Are our knowledge graphs trustworthy?

Elena Simperl
King’s College London
@esimperl

International workshop on knowledge science

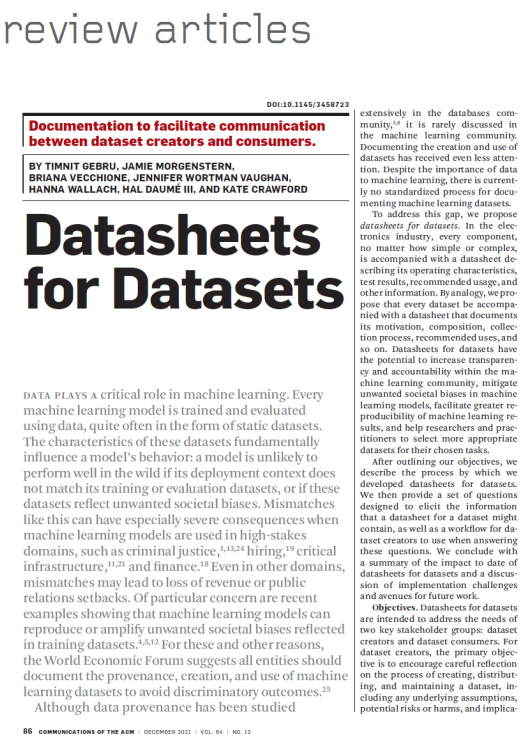
March 2022, online

Knowledge graphs are machine-readable data organised for general-purpose use

Applications in search, recommenders, virtual assistants, enterprise data integration etc.

Used in AI systems alongside machine learning as source of domain knowledge, transfer learning, explanations

Can we trust them?



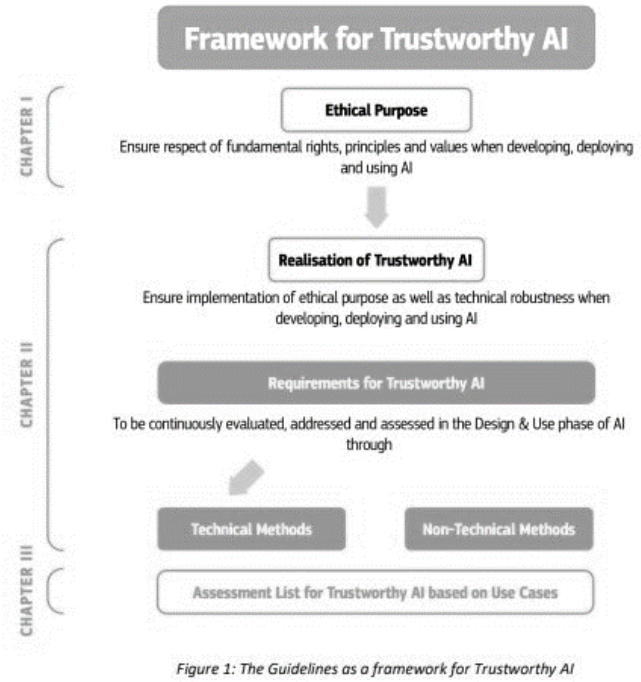
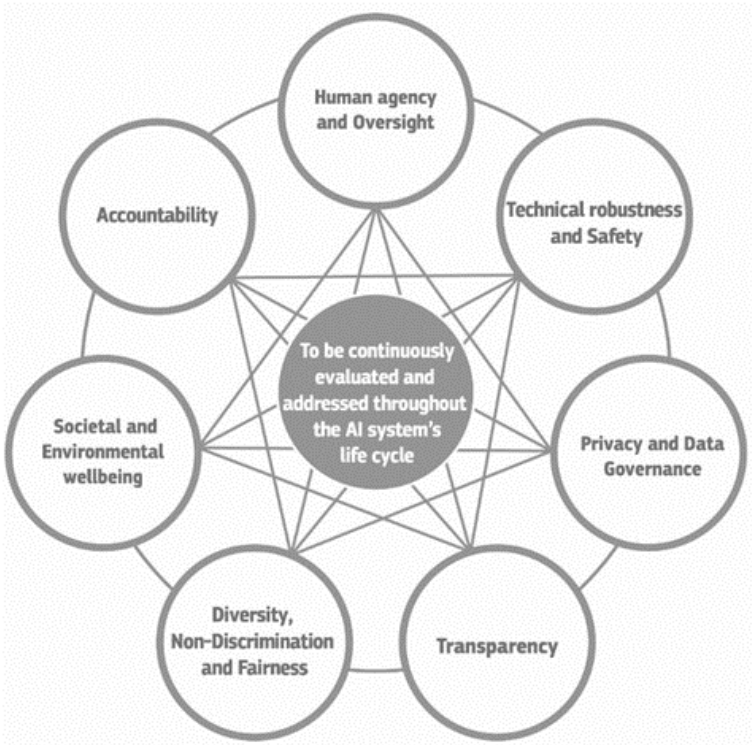


