and provide an impact on society.

The major issue identified in this area is the lack of availability of relevant and required data and language resources, as well as Knowledge Graphs in all EU member states languages, thereby in the official EU 24 languages. At the moment sufficient data is available mainly in English language and somehow also in the 5 major EU languages: English, German, French, Spanish and Italian. But already in these languages data gaps exist that hinder a digital language equality.

Looking further into this area it is easy to identify an even greater gap of the availability of data regarding dialects to the EU languages as well as to regional language specifics. Although such dialects or regional / local languages or language specifics are in place, actively used and form a part of identity and culture. Diversity in language specifics is so strong that sometimes in a small region several different terminology and/or language specifics are used.

A simple example for this statement is a speech to text engine that is trained in German language and works properly with a German native speaker that speaks clearly and without any local specifics or dialects in language versus an Austrian native speaker located in Vienna who speaks German with a slightly difference in pronunciation and is using some Austrian / Viennese terminology specifics. The results of the speech to text engine clearly start to create errors for the Austrian / Viennese user and produce text that is not precisely showing the input of speech anymore.

In addition there is very little data available for sign languages which is a clear issue regarding inclusion of those with disabilities, and also little to no respective data available for non-EU languages that are widely spoken in the EU, like for example Turkish or Arabic. There are efforts to translate most important content e g by public administration on some regional level to these languages (e. g., for public transport systems, or public administration websites, or by telekom providers that serve their customers that are speaking such languages by web-based self service systems), but this is human translation work only and too little efforts are in place to also develop, collect or acquire relevant data and language resources for such languages, that would allow to develop proper language technology applications like for instance chat bots or other means of conversational AI.

Digital Language Equality is an important part regarding diversity and inclusion and thereby

for a well working European society. Inside of single EU Member States and even more across Europe with its colourful cultures and identities, its regional and local specifics.

The lack of a digital European Language Equality can divide society as it fosters misunderstanding and supports even the promotion of toxic content like fake news or the wrong interpretations of regional policies and regulations or the misinterpretation of scientific results in a crisis. Like for example the current covid-19 pandemic that shows the power of and the need for a digital European Language Equality and thereby clearly shows the strong need of the availability of full-blown multilingual and crosslingual data and language resources, to develop and provide language technology based applications. Following the given pandemic example: communities that are not speaking English or not speaking / understanding the official national languages sufficiently, are excluded from latest information in the field like for instance information about available vaccines or the mode of action of a vaccine or a specific medication. This lack of information automatically provides a good ground for misinformation, toxic content and bias.

Looking even further into the topic, European society has a strong diversity regarding cultural specifics, regarding gender specifics and thereby a strong diversity in, e. g., the meaning of gestures or mimic et al. For future language technology based solutions this area becomes more and more important as well and needs to be taken into account for a European Language Equality.

Taking all the above mentioned viewpoints and specifics into account results in a strong need for action in regard to the acquisition, development and provision of such data and language resources for a broad range of data types required to enable the development of language technology that enables digital language equality. The range of data types needed include, but are not limited to, mono-lingual, bi-lingual and multilingual dictionaries and all types of lexicographical resources, language pairs, annotated corpora, a broad range of language models as well as translation memories and so on and so forth.

Beside technology and data specific requirements the problem needs to be solved how such data and language resources can be acquired or developed and what efforts and actions are needed to allow this? We have identified the following 3 approaches:

**(i) Language Equality Action Plan:** more funding and support by the local and national governments and/or the European Union to support the development of digital language equality in Europe by a clear action plan over the years to come for EU languages as well as regional language specifics and dialects (and if possible for non-EU languages), e. g., in the form of a data and language resource matrix that shows which data and language resources should be available when for which languages. Such programmes should be defined and established by involving all relevant stakeholders from (i) academia (thereby a broad range of research and education), (ii) industry, and (iii) government / public administration and finally (iv) citizen representatives. The implementation can for example follow existing networks like academies of sciences and/or digital innovation hubs et al.

**(ii) Crowdsourcing and citizen science:** the creation of required data needs the support of native speakers as well as linguists with respective language experience and skills, and the support by (data governance) experts providing guidance in regard to the creation of useful and high quality data needed. A good and sustainable approach for this can be the establishment of more crowdsourcing and citizen science platforms, mechanisms and programmes for more languages and dialects. At the moment such mechanisms (e. g., mechanical turk) are available for the major languages only. The development and provision of motivating incentive programmes for such crowdsourcing activities are needed, that can be (i) direct incentives for the work done / time invested as well as (ii) the enablement of clear awareness that such work supports the digital language

equality and thereby new and innovative language technology solutions in respective languages, what thereby contributes back to the contributors.

**(iii) Data related business models:** in the field to foster data creation and acquisition for minor languages and dialects by industry and the private sector. Only if such business models are in place will industry and the private sector invest into the creation and development of data and language resources in more languages and dialects.

In addition – and to allow language equality for certain domains like for instance health or others – there is a strong need in the continuous development and maintenance of monolingual, bi-lingual and multilingual domain specific vocabularies and Knowledge Graphs, to allow the multilingual and cross-lingual development of innovative domain specific applications and solutions that provide value to economy and society in the field of conversational artificial intelligence and beyond (as one major field of future applications for businesses in b2b, and b2c).

Domain specific vocabularies and Knowledge Graph enable interoperability and as of the structured and interlinked approach of its data management bring clear benefits to respective applications in all data- and information-driven fields and solutions, from data- and information management, over metadata management or content management, to knowledge management and learning systems, recommender engines or other horizontal solutions. In regard to vertical solutions really any industry sector can benefit, but also government and public administrations.

The approach to purely translate or even machine translated vocabularies or Knowledge Graphs does not work properly as of the richness in meaning and context of such models, that can be easily lost by a 1:1 translation exercise. This means there is a need to (i) involve domain experts that are usually available in the respective domain / industry field, and to (ii) the further research and development of innovative new methodologies and technologies in semantic context and meaning modeling are needed.

The same applies regarding the need of modeling mechanisms of cultural aspects, gesture and more.

When such action plans and innovations as described above are developed, attention is required on secure and easy-to-understand solutions for personal data collection and sharing, as well as for well working privacy mechanisms and the respective legal framework.

The idea of ownership and sovereignty in personal data, which is getting more and more important at the moment, should be taken into account and supported. This enables data collection and -sharing of personal data for, e. g., important research in the health domain (where patient data is required). The MyData movement[107] as well as Tim Berners-Lee's project SOLID[108] need to be mentioned as potential ideas in the field that should be taken into account to move forward.

A similar importance is required for (i) data and ethics, as well as (ii) an explainable and responsible artificial intelligence. The area of (iii) data literacy as well as of (iv) the need to develop mechanisms and technologies for "toxic data detection" not to forget, to allow end users to make use of such language technology applications and solutions with a clear benefit, and to be able to make informed decisions about the use of such applications.

[107] https://mydata.org
[108] https://solid.mit.edu

| Scenario Characteristic | Challenge scenario | Baseline scenario | High growth scenario |
|---|---|---|---|
| Data innovation | Low level | Healthy growth | High level |
| Concentration of power | A moderate level due to digital markets fragmentation | Moderate by data providers | Low data power concentration |
| Data governance model | Unclear | Protecting personal data rights | Open and transparent |
| Distribution of data innovation benefits in the society | Uneven | Uneven but rather wide | Wide |

Table 2: The 2025 scenarios for data market and data economy

# 7 Data, Language Resources, Knowledge Graphs: Main Breakthroughs Needed

This section provides an overview of the main breakthroughs that are needed in the field of: data, language resources and knowledge graphs regarding a digital language equality.

## 7.1 Data infrastructure, data spaces and datamarkets

Based on the identified components, the state of the art analysis as well as the gap analysis above the area of data infrastructures, data spaces and datamarkets are one major area for the future technology visions and the breakthroughs needed in the field, as this area provides the overall umbrella for the availability and accessibility of required data for powerful (natural) language technologies, as well as for digital language equality.

