Deep Knowledge Group

Data Sources, Quality Assurance and Accuracy Overview

AI Industry Ecosystem Governance and Policy Dashboard

November 2024

Table of Contents

1. General Overview

1.1. **Deep Knowledge Group's** primary competitive edge lies in its analytical approach to sourcing data. In pursuit of this objective, we employ a diverse range of algorithms for collecting and processing data. Our team of experts is continuously innovating, creating new algorithms, and enhancing existing ones to further bolster our capabilities.

1.2. Our data collection process employs a range of sophisticated algorithms designed to efficiently gather information from diverse sources. These algorithms are tailored to specific data types and sources, enabling comprehensive data retrieval across various domains.

1.3. By utilising advanced parsing techniques and leveraging **APIs, NLP** and **web scraping** technologies, we ensure the accuracy and reliability of the collected data. These algorithms play a crucial role in enabling us to access, extract, and organise vast amounts of data, empowering us to derive valuable insights and make informed decisions.

2. Search Engine (Sources)

2.1. **Bing Search Parser:** The primary competitive edge lies in its analytical approach to sourcing data. In pursuit of this objective, we employ a diverse range of algorithms for collecting and processing data. Our team of experts is continuously innovating, creating new algorithms, and enhancing existing ones to further bolster our capabilities.

Bing Search is a publicly available search engine that indexes a broad range of information, including data on companies, financial metrics, and related business news. It provides access to company profiles, stock prices, and other financial data aggregated from various sources, including third-party financial databases and news outlets.

Coverage: 8/10. Bing Search covers a broad spectrum of publicly available information about companies, including news, stock prices, and basic financial data. However, its coverage may not be as exhaustive as specialized financial platforms, particularly for in-depth analysis or lesser-known companies.

Data quality: 7/10. The quality of the data depends on the sources indexed by Bing. While it can provide accurate and timely information, the data might lack the consistency and reliability of dedicated financial databases. There is also variability in the credibility of sources.

Structured data: 4/10. Data retrieved through Bing Search can vary significantly in format, as it aggregates information from multiple sources with different structures. This makes parsing and automated extraction more complex compared to platforms with more standardized data formats.

Accessibility: 10/10. Bing Search is freely accessible and widely available, making it a convenient tool for gathering information about companies without any cost barriers. This is advantageous for general research and quick lookups.

Overall: 8/10.

2.2. **Yahoo Finance**. Yahoo Finance is a publicly available financial news and data platform that provides comprehensive information about companies, stocks, markets, and financial news. It covers a broad spectrum of financial data, including stock prices, historical performance, company profiles, financial statements, and news articles. Yahoo Finance aggregates data from various sources, making it a widely used tool for investors and financial analysts.

Coverage: 8/10. Covers a wide range of financial data, including stock prices, market indices, company financials, news, and analysis. However, it primarily focuses on publicly traded companies.

Data Quality: 7/10. Data is generally reliable and comes from reputable sources, but there can be occasional discrepancies, especially in real-time data. News articles are aggregated from various sources, which may vary in quality.

Structured Data: 8/10. Data is well-structured, making it relatively easy to parse for financial analysis. Information is categorized by company, financials, market data, etc.

Accessibility: 9/10. Free access to most data, with some advanced features available through a paid subscription (Yahoo Finance Premium). Free access is sufficient for most basic analysis and data extraction.

Overall: 8.5/10.

Datapoints: Stock prices, historical performance, company profiles, financial statements (income statement, balance sheet, cash flow statement), market capitalization, P/E ratio, EPS, news articles, analyst recommendations, insider transactions. Information about company executives, board members, industry sector, and competitors.

2.3. **Google Finance**. Google Finance is a publicly available platform that provides real-time and historical data on financial markets and companies. It aggregates information from various financial data providers, offering insights into stock prices, financial statements, news, and other market data.

Coverage: 7/10. Google Finance covers a wide range of publicly traded companies, offering stock prices, financial metrics, and related news. However, it is less comprehensive compared to dedicated financial databases like Bloomberg or FactSet, particularly for smaller companies or international markets.

Data quality: 7/10. The data provided is generally reliable for basic financial information, but it lacks the depth and granularity found in specialized financial databases. Some data, like analyst ratings or in-depth financial analysis, may be limited or not as frequently updated.

Structured data: 6/10. Google Finance offers data in a structured format that is accessible via the web interface. However, it lacks API support for automated data extraction, making large-scale parsing more challenging compared to other financial data platforms.