Figure 1: The Guidelines as a framework for Trustworthy AI



What do we mean by trust?

Process (knowledge engineering) vs outcome (knowledge graph as a resource)

Frameworks, methods, guidance on trustworthy AI, data quality, data interoperability standards

We know how good the data is

What we talk about when we talk about Wikidata quality: a literature survey

Alessandro Piscopo*
University of Southampton
Southampton, United Kingdom
alessandro.piscopo@bbc.co.uk

Elena Simperl
University of Southampton
Southampton, United Kingdom
E.Simperl@soton.ac.uk

ABSTRACT

Launched in 2012, Wikidata has already become a success story. It is a collaborative knowledge graph, whose large community has produced so far data about more than 55 million entities. Understanding the quality of the data in Wikidata is key to its widespread adoption and future development. No study has investigated so far to what extent and which aspects of this topic have been addressed. To fill this gap, we surveyed prior literature about data quality in Wikidata. Our analysis includes 28 papers and categorise by quality dimensions addressed. We showed that a number of quality dimensions has not been yet adequately covered, e.g. accuracy and trustworthiness. Future work should focus on these.

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; • **Information systems** → **Collaborative and social computing systems and tools**; **Wikis**; *Graph-based database models.*

KEYWORDS

Wikidata, data quality, literature survey

ACM Reference Format:
Alessandro Piscopo and Elena Simperl. 2019. What we talk about when we talk about Wikidata quality: a literature survey. In *The 15th International Symposium on Open Collaboration (OpenSym '19), August 20–22, 2019, Skövde, Sweden.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3306446.3340822

*Also with British Broadcasting Corporation.

OpenSym '19, August 20–22, 2019, Skövde, Sweden

ACM ISBN 978-1-4503-6319-8/19/08…$15.00
https://doi.org/10.1145/3306446.3340822

1 INTRODUCTION

Wikidata is a relatively young project—it was launched in 2012—but it is already considered by many a success story. It is a collaborative knowledge graph which has already grown up to include more than 55 million data items[1] and has recently overtaken the English Wikipedia as the most edited Wikimedia website.[2]

Knowledge graphs are graph-based knowledge representations which describe real world entities and the relations between them [39]. Numerous knowledge graphs have been developed prior to Wikidata, with notable examples being DBpedia [10] and YAGO [55]. Whereas Wikidata shares a number of features with these, e.g. releasing all data under an open licence, which allows anyone to share and reuse it, it differs with respect to others. Possibly the most significant is its completely collaborative, bottom-up approach to knowledge engineering—a task typically carried out by trained experts [46]. Anyone can edit Wikidata, either registered or anonymously. These features, combined with a large existing community around the Wikimedia ecosystem and the lessons learned from previous knowledge engineering projects, are likely to be among the determinants of Wikidata's success [46].

The growth of Wikidata in terms of size and visibility has already led to its adoption as a knowledge resource for a variety of purposes. For example, already in 2016 the Finnish Broadcasting Company (Yle) started using Wikidata identifiers to annotate content. [3] It is thus not surprising that substantial efforts around Wikidata have been dedicated to its quality and the approaches to evaluate it. Several community initiatives have attempted to gauge quality of the data in the graph, e.g. the item grading scale used in [44]. Data quality was one of the most debated topics at the first WikidataCon, a conference celebrating the 5th year of Wikidata organised by Wikidata Germany in collaboration with the Wikidata community.[4] More recently, a workshop has been dedicated specifically to Wikidata quality, bringing together

[1] https://www.wikidata.org/wiki/Special:Statistics, accessed 30 March 2019.
[2] https://www.wikidata.org/wiki/Wikidata:News, accessed 30 March 2019.
[3] http://wikimedia.fi/2016/04/15/yle-3-wikidata/, accessed 30 March 2019.
[4] https://www.wikidata.org/wiki/Wikidata:WikidataCon_2017, accessed 30 March 2019.