Parts of the following analysis are citepd from the publication: TRUSTS – D2.1 'Definition and analysis of the EU and worldwide data market trends and industrial needs for growth'[109].

**Industry perspective: a closer look at the market trends and the market size** A recent study by the European Commission (Cattaneo et al., 2020) examines trends of data markets. The study measures *the value of 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 data economy*, i. e., "measures 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. The scenarios are summarized in Table 2.

The EU data market value is likely to increase after 2019 (see Figure 8). It will reach €72 billion and €83 billion in the 2025 challenge scenario and 2025 baseline scenario, respectively. In the most optimistic scenario, it will grow by 10.% compared to 2019 (i. e., to reach €107 billion).

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

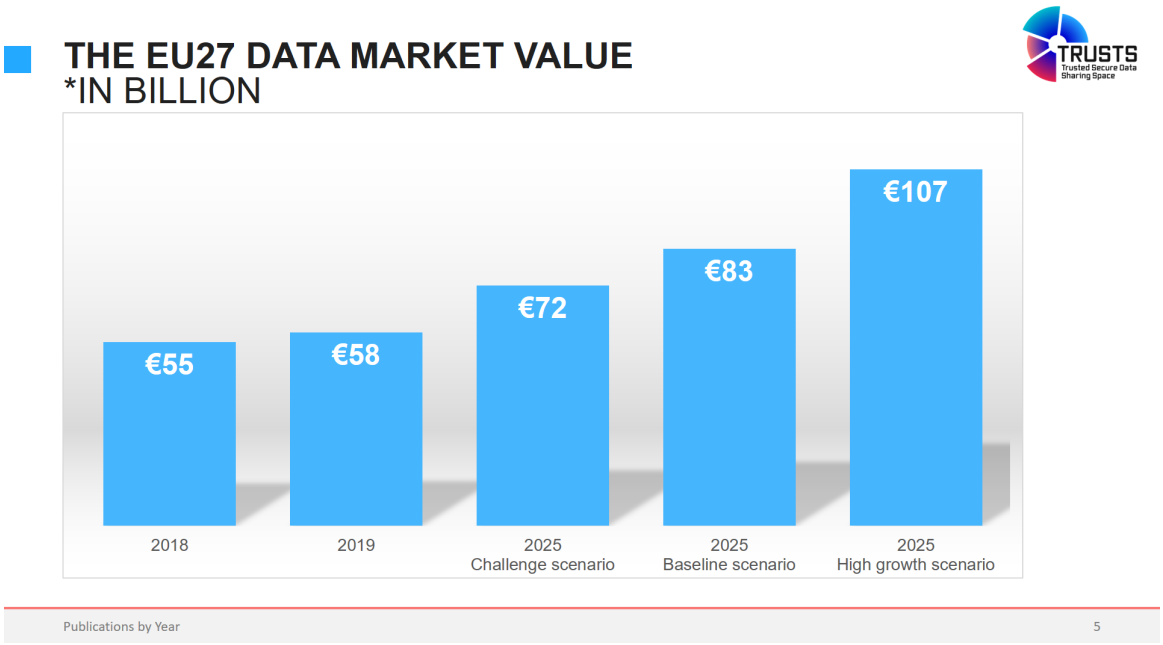


Figure 8: The EU data market value adapted from European Commission

Similar to the general trend of the EU data market value, the data economy's value is also expected to grow positively between 2020 and 2025, as shown in Figure 9. It will reach a value of €432 billion in the 2025 challenge scenario. With a compound annual growth rate of 9,2%, the EU data economy value will increase to €550 billion in the 2025 baseline scenario. In the 2025 high growth scenario, it will reach a value of €827 billion.

The growth trend in the data market and data economy brings several implications. According to the European Commission,[110] for instance, the total number of data professionals (i. e., those who deal with data endeavours as their primary task) will also consistently rise. Many opportunities will open in data-related jobs, and more knowledge workers are needed. Despite its positive trend, there is still a potential lack of data professionals' supply in the high-growth scenario. Following this, companies taking the role as data providers and data buyers will also grow in overall number and share.

**Academic publications** The trend of data marketplace publications in our database is as illustrated in Figure 10.

The trend of publication can be divided into two clusters: the initial phase and the take-off phase. Data marketplaces have emerged in the take-off phase (i. e., after 2010), when the number of publications rapidly increased. Publications regarding stolen data markets (Holt and Lampke, 2010), data markets in the cloud (Balazinska et al., 2011), data as-services (Vu et al., 2012), survey in data marketplaces (Schomm et al., 2013) are among ones that triggered the hype of data marketplaces. Considering the EU investment in data marketplaces, and the projected trends in the data economy, the increasing trend of data marketplace publications is also predicted to continue in the future.

We are now moving to consider publication trends based on co-authorship countries. Figure 11 presents the top 10 countries that actively publish data marketplace articles. Authors from the United States published the most, followed by authors from Germany and China.