Accessibility: 9/10. The platform is freely accessible, offering a user-friendly interface for viewing basic financial data, which is a significant advantage for casual users and small-scale investors.

Overall: 7.2/10.

Datapoints: Company name, stock symbol, market price (real-time and historical), market capitalization, P/E ratio, dividend yield. Financial statements, including revenue, net income, EPS (Earnings Per Share). Recent news articles related to the company, provided by various news outlets. Industry and sector classification of the company. Key metrics such as 52-week high/low, volume, and average trading volume. Limited analyst ratings and recommendations.

2.4. **LinkedIn**. LinkedIn is a professional networking platform widely used for accessing information about individuals, companies, and industry trends. It serves as a hub for professional profiles, company pages, job postings, and industry news, making it a valuable tool for both professionals and organizations.

Coverage: 9/10. LinkedIn provides extensive coverage of professional and company-related information. It includes detailed profiles of professionals, company pages, job postings, and industry-specific groups. However, its coverage is primarily focused on professional data, which may limit its utility for broader business metrics or financial analysis. It is highly effective for networking and accessing career-related information but may not provide in-depth company financials or stock market data.

Data Quality: 8/10. The quality of data on LinkedIn is generally high, as most information is self-reported by users or managed by companies. Profiles are often updated regularly, ensuring that data is relatively current. However, the accuracy and completeness of the information can vary depending on the user's diligence in maintaining their profile. Some data might be exaggerated or incomplete, especially in less regulated regions or for lesser-known professionals.

Structured Data: 7/10. LinkedIn provides structured data through well-defined fields in user profiles and company pages, such as experience, education, and skills. This makes it easier to parse and extract data for automated processes compared to unstructured sources. However, variations in how users fill out their profiles can introduce inconsistencies, and LinkedIn's limitations on data export can restrict large-scale data analysis.

Accessibility: 7/10. LinkedIn offers both free and premium access. While basic information is accessible for free, premium features, such as advanced search filters, InMail, and detailed analytics, require a subscription. This tiered access can be a barrier for those seeking comprehensive data without incurring additional costs.

Overall: 8/10. LinkedIn is a powerful tool for accessing professional and company information, particularly for networking and career development. Its high-quality, structured data and broad coverage make it an essential resource for professionals and organizations alike, though some limitations in data completeness and accessibility should be noted.

2.5. **Companies Websites**. Companies Websites are the primary source of direct information about a company's operations, products, services, and leadership. They offer the most authentic and up-to-date details, often including press releases, financial reports, and executive biographies, making them a vital resource for obtaining accurate data directly from the source.

Coverage: 9/10. Company websites provide comprehensive coverage of their own activities, including detailed information on their products, services, corporate structure, and latest news. However, they may lack third-party analysis or broader market context that other sources might provide.

Data Quality: 9/10. The data from company websites is generally reliable, as it is directly published by the company. However, the data can be biased towards positive portrayals, and sometimes less critical or negative information might be underrepresented.

Structured Data: 6/10. The information on company websites is typically well-structured for human consumption but can vary significantly in format and complexity. This variability can make automated data extraction challenging, particularly when comparing across different companies.

Accessibility: 10/10. Company websites are freely accessible and provide immediate and unrestricted access to a wealth of company-specific information, making them an indispensable tool for detailed company research.

Overall: 8.5/10. Company websites are a highly valuable resource for obtaining direct, high-quality information about a company, though their lack of third-party perspectives and potential data structuring challenges slightly lower their overall utility for some research purposes.

2.6. **Semantic Scholar**. Semantic Scholar is a publicly available academic search engine that indexes a vast range of scholarly articles, conference papers, and other academic publications. It provides access to research papers across various fields of study, including computer science, biology, medicine, and social sciences, aggregated from multiple academic publishers, research repositories, and open-access platforms.

Coverage: 9/10. Semantic Scholar covers a wide array of academic disciplines, offering access to millions of research papers and articles. Its extensive database includes both well-known journals and lesser-known conference proceedings. However, some specialized fields or recent publications may be underrepresented compared to databases like Scopus or Web of Science.

Data quality: 8/10. The data quality on Semantic Scholar is generally high, as it sources content from reputable academic publishers and research institutions. However, the quality and depth of the metadata (e.g., abstracts, citations) may vary, especially for older or less-cited publications.

Structured data: 7/10. Data on Semantic Scholar is relatively well-structured, with comprehensive metadata that includes author information, citations, and related works. This structure supports advanced search features and citation analysis. However, variability in the completeness