We know where the data comes from

Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach

GABRIEL AMARAL, King's College London, United Kingdom
ALESSANDRO PISCOPO, BBC, United Kingdom
LUCIE-AIMÉE KAFFEE, University of Southampton, United Kingdom
ODINALDO RODRIGUES and ELENA SIMPERL, King's College London, United Kingdom

Wikidata is one of the most important sources of structured data on the web, built by a worldwide community of volunteers. As a secondary source, its contents must be backed by credible references; this is particularly important, as Wikidata explicitly encourages editors to add claims for which there is no broad consensus, as long as they are corroborated by references. Nevertheless, despite this essential link between content and references, Wikidata's ability to systematically assess and assure the quality of its references remains limited. To this end, we carry out a mixed-methods study to determine the relevance, ease of access, and authoritativeness of Wikidata references, at scale and in different languages, using online crowdsourcing, descriptive statistics, and machine learning. Building on previous work of ours, we run a series of microtasks experiments to evaluate a large corpus of references, sampled from Wikidata triples with labels in several languages. We use a consolidated, curated version of the crowdsourced assessments to train several machine learning models to scale up the analysis to the whole of Wikidata. The findings help us ascertain the quality of references in Wikidata and identify common challenges in defining and capturing the quality of user-generated multilingual structured data on the web. We also discuss ongoing editorial practices, which could encourage the use of higher-quality references in a more immediate way. All data and code used in the study are available on GitHub for feedback and further improvement and deployment by the research community.

CCS Concepts: • **Human-centered computing → Collaborative content creation; Asynchronous editors;** • **Information systems → Crowdsourcing; Wikis;**

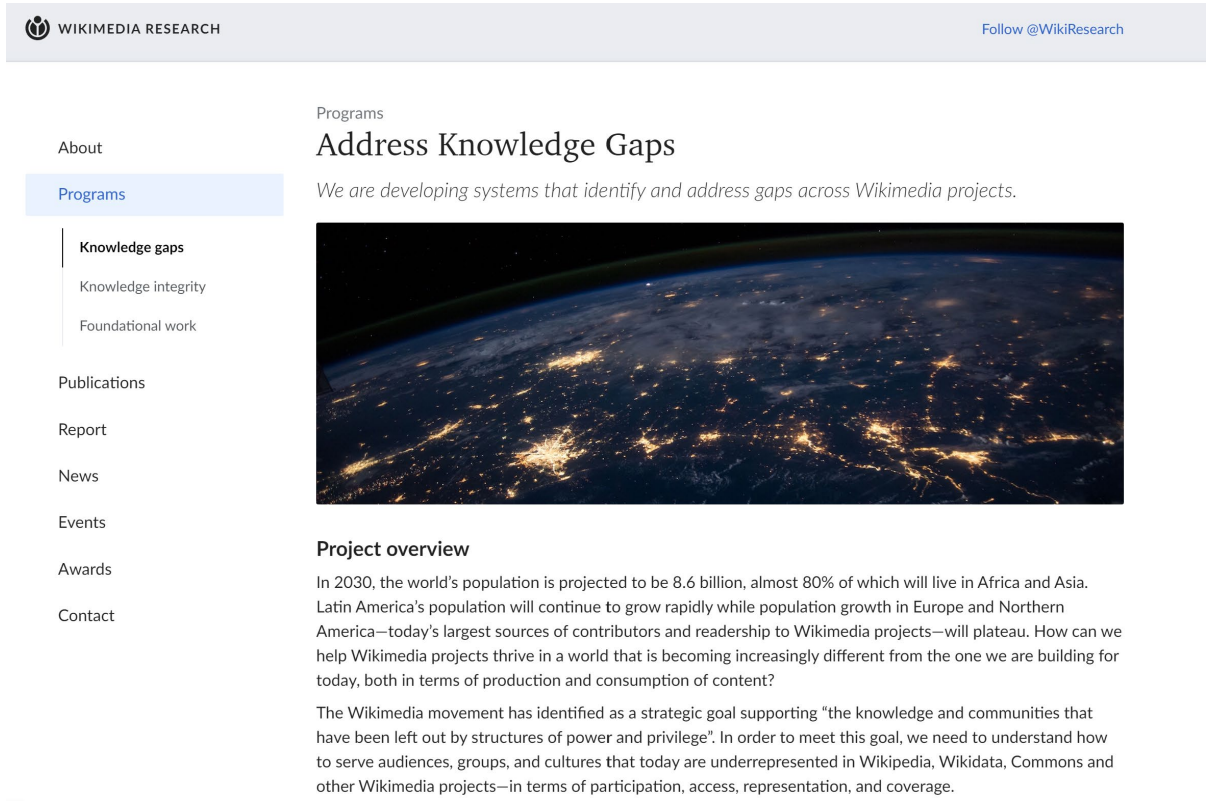
Additional Key Words and Phrases: Wikidata, crowdsourcing, verifiability, data quality, knowledge graphs