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

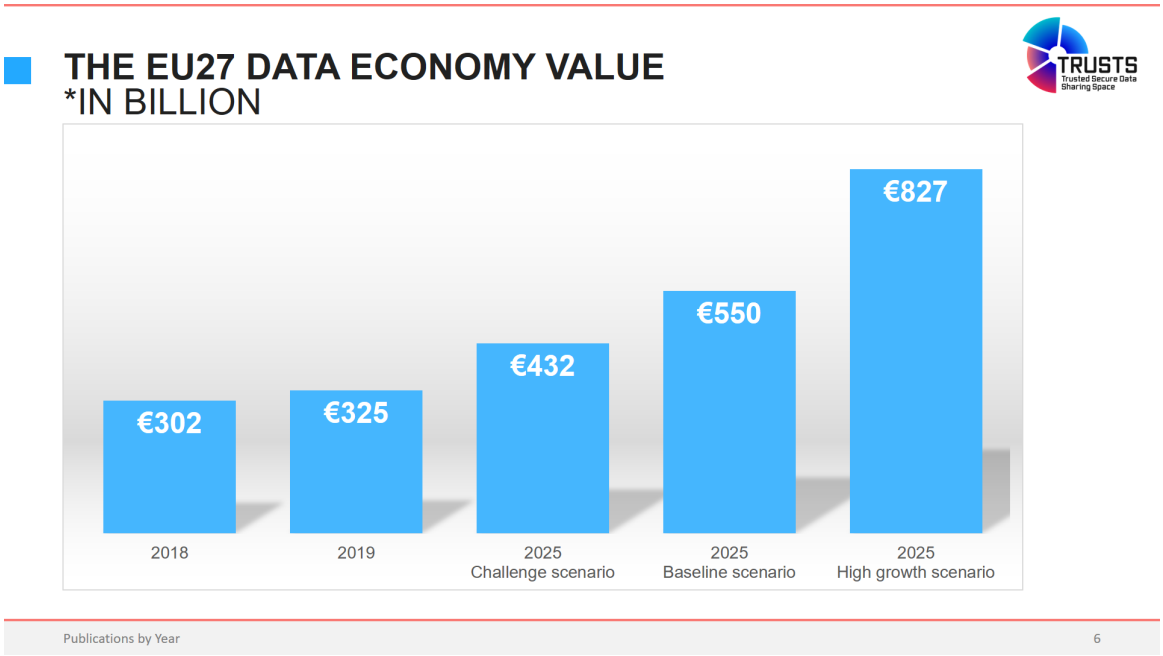


Figure 9: The EU data economy value adapted from European Commission

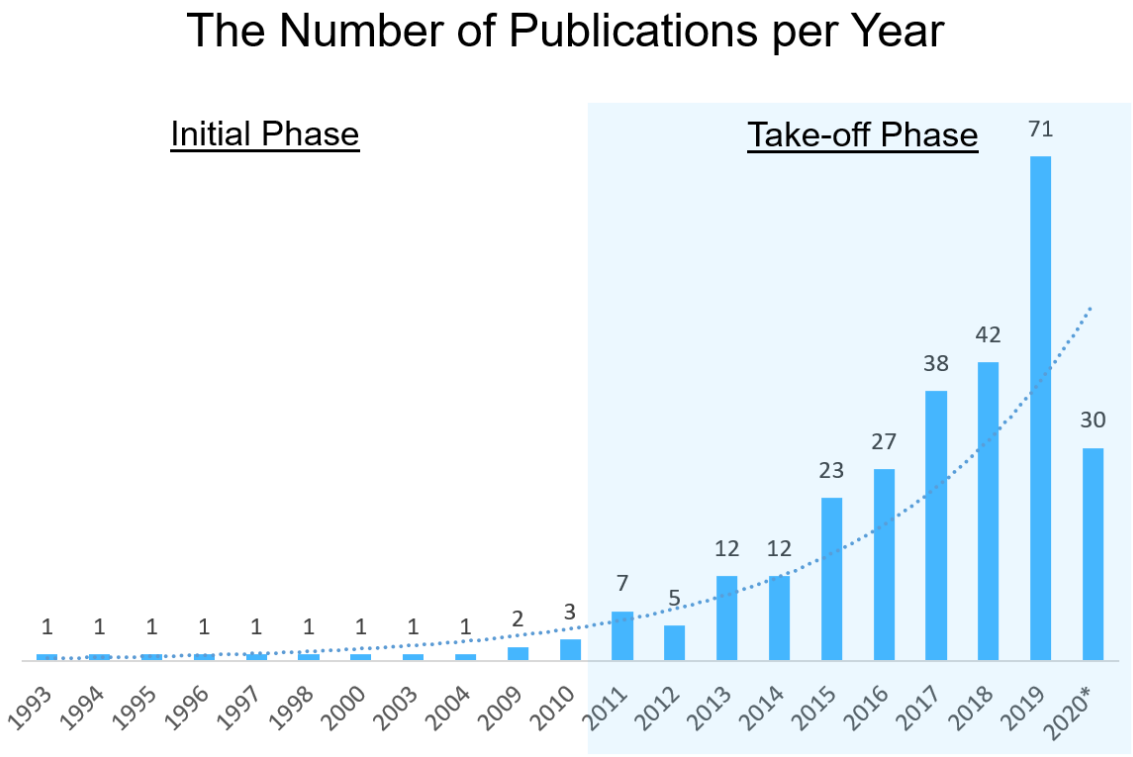


Figure 10: The number of Publications per Year

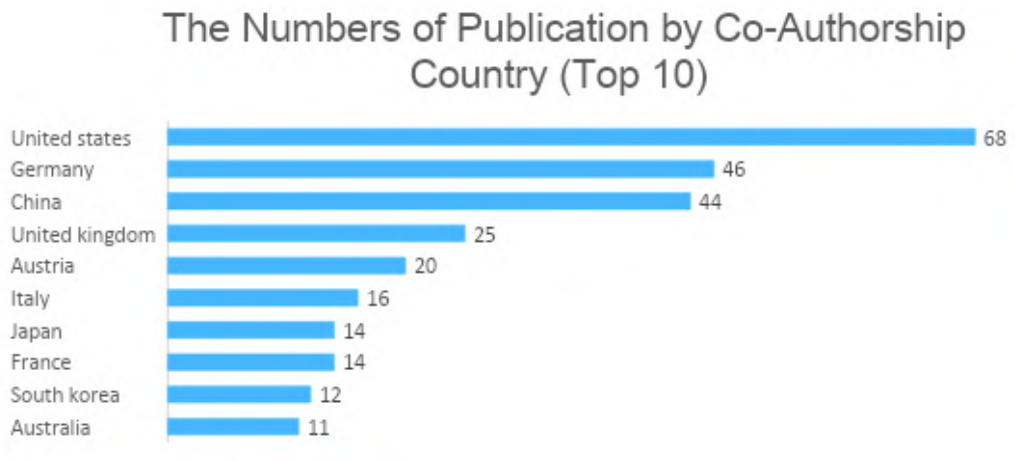


Figure 11: The Numbers of Publication by Co-Authorship Country (Top 10)

From the regional perspective, i. e., continent level, authors from the EU and the UK published the most. The trend may correlate to the EU vision to keep increasing the data economy's value, resulting in many grants being available in the data marketplace domain.

This shows the importance of data spaces and datamarkets in the European landscape regarding data collection and sharing as well as the related trends.

**Main breakthroughs needed: data infrastructures and data spaces**

- Directives and Regulations in a clearly EU-wide legal framework for data infrastructures, data spaces and datamarkets.
- The development, promotion and use of standards regarding secure data sharing and data collaboration that are based on existing standards or expand existing STANDARDS to avoid the development of several new ones that are similar to existing ones. All stakeholders need to be involved in such developments.
- Fostering metadata and data interoperability to ensure easy access and use of the available data for powerful (natural) language technologies with clearly specified licenses, and with fair costs and conditions for all stakeholders AND for all data (and languages).
- Further developments regarding data related services and tools that are provided in the same data infrastructure as the data itself to avoid high costs and high efforts in data evaluation, cleaning, and integration work.
- Successful and sustainable business models for data sharing, data trading and data collaborations that enable the operators of data spaces and data infrastructures to establish a growing business. And the development of business models that distinguish between the use of the data in a data infrastructure by, e. g., research or industry to meet the requirements of the stakeholders.
- Provide models for incentives for data related activities like crowd sourced data cleaning or manual annotations et al for all languages required for a digital European language equality.
- Clear commitment and support by governments in the form of (i) the implementation of the legal frameworks, and (ii) public investments and funding, to establish a sustainable data infrastructure.

- A stronger connection of the fragmented landscape of data infrastructures by interconnecting scientific data infrastructures (like EOSC), industrial data spaces, and data portals provided by governments (e. g., open data catalogues like the European Data Portal). And thereby also provide the needed data for the development of models and algorithms for AI and machine learning.

## 7.2 Knowledge Graphs & Semantic AI

Knowledge Graphs and Semantic AI combined and provided as part of a data infrastructure can bring clear value to such data infrastructure – or better: should be a part of any data infrastructure in the future.

**Industry Trends in Graph Technology** Gartner Research states that from 2021 onwards: graphs form the foundation of modern data and analytics with capabilities to enhance and improve user collaboration, machine learning models and explainable AI. Although graph technologies are not new to data and analytics, there has been a shift in the thinking around them as organizations identify an increasing number of use cases. In fact, as many as 50% of Gartner client inquiries around the topic of AI involve a discussion around the use of graph technology.[111]

And already in 2020: By 2023, graph technologies will facilitate rapid contextualization for decision making in 30% of organizations worldwide.[112]

BFortune Business Insights stated that the global Artificial Intelligence (AI) market size is expected to gain momentum by reaching USD 360.36 billion by 2028 while exhibiting a CAGR of 33.6% between 2021 to 2028. In its report titled, "Artificial Intelligence (AI) Market Size, Share & COVID-19 Impact Analysis, By Component (Hardware, Software, and Services), By Technology (Computer Vision, Machine Learning, Natural Language Processing, and Others), By Deployment (Cloud, On-premises), By Industry (Healthcare, Retail, IT & Telecom, BFSI, Automotive, Advertising & Media, Manufacturing, and Others), and Regional Forecast, 2021-2028" Fortune Business Insights mentions that the market stood at USD 35.92 billion in 2020.[113]

#### Main breakthroughs needed: Knowledge Graphs and Semantic AI

- Develop Knowledge Graph 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.
- Foster the development of multilingual Knowledge Graphs to be available for fair conditions and costs for use and re-use.
- Foster the development of domain specific Knowledge Graphs to be available for fair conditions and costs for use and re-use.
- Knowledge Graphs need a higher level of automation in its creation and maintenance as well as needs to take into account more formats of data beside textual data like audio or video.
- High level of deep and continuous learning enables Knowledge Graphs to maintain itself over time regarding new domain specific and language specific terminology. Means

[111] https://www.gartner.com/smarterwithgartner/gartner-top-10-data-and-analytics-trends-for-2021
[112] https://info.tigergraph.com/gartner-graph-steps-onto-the-main-stage-of-data-and-analytics
[113] https://www.fortunebusinessinsights.com/press-release/artificial-intelligence-market-9227

that new terms are being identified e g via corpora, analysed and put into the graph in the correct position, as well as being applied to the applications used by the Knowledge Graph.