ACM Reference format:
Gabriel Amaral, Alessandro Piscopo, Lucie-Aimée Kaffee, Odinaldo Rodrigues, and Elena Simperl. 2021. Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach. *J. Data Inform. Quality* 13, 4, Article 23 (October 2021), 35 pages.
https://doi.org/10.1145/3484828

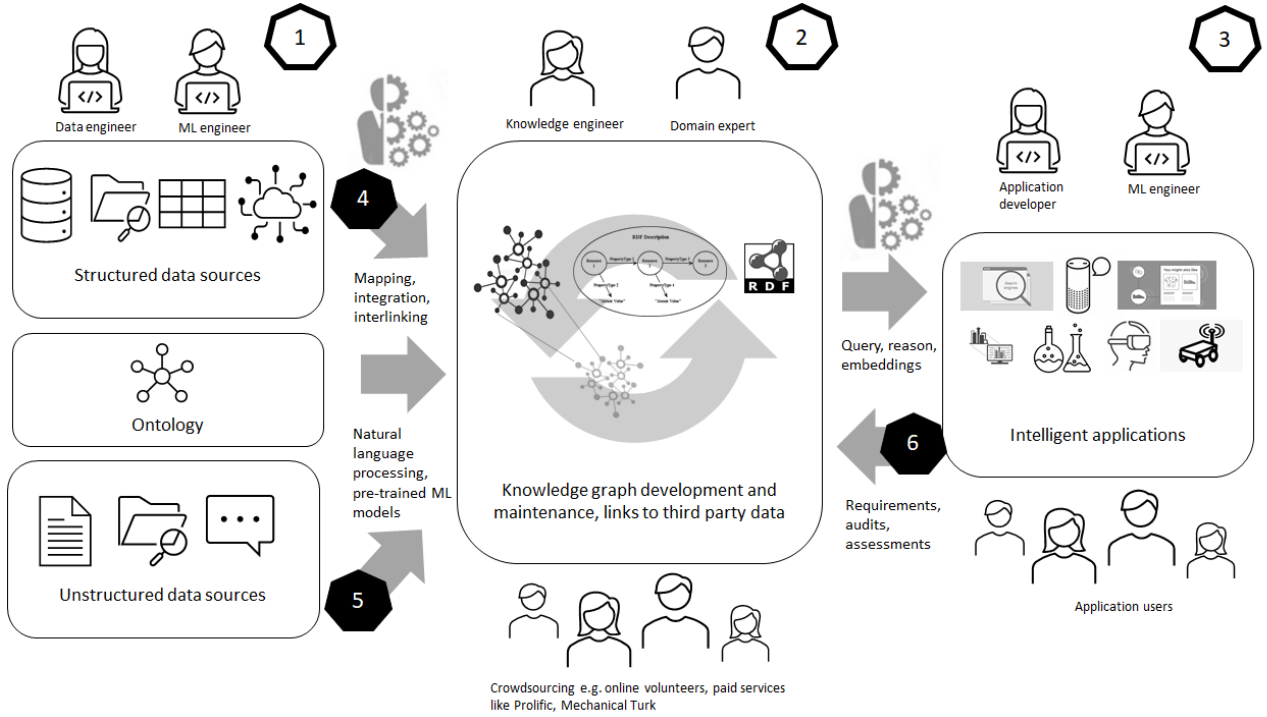
The project leading to this publication has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 812997 (Cleopatra).
Authors' addresses: G. Amaral, O. Rodrigues, and E. Simperl, King's College London, London WC2R 2LS, UK, London, United Kingdom, WC2R 2LS; emails: gabriel.amaral@kcl.ac.uk, odinaldo.rodrigues@kcl.ac.uk, elena.simperl@kcl.ac.uk; A. Piscopo, BBC, London W12 7TQ, London, United Kingdom; email: alessandro.piscopo@bbc.co.uk; L.-A. Kaffee, University of Southampton, Southampton SO17 1BJ, UK, Southampton, United Kingdom; email: kaffee@soton.ac.uk.

1936-1955/2021/10-ART23 $15.00
https://doi.org/10.1145/3484828

We audit our data to make it less biased



We know how the data was created



We know who created the data

Who Models the World? Collaborative Ontology Creation and User Roles in Wikidata

ALESSANDRO PISCOPO, University of Southampton, United Kindgdom
ELENA SIMPERL, University of Southampton, United Kingdom

Wikidata is a collaborative knowledge graph which is central to many academic and industry IT projects. Its users are responsible for maintaining the schema that organises this knowledge into classes, properties, and attributes, which together form the Wikidata 'ontology'. In this paper, we study the relationship between different Wikidata user roles and the quality of the Wikidata ontology. To do so we first propose a framework to evaluate the ontology as it evolves. We then cluster editing activities to identify user roles in monthly time frames. Finally, we explore how each role impacts the ontology. Our analysis shows that the Wikidata ontology has uneven breadth and depth. We identified two user roles: contributors and leaders. The second category is positively associated to ontology depth, with no significant effect on other features. Further work should investigate other dimensions to define user profiles and their influence on the knowledge graph.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; **Wikis**; • **Information systems**;

Additional Key Words and Phrases: Collaborative knowledge engineering, Wikidata, user roles, ontologies

ACM Reference Format:
Alessandro Piscopo and Elena Simperl. 2018. Who Models the World? Collaborative Ontology Creation and User Roles in Wikidata. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 141 (November 2018), 18 pages. https://doi.org/10.1145/3274410

1 INTRODUCTION

Wikidata is the collaborative knowledge graph (KG) initiated by the Wikimedia Foundation in 2012 [45]. Since its launch, its community has grown up to more than 100 thousand registered editors, who have contributed knowledge about around 46 million entities.

Knowledge graphs are a technology used to add context and depth to anything from web search to product recommendations and intelligent assistants. They describe real-world entities, their relationships, and attributes [32]. A KG typically spans across several domains and is built on top of a conceptual schema, or *ontology*, which defines what types of entities (*classes*) are allowed in the graph, alongside the types of *properties* they can have.

Creating KGs is not trivial. It requires a mix of sophisticated machine algorithms and human input. DBpedia, for example, is automatically extracted from Wikipedia via mapping rules created by knowledge engineers and domain experts [24]. YAGO is extracted from Wikipedia as well, but relies on other resources to recognise the type of each entity in the KG [26]. Freebase, a Google project closed in 2015, was built by a team of experts supported by crowdsourcing [32]. While many of these KGs overlap from a content point of view, they differ in their sociotechnical fabric. Each of

Authors' addresses: Alessandro Piscopo, University of Southampton, Southampton, United Kindgdom, alessandro.piscopo@gmail.com; Elena Simperl, University of Southampton, Southampton, United Kingdom, E.Simperl@soton.ac.uk.