- Bringing together the 2 main AI communities of statistical AI and symbolic AI to work together on future Semantic AI approaches.
- Develop the areas of Responsible AI and Explainable AI by making use of the combination of statistical and symbolic AI (also called: Semantic AI) in multilingual environments to provide AI based applications that bring correct results and benefits for research, industry and society.

## 7.3 Innovative data and metadata management tools

**Industry market trends and market size: Data Management** The global enterprise data management market size is expected to reach USD 208.87 billion by 2028, registering a CAGR of 13.8

The increasing need for on-time data delivery with accuracy is the major driving factor of the overall market growth. Enterprise Data Management (EDM) is a data management platform wherein the data is pulled across from multiple business and enterprise locations and siloes to a central hub.

Therefore, it integrates with any database to store as well as fetch data. EDM allows organizations to implement effective data governance and data quality rules and norms. These rules help businesses and enterprises organize and manage their data better, and enable efficient data-based decision making.[114]

**Industry market trends and market size: Metadata Management** The global Enterprise Metadata Management Market is forecasted 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 control and clarity needed to control the change that often accompanies a complex enterprise data ecosystem. EMM and the various management software created for it provide administration for data integration and allow users to inspect the metadata's links and roles.[115]

**Main breakthroughs needed: Innovative data and metadata management tools**

- Development of tools that can be easily integrated with data infrastructures, data spaces and datamarkets.
- Develop technologies and tools that can identify and remove bias, toxic content as well as fake data from data and content.
- Provision of tools in the field of Semantic AI, and thereby the combination of statistical and symbolic artificial intelligence, that provide out-of-the-box responsible and explainable AI capability.
- Develop towards a tool landscape that can create models and algorithms based on Semantic AI and thereby can create these models and algorithms with clearly less data.
- Tools for data and metadata management that work not only in major languages like English but also can be easily adapted with low costs to small languages, dialects and/or old languages.

[114] https://www.researchandmarkets.com/reports/5415471/global-enterprise-data-management-market-size
[115] https://www.reportsanddata.com/report-detail/enterprise-metadata-management-market

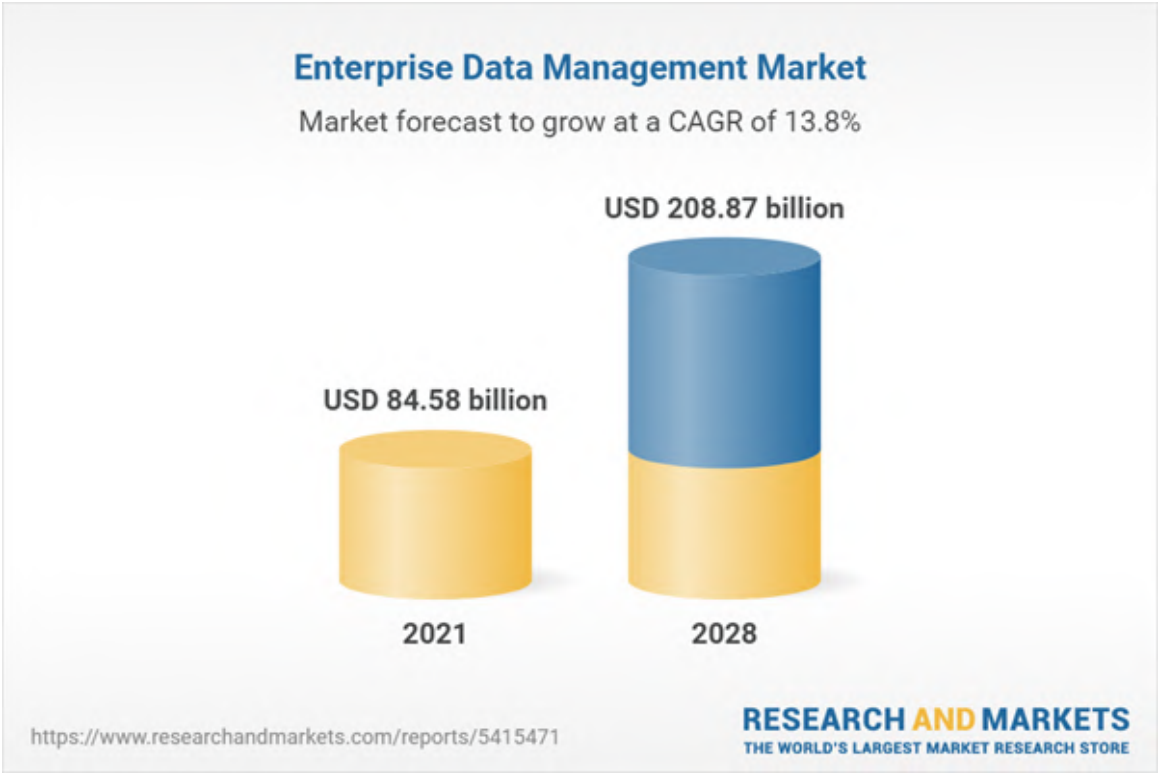


Figure 12: Source: Research and Markets, July 2021.

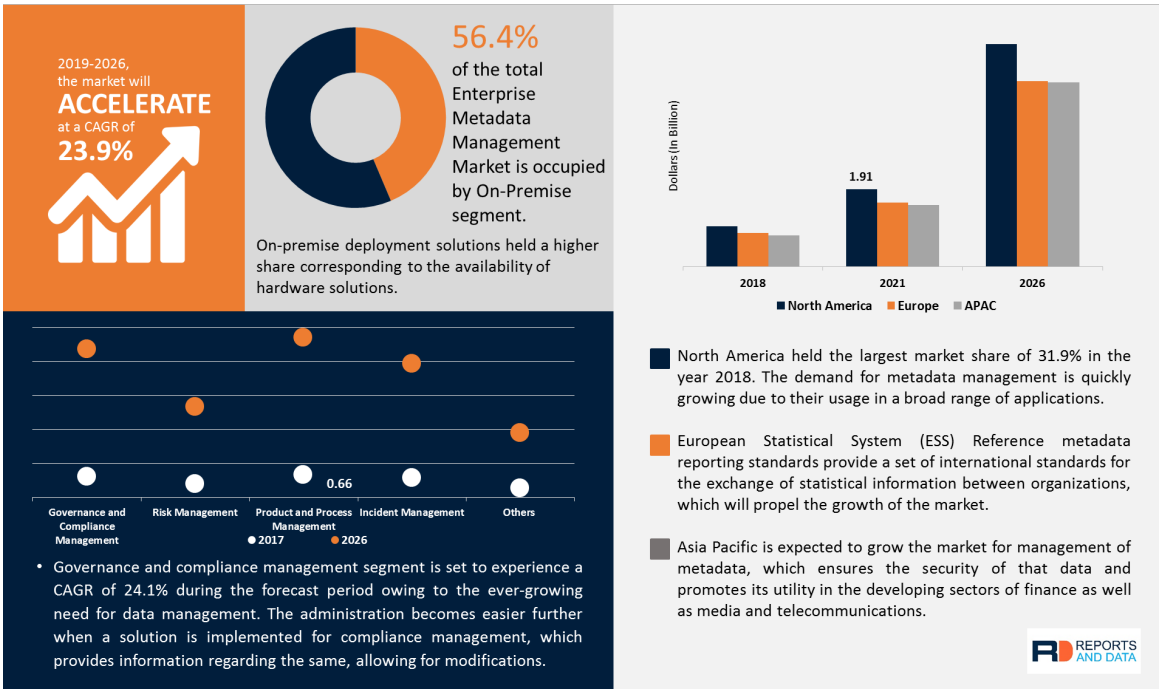


Figure 13: Source: Reports and Data, September 2020.

- Tools that allow deeper modeling of cultural aspects, gender aspects et al to avoid bias in data.
- Tools that are able to combine input from various types of data like textual, images, audio and video but also gestures or behaviour of human beings.
- Tools along the whole data life cycle for all languages and all relevant use cases are required to ensure powerful (natural) language technology and enable digital language equality.

# 8 Data, Language Resources, Knowledge Graphs: Main Technology Visions and Development Goals

This section provides technology visions and development goals for the area of data, language resources and knowledge graphs regarding digital language equality. It starts with a comprehensive list of identified use cases in the field and highlighting the related requirements in data, language resources and Knowledge Graphs to realise such future thinking use cases. Finally the main technology visions and development goals are summarised.

## 8.1 Future Use Cases and Related Requirements

*Use Case: Common Sense Knowledge & Data – availability in all languages* Interfaces and content should be available in ALL languages via the web, meaning that the information available about a specific object or person provides the same amount of information in all languages (for example in Wikipedia: an article about Barack Obama has same and comprehensive information available in English as well as in Finish and other languages).

In addition such a system allows to (i) write content in its own language and provide the content in all languages, as well as consume content in its own language although such content has been created in another language.

In addition Oral Knowledge should become available for all languages, this means knowledge and content that is available in the form of audio files that can be easily consumed by a recipient. The source data for such Oral Knowledge should be taken from textual data as well as from audio-visual data available. This approach of Oral Knowledge supports the idea of a digital language equality in Europe and beyond.

*Data related requirements for this future use case* Tools like machine translation and multi- and cross-lingual summarisation as well as respective speech to text. For this the underlying data in ALL languages need to be available to train and develop proper mono-lingual, bilingual and multilingual models and memories that cover the type of knowledge (domain specific) and the type of language required for the use case.

*Use Case: Speech2Text in the medical domain* A speech-speech translation system in the medical domain should be available to facilitate interaction between medical personnel and non-fluent speakers, thus enabling foreigners/ newly arrived immigrants to access the local health services properly.

*Data related requirements for this future use case* For the machine translation component: parallel datasets (for language pairs) in the medical domain need to be available. The medical domain might be too broad and thereby need to be subdivided/tuned. Speech systems: are required that are trained on data that has sufficient medical coverage/ age ranges/ gender etc. If this is not possible by end-to-end systems, NLP pipeline that are trained in the medical domain are needed.

*Use Case: Multilingual search on public administration/ services websites* For example: if a person is new to a country it facilitates the same level of access to information that the

citizens of that country have. As a remark: currently people use expat groups on social media, or other sources. Thereby also: a post that is made on social media could automatically provide potential answers with links to relevant websites?

*Data related requirements for this future use case* This case is specifically related to public services but can apply more widely. Such Question-Answering systems rely on digitally accessible data in public administration websites/archives. Reliable machine translation systems will need to run in the background to localise the correct language to the specific query. Efficient APIs (application programming interfaces) are needed to integrate such data and systems with social media platforms.

*Use Case: Screen-readers for all languages (with summarisation tools combined)* The amount of content we need to read online is creating screen fatigue, and screen-readers are also needed for visually impaired, reading disabilities.

*Data related requirements for this future use case* One requirement is working speech synthesis tools. Furthermore summarisation tools in all required languages, which require NLP pipelines of tokenizers, taggers, parsers etc., which require labelled linguistic datasets (e. g., treebanks) and evaluation sets available.

*Use Case: Virtual Assistant for healthcare to assist diagnostics and treatment.*

*Data related requirements for this future use case* Pseudonymised clinical data for all EU languages, as well as annotated corpora for clinical text, and finally: multilingual and comprehensive ontologies in the health domain. Furthermore pre trained and fine-tuned BERT models for the medical domain for at least all EU 24 languages

*Use Case: Virtual Assistant for for visually impaired and for deaf people* To allow them to deal with daily activities, as well as to include them in all activities around education and work.

*Data related requirements for this future use case* Text to Speech resources for common vocabularies / terminologies are required, as well as computer vision for sign languages. Knowledge Graphs for common concepts, events description for daily activities, and finally patterns for frequent questions are needed.

*Use Case: Virtual Assistant for emigrants/refugees/minorities* People that are not quite familiar with the official language of the country that will assist their daily communication, as well as communication with the administrative offices and authorities.

*Data related requirements for this future use case* Language resources (speech, text) for EU languages AND besides the EU languages are needed. As well as for EU languages and Non-European languages: for some dialects and languages of minorities.

*Use Case: Virtual Assistant as a social assistant for elderly people, and people with Alzheimer, dementia etc.* That will assist their daily activities and even can act as a friend for social communication to not be lonely.

Remark: the list of such Virtual Assistants can be expanded for a huge number of use cases as for instance: a Virtual Assistant for Industry, or a Virtual Assistant for law and legal issues at a court, and many more.

*Use Case: speech translation connected to a chatbot for some public administration* For example the National Health Service England, where a chatbot can manage the call or in-person demand for a user that is non-speaking the local language.

*Data related requirements for this future use case* Here we would need to handle personal information and thereby a strong need for anonymisation is in place. Furthermore speech models that address 24 (EU) languages are required. As well as data and models which address gender bias and / or minority bias et al..

*Use Case: Automatic adaptation of content to match permanent or situational cognitive abilities of the information consumer*

*Data related requirements for this future use case*

1. Model(s) of cognitive abilities

2. Adaptive content generation or transformation

*Use Case*: Anonymizer/personal data advisor, that warns a user if the contributed content may reveal personal information, or if an input consuming service may not comply with GDPR or other privacy rules or data related regulations, or even starts collecting personal identifiable information (PII).
*Data related requirements for this future use case*

1. Model(s) of personal traits that might get exposed via text or voice
2. Model(s) of GDPR regulations

*Use Case: Challenges/games that allow citizen scientists to create, curate, or assess language-related assets.*
*Data related requirements for this future use case*

1. Model(s) of language-related assets
2. Model(s) of assessment for language-related assets

*Use Case:* Conversational agent, that can mediate between me and a friend in a different language and cultures by providing a high precision of understanding of each other. For instance a person speaking English and a person speaking Japanese: cultural and language specifics are added to the conversation that is guided by the conversational agent.
*Data related requirements for this future use case* More raw data for the respective languages, as well as the ability that modelling will evolve faster than language changes, as well as the ability of modeling cultural specifics.
*Additional collected data related requirements* The lack of annotated data is a huge barrier for many systems at the moment and thereby this needs to be improved for the future. There is much need for better designed crowdsourcing platforms to enable more citizen science efforts towards building speech and language systems. Integrated with mobile applications where a push notification requests a user's response to a small task. On a large scale, this could be very effective (e. g., collecting speech data).
Another general requirement for development of many systems: usually the challenge for many industries is the lack of domain specific 'annotated' data. Premise: when dealing with specific terminology, the client knows their data better than 3rd party vendor annotators. Inhouse data curation and annotation tools would therefore be a game-changer for future developments.
Additional future use cases have been collected as follows (all with the minimum requirements of multilingualism):

**Event Detection:** detecting notable events around the world to alert governments and newsrooms, for example, a natural disaster such as an earthquake or a volcano is usually discussed on social media before any newsrooms/governments have details of the event. Similarly: alerts to first responders regarding potential riots, demonstrations etc. and automated detection and information in local languages.

**Anti Money Laundering systems** require proper language data and resources as this is done across borders.

**Urgency Detection** for example, in customer support ticket system/ flagging urgent issues to prevent such tickets from being lost in a backlog.

**Brand Harming Alerts:** monitoring social media for clients to ensure their brand isn't being unduly held in a negative light. Identification and interpretation is required in many languages as well as in dialects.

**Document retrieval:** for example in a manufacturing industry that creates new contracts on a regular basis. NLP that is trained with available language resources in all relevant languages meet the need for referring to previous similar contracts.