2573-0142/2018/11-ART141 $15.00
https://doi.org/10.1145/3274410

What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata

Alessandro Piscopo, Chris Phethean, Elena Simperl

University of Southampton,
Southampton, United Kingdom
[A.Piscopo, C.J.Phethean, E.Simperl]@soton.ac.uk

Abstract. Wikidata is a community-driven knowledge graph which has drawn much attention from researchers and practitioners since its inception in 2012. The large user pool behind this project has been able to produce information spanning over several domains, which is openly released and can be reused to feed any information-based application. Collaborative production processes in Wikidata have not yet been explored. Understanding them is key to prevent potentially harmful community dynamics and ensure the sustainability of the project in the long run. We performed a regression analysis to investigate how the contribution of different types of users, i.e. bots and human editors, registered or anonymous, influences outcome quality in Wikidata. Moreover, we looked at the effects of tenure and interest diversity among registered users. Our findings show that a balanced contribution of bots and human editors positively influence outcome quality, whereas higher numbers of anonymous edits may hinder performance. Tenure and interest diversity within groups also lead to higher quality. These results may be helpful to identify and address groups that are likely to underperform in Wikidata. Further work should analyse in detail the respective contributions of bots and registered users.

Keywords: Wikidata, collaborative knowledge graphs, group composition

1 Introduction

Peer production systems have been experimented with successfully in several fields. Wikipedia is probably the most well-known example, but the efforts of communities of users are behind diverse projects, including open source software (e.g. Apache or Linux), database management systems (e.g. PostgreSQL), or question-answer sites (e.g. Stack Overflow). Wikidata is a recent addition to this already large list. It is a community-driven knowledge graph started by the Wikimedia foundation in October 2012. Since its inception, it has gathered a user pool of around 100 thousand registered users, who are able to add facts about more than 24 million entities. Because of these and other features, Wikidata has drawn the attention of researchers and practitioners alike.

Knowledge Graphs (KGs) are large collections of structured data, encoded as terms describing entities and the relationships existing between them [26]. KGs are important as they provide data that can be processed by machines to create new, tailored

We know how the data is used

Semantic Web 13 (2022) 163–194
DOI 10.3233/SW-210431
IOS Press

Using natural language generation to bootstrap missing Wikipedia articles: A human-centric perspective

Lucie-Aimée Kaffee [a,*], Pavlos Vougiouklis [b,**] and Elena Simperl [c]
[a] *School of Electronics and Computer Science, University of Southampton, UK*
E-mail: kaffee@soton.ac.uk
[b] *Huawei Technologies, UK*
E-mail: pavlos.vougiouklis@huawei.com
[c] *King's College London, UK*
E-mail: elena.simperl@kcl.ac.uk

Editor: Philipp Cimiano, Universität Bielefeld, Germany
Solicited reviews: John Bateman, Bremen University, Germany; Leo Wanner, Pompeu Fabra University, Spain; Denny Vrandecic, Wikimedia Foundation, USA

Abstract. Nowadays natural language generation (NLG) is used in everything from news reporting and chatbots to social media management. Recent advances in machine learning have made it possible to train NLG systems that seek to achieve human-level performance in text writing and summarisation. In this paper, we propose such a system in the context of Wikipedia and evaluate it with Wikipedia readers and editors. Our solution builds upon the ArticlePlaceholder, a tool used in 14 under-resourced Wikipedia language versions, which displays structured data from the Wikidata knowledge base on empty Wikipedia pages. We train a neural network to generate an introductory sentence from the Wikidata triples shown by the ArticlePlaceholder, and explore how Wikipedia users engage with it. The evaluation, which includes an automatic, a judgement-based, and a task-based component, shows that the summary sentences score well in terms of perceived fluency and appropriateness for Wikipedia, and can help editors bootstrap new articles. It also hints at several potential implications of using NLG solutions in Wikipedia at large, including content quality, trust in technology, and algorithmic transparency.

Keywords: Wikipedia, Wikidata, ArticlePlaceholder, multilingual, natural language generation, neural networks

1. Introduction

Wikipedia is available in 301 languages, but its content is unevenly distributed [31]. Language versions with less coverage than e.g. English Wikipedia face multiple challenges: fewer editors means less quality control, making that particular Wikipedia less attractive for readers in that language, which in turn makes it more difficult to recruit new editors from among the readers.

*Corresponding author. E-mail: kaffee@soton.ac.uk.
**Work done while working at University of Southampton.

- Empirical knowledge engineering
- Knowledge graphs and language models
- Regulatory compliance