**Scientific Research/Legal Research:** finding relevant documents from previous cases.

**Social listening/ market intelligence:** monitoring social media to assess how the brand/company/product is being regarded/discussed across languages and cultures.

**Social media monitoring:** removing toxic content, hate speech, fake news, thus making the internet a safer and more positive environment. Or identifying and highlighting such content and using this for media literacy approaches.

**Multi- and cross-lingual text summarisation:** for example, for newsrooms and/or medical research

**Natural Language Generation (NLG):** automatic report generation/ automatic FAQ generation , that is based on queries received either in phone conversation or via email/social media

As well as:

**Effective Podcast search:** keyword search (text or audio prompt) through audio files, especially useful when metadata or sufficient description of podcast is not present.

**Actionable Audio Ads:** also referred to as Voice Commerce. The idea is that users can interact with any audio ads that arise during audio streaming, e. g., you're out jogging/walking while listening to audio online and you hear an ad you're interested in. You can interrupt and ask to hear more/ order something there and then.

**Voice controlled shopping:** similar to above but not through ad prompts, but actual commerce sites.

**Intelligent cars:** for example with a broken tyre on the road, as manuals are often black & white text that are unhelpful for navigation, it takes a while to figure out what an alert could exactly mean, and after pumping up the tyre, and searching the web for a video to help to figure out how to manually reset the alert needs et al. Here, having a conversation with the car in this instance would be much more effective! Or if the car wasn't that high-tech, maybe a car-support platform that allows you to submit a photo of the alert light and receive a series of steps to resolve the issue.

This longer list of future use cases have been collected to provide a better picture on the Technology Visions for (Natural) Language Technology for 2030 and beyond, and thereby to better understand the need regarding data, language resources, and Knowledge Graphs.

All in all the majority of these use cases raise the topic of human-machine communication and interaction, as well as human to human communication and interaction by making use of digital tools and thereby can be categorised by the technology concepts of *Conversational AI / Platforms and Insight Engines*, that are well covered by other deep dives.

## 8.2 Future Technology Vision

In regard to the future technology visions in the area of: data, language resources and Knowledge Graphs the previous section providing the main breakthroughs for the following areas should be studied in detail:

- Data infrastructures, data spaces and datamarkets

- Knowledge Graphs and Semantic AI
- Innovative metadata and data management tools

A starting point for this Technology Vision in data for 2030, to support (Natural) Language Technology is the vision of a *Semantic Data Fabric*.

Gartner Research identified Data Fabrics as a clear trend in their Top 10 Data and Analytics Technology Trends already for 2019[116] but first real world implementations are not available before 2022 or even 2023.

Data fabric enables frictionless access and sharing of data in a distributed data environment. It enables a single and consistent data management framework, which allows seamless data access and processing by design across otherwise siloed storage. Through 2022, bespoke data fabric designs will be deployed primarily as a static infrastructure, forcing organisations into a new wave of cost to completely re-design for more dynamic data mesh approaches.

**The Data Fabric approach** Gartner Research provides specification of and insights into the Data Fabric approach as follows:[117]

> Data management teams are under constant pressure to provide faster access to integrated data across increasingly distributed landscapes. Data and analytics leaders must upgrade to a data fabric design that enables dynamic and augmented data integration in support of their data management strategy.
>
> Impacts and Recommendations
>
> - ML-Augmented Data Integration is making active metadata analysis and semantic knowledge graphs pivotal parts of the data fabric,
> - Data Fabric must have the ability to collect and analyse all forms of metadata,
> - Data Fabric must have the ability to analyse and convert passive metadata to active metadata,
> - Data Fabric must have the ability to create a knowledge graph that can operationalise the data fabric design,
> - Data Fabric must enable business users to enrich the data models with semantics,
> - Extreme levels of distribution, scale and diversity of data assets add complexity to Data Integration Design and Delivery,
> - A strong Data Integration Backbone is necessary for versatile Data Sharing in support a Data Fabric Design,
> - Core Data Fabric functionalities now appear in many separate data management tools; Distinction among them is blurring,
> - Delivering the Data Fabric with a combination of tools and capabilities.

**Towards a Semantic Data Fabric** A Semantic Data Fabric is a new solution to data silos that combines the best-of-breed technologies, data catalogs and knowledge graphs based on Semantic AI.[118]

[116] https://www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo
[117] https://www.gartner.com/en/documents/3978267/data-fabrics-add-augmented-intelligence-to-modernize-you
[118] https://www.poolparty.biz/what-is-a-semantic-data-fabric

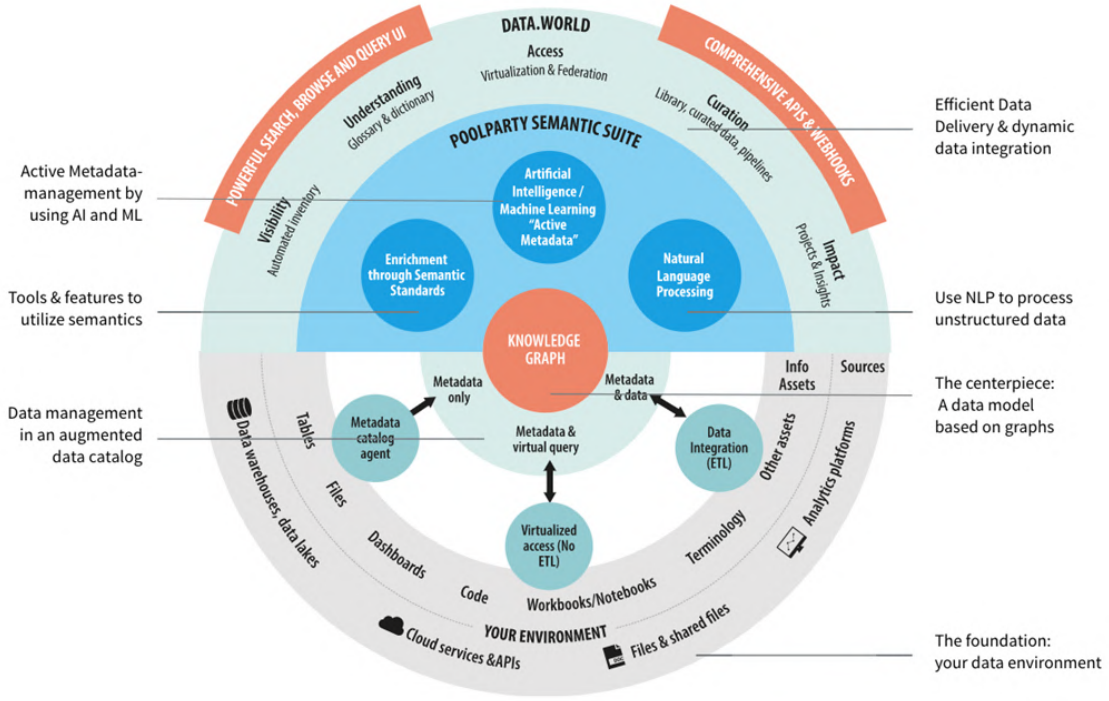


Figure 14: The Semantic Data Fabric, White Paper, A New Solution to Data Silos. Source: data.world & PoolParty Semantic Suite

"The Semantic Data Fabric combines the respective advantages of Data Lakes and Data Warehouses and complements them especially with the advanced linking methods that Semantic Graph Technologies bring with them".[119]

Data.world,[120] a US based data catalogue vendor and Semantic Web Company,[121] an Austrian semantic middleware vendor (of PoolParty Semantic Suite) has adapted and expanded this approach to an integrated view as depicted in the Figure 14.

As this figure above shows the main cornerstones of the Semantic Data Fabric approach are:

**Active Metadata:** by making use of machine learning and AI to enable dynamic metadata over time and not only once when importing a metadata set manually.

**Tools and features to use semantics:** to provide context for data sets and data objects

**Use NLP to process unstructured data:** the main part of data in organisations is still unstructured data like documents. Thereby powerful analysis mechanisms are required to get the full benefit of such (unstructured) data.

**A data model based on graphs:** graph technologies can bring more benefits to data management than traditional data models, as of the organisation of data and data objects in a graph-based structure, that allows complex querying as well as easy and fast querying.

[119] https://www.poolparty.biz/the-knowledge-graph-cookbook/
[120] https://data.world
[121] https://www.semantic-web.com

To make use of the Semantic Data Fabric ideas and principles for data infrastructures, data spaces and datamarkets should be the next step in the evolution in this area. Many parts and components are used in different data infrastructures already today, but the integrated combination of all of them could – from a technology perspective – be the main breakthrough and the technology vision for the future management of metadata and data, as well as of language resources, that can act as the backbone for powerful Language Technologies. To realise the envisaged digital language equality in Europe and beyond. Existing language technology data infrastructure providers, as for instance the European Language Grid or ELDA and others, as well as newly developed data spaces can thereby easily be used in a kind of a federated data infrastructure / network, by means of interoperability, mainly provided by Knowledge Graph technology. And thereby the future language technology use cases in the fields of conversational AI and insight engines can be realised.

# 9 Data, Language Resources, Knowledge Graphs: Towards Deep Natural Language Understanding

A lot of areas of this deep dive on: data, language resources, and Knowledge Graphs have already provided state of the art, gap analysis and outlook for a deep natural language understanding.

The way to come there is for sure again the listed components for data:

- Availability of data and metadata
- Accessibility of data
- Quality of data
- Data Interoperability
- Licenses and data related regulations
- Data and ethics
- Data literacy

And related to these components regarding data and metadata the following related technology concepts, methodologies and tools that have been identified:

- Data infrastructures, data spaces and datamarkets
- Knowledge Graphs
- Semantic AI: statistical and symbolic AI in combination
- Innovative data and metadata management tools.

In addition the following key areas are of high importance in regard to data for a Deep Natural Language Understanding.

- The ability to model emotions
- The ability to model cultural specifics and thereby cross-cultural understanding
- The availability of world knowledge and situation knowledge in as many languages as possible

- And thereby tools that allow the modeling as well as the continuous learning of such attributes

Continuous adaptations of the language resources, in all languages and by means of automated and manual / handcrafted mechanisms are key for deep natural language understanding, to ensure new terminology appearing is immediately taken into account and provided in mono-, bi-, and multi-lingual data to ensure that new topics, as for instance the COVID-19 pandemic, can be handled properly but also the development and the impact can be fully understood by a broad population to e. g. avoid bias, but also a split in society, as issues in language equality clearly support the divide of societies.

# 10 Summary and Conclusions

This technology deep dive: data, language resources, and Knowledge Graph provides a kind of an add-on document to the other deep dives of the ELE project, as data builds the basis and backbone for technologies and solutions in the area of (Natural) Language Technology and thereby for a Digital European Language Equality.

This report provides a technology deep dive with a clear focus on data and language resources required for a full language equality in Europe by 2030. The document on hand provides insights into: (i) the main components of this technology deep dive, (ii) the current state of the art, (iii) the main gaps identified in the field, (iv) a chapter about the contribution to digital language equality and the impact on society, (v) an analysis of the main breakthroughs needed, and (vi) the main technology visions and development goals identified, as well as as (vii) a chapter regarding deep natural language understanding and data, and is finally closed (viii) by a summary and conclusions section.

The methodology for the creation of this deliverable has taken into account (i) desktop research, (ii) two virtual workshops with ELE consortium members on the topics: (a) state of the art and main gaps and (b) future use cases and requirements regarding data, as well as future technology visions, and (iii) discussions with industry representatives about the topics listed in (ii).

*The following components have been identified for this deep dive:*

The main components of this technical deep dive: Data, Language Resources, Knowledge Graphs have been identified as follows:

- Availability of data and metadata
- Accessibility of data
- Quality of data
- Data Interoperability
- Licenses and data related regulations
- Data and ethics
- Data literacy

All of these components need to be tackled in the future to allow data collection and provision with fair conditions and costs for all relevant stakeholders to develop towards a digital European Language Equality.

Related to these components regarding data and metadata the following related technology concepts, methodologies and tools have been identified, that are currently on the rise and will also be part of the technology vision for 2030 in this deep dive document:

- Data infrastructures, data spaces and datamarkets
- Knowledge Graphs
- Semantic AI: statistical and symbolic AI in combination
- Innovative data and metadata management tools.

As an add-on component in this deep dive the topic of *data-related business models* is tackled throughout the document, as we have identified the importance of working and sustainable data-related business models as a prerequisite for a working data economy and ecosystem that thereby stimulates and fosters the above listed data related components, and thereby a well functioning language technology landscape as a basis for a digital European language equality.

These components have been (i) defined, (ii) used for a state of the art analysis and (iii) a gap analysis and build a part of the forward looking chapters of this document that are (a) the contribution to European Language Equality and the impact on society, (b) the main breakthroughs needed, and (c) the technology visions of this deep dive.

A quite long list of future use cases have been collected and described, the related data requirements specified, and the main technological areas have been identified as (i) conversational AI, and (ii) insight engines.

Beside technology, interoperability or data related attributes there must be a strong focus established on applying all these mechanisms and methodologies to the widest range of languages possible, at least to EU languages but also local and regional dialects of these languages, as well as to non-EU languages that are wide-spread across Europe. Without such data and language resources in place a digital language equality cannot be reached.

To fill the identified gaps in data, language resources, and Knowledge Graphs we recommend and suggest a future path for Europe towards comprehensive and interlinked data infrastructures. These infrastructures have to provide interoperability out-of-the-box by following harmonised and well-proven standards, regarding (i) data (semantic data) interoperability as well as (ii) services and (iii) innovative metadata and data management tools that are available along all steps of the data life cycle.

Metadata, data, data-driven services and data-driven tools to be easily docked into these data infrastructures, without todays' huge efforts in data cleaning and data integration, or service- and tool integration. This future technology vision of integrated and interoperable data infrastructures shall follows the idea of a Semantic Data Fabric including rich semantics, and thereby context and meaning as well as dynamic metadata and augmented metadata- and data management. By this approach a federated network and infrastructure of interlinked data spaces for language technology can be realised. Existing data spaces as well as newly developed ones should be integrated, where appropriate and possible.

In such a federated ecosystem relevant data regarding a domain and/or language can easily be identified, loaded, and evaluated for specific use cases. Data driven services are provided and can be used along end users requirements. Integrated crowdsourcing and/or citizen science mechanisms allow human-machine interaction to foster data acquisition, cleaning and enrichment (e. g., annotation, classification, quality validation and repair, domain specific model creation, et al.). Raw data can be loaded into available tools to train algorithms or create memories and/or (language) models for specific use cases, but also existing algorithms, models or vocabularies are available and can be easily loaded and re-used to avoid unnecessary energy consumption / computing power to foster the idea of energy efficient data management.

In addition high importance needs to be put on privacy protection (related to personal identifiable information, PII and beyond), the avoidance of bias (for example on gender et al.), and on data sovereignty.

The approach of such data infrastructures require working and sustainable business models that allow data trading, -sharing and collaboration. And require supporting policies, as well as sustainable data governance models around data creation, data provision and data sharing. Well targeted publicly funded/supported programmes and activities in the area of data literacy are required from early education onward, to ensure that sufficient human resources in the field are available in the future.

In addition an action plan for the collection and the development of data and language resources that are relevant for language technology, as well as for Knowledge Graphs is needed to ensure the availability of sufficient data in the EU languages, as well as in dialects and important non-EU languages. The recommendation for this is to look into three areas, as: (i) Language Equality Action Plan by means of targeted national and European funding along a matrix of relevant resources and languages, combined with (ii) more measures in the fields of crowdsourcing and citizen science, and (iii) the development of functioning data related business models.

Europe has difficult prerequisites for digital language equality but at the same time a huge potential to become a world leader in language technology and thereby also a role model for digital language equality. Reasons for this are (i) the specifics of the European language space with the EU official languages, a broad range of dialects and old languages, as well as a high number of Non-EU languages in use by a growing number of citizens across the continent and the European Union, (ii) the European societal characteristics with rich variety and diversity in cultures and thereby in the European society, and (iii) the overall challenging requirements of the continuous digitization in a more and more globalised world and the related strong needs for an efficiently working (language) data infrastructure, that provides a rich, easy to use and sustainable backbone for the European (natural) language technology.

The availability of high quality data, language resources and knowledge graphs in at least EU 24 languages, but furthermore 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 can support clear benefits and competitive advantage for the stakeholders.

For the European research community to foster innovations in the field, for the industry to successfully compete in a global market, and thereby for the European citizens and its society, that is constantly growing in regard to its diversity and a wide and increasing variety of languages. Data, language resources, and Knowledge Graphs are thereby a crucial factor on our way to digital European Language Equality.

## References

Bilal Abu-Salih. Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, 185:103076, 2021.

Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, WIMS '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450354899. doi: 10.1145/3227609.3227689. URL https://doi.org/10.1145/3227609.3227689.

Magdalena Balazinska, Bill Howe, and Dan Suciu. Data markets in the cloud: An opportunity for the database community. *Proceedings of the VLDB Endowment*, 4(12):1482–1485, 2011.

Andreas Blumauer and Helmut Nagy.

Kurt D. Bollacker, Robert P. Cook, and Patrick Tufts. Freebase: A shared database of structured general human knowledge. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence,*

*July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 1962–1963. AAAI Press, 2007. URL http://www.aaai.org/Library/AAAI/2007/aaai07-355.php.

Harry Bouwman, Henny de Vos, and Timber Haaker. *Mobile service innovation and business models*. Springer Science & Business Media, 2008.

Gabriella Cattaneo, Giorgio Micheletti, Mike Glennon, Carla La Croce, and Chrysoula Mitta. The european data market monitoring tool. *Key facts & figures, first policy conclusions, data landscape and quantified stories. Final study report. Luxembourg: Publications Office of the European Union*, 2020.

Organisation For Economic Co-operation and Development. *OECD Glossary of Statistical Terms*. OECD PUBLICATIONS, 2008.

Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294*, 2021.

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic AI: The 3rd Wave. *arXiv e-prints*, art. arXiv:2012.05876, December 2020.

Artur S. d'Avila Garcez, Krysia Broda, and Dov M. Gabbay. Neural-symbolic learning systems - foundations and applications. In *Perspectives in neural computing*, 2002.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Claudia d'Amato. Machine learning for the semantic web: Lessons learnt and next research directions. *Semantic Web*, 11:195–203, 2020.

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

Jennifer Golbeck. Chapter 9 - entity resolution and link prediction. In Jennifer Golbeck, editor, *Analyzing the Social Web*, pages 125–149. Morgan Kaufmann, Boston, 2013. ISBN 978-0-12-405531-5. doi: https://doi.org/10.1016/B978-0-12-405531-5.00009-2. URL https://www.sciencedirect.com/science/article/pii/B9780124055315000092.

Thomas J Holt and Eric Lampke. Exploring stolen data markets online: products and market forces. *Criminal Justice Studies*, 23(1):33–50, 2010.

Georgy Ishmaev. The ethical limits of blockchain-enabled markets for private iot data. *Philosophy & Technology*, 33(3):411–432, 2020.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Pantelis Koutroumpis, Aija Leiponen, and Llewellyn DW Thomas. Markets for data. *Industrial and Corporate Change*, 29(3):645–660, 2020.

Sebastian Lawrenz, Priyanka Sharma, and Andreas Rausch. Blockchain technology as an approach for data marketplaces. In *Proceedings of the 2019 International Conference on Blockchain Technology*, pages 55–59, 2019.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. doi: 10.3233/SW-140134. URL https://doi.org/10.3233/SW-140134.

H Steve Leslie and Natalie A Johnson-Leslie. Gender equity in business schools–perception or reality: A conventional content analysis. *Global Journal of Business Disciplines*, 4(1):44, 2020.

Xinyu Li, Mengtao Lyu, Zuoxu Wang, Chun-Hsien Chen, and Pai Zheng. Exploiting knowledge graphs in industrial products and services: A survey of key aspects, challenges, and future perspectives. *Computers in Industry*, 129:103449, 2021.

Kang Liu, Wuhui Chen, Zibin Zheng, Zhenni Li, and Wei Liang. A novel debt-credit mechanism for blockchain-based data-trading in internet of vehicles. *IEEE Internet of Things Journal*, 6(5):9098–9111, 2019.

Chaoyue Niu, Zhenzhe Zheng, Fan Wu, Shaojie Tang, and Guihai Chen. Online pricing with reserve price constraint for personal data markets. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: Lessons and challenges. *Communications of the ACM*, 62 (8):36–43, 2019. URL https://cacm.acm.org/magazines/2019/8/238342-industry-scale-knowledge-graphs/fulltext.

Hyeontaek Oh, Sangdon Park, Gyu Myoung Lee, Hwanjo Heo, and Jun Kyun Choi. Personal data trading scheme for data brokers in iot data marketplaces. *IEEE Access*, 7:40120–40132, 2019.

Marcelo Iury S Oliveira, Glória de Fátima Barros Lima, and Bernadette Farias Lóscio. Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems*, 61(2):589–630, 2019.

Carla Parra Escartín, Teresa Lynn, Joss Moorkens, and Jane Dunne. Towards transparency in NLP shared tasks. *arXiv e-prints*, art. arXiv:2105.05020, May 2021.

Gowri Sankar Ramachandran, Rahul Radhakrishnan, and Bhaskar Krishnamachari. Towards a decentralized data marketplace for smart cities. In *2018 IEEE International Smart Cities Conference (ISC2)*, pages 1–8. IEEE, 2018.

Xiaoqi Ren, Palma London, Juba Ziani, and Adam Wierman. Datum: Managing data purchasing and data placement in a geo-distributed data market. *IEEE/ACM Transactions on Networking*, 26(2):893–905, 2018.

Md. Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-symbolic artificial intelligence: Current trends. *ArXiv*, abs/2105.05330, 2021.

Fabian Schomm, Florian Stahl, and Gottfried Vossen. Marketplaces for data: an initial survey. *ACM SIGMOD Record*, 42(1):15–26, 2013.

Laura Sebastian-Coleman. *Measuring data quality for ongoing improvement: a data quality assessment framework*. Newnes, 2012.

Jan-Terje Sørlie and Jörn Altmann. Sensing as a service revisited: A property rights enforcement and pricing model for iiot data marketplaces. In *International Conference on the Economics of Grids, Clouds, Systems, and Services*, pages 127–139. Springer, 2019.

Ahmet Soylu, 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. *Enhancing Public Procurement in the European Union Through Constructing and Exploiting an Integrated Knowledge Graph*, pages 430–446. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Science and Business Media Deutschland GmbH, Germany, 2020. ISBN 9783030624651. doi: 10.1007/978-3-030-62466-8_27. 19th International Semantic Web Conference, ISWC 2020 ; Conference date: 02-11-2020 Through 06-11-2020.

Markus Spiekermann. Data marketplaces: Trends and monetisation of data goods. *Intereconomics*, 54 (4):208–216, 2019.

Shefali Virkar, Gabriela Viale Pereira, and Michela Vignoli. Investigating the social, political, economic and cultural implications of data trading. In *International Conference on Electronic Government*, pages 215–229. Springer, 2019.

Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

Quang Hieu Vu, Tran-Vu Pham, Hong-Linh Truong, Schahram Dustdar, and Rasool Asal. Demods: A description model for data-as-a-service. In *2012 IEEE 26th International Conference on Advanced Information Networking and Applications*, pages 605–612. IEEE, 2012.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

Dehai Zhang, Menglong Cui, Yun Yang, Po Yang, Cheng Xie, Di Liu, Beibei Yu, and Zhibo Chen. Knowledge graph-based image classification refinement. *IEEE Access*, 7:57678–57690, 2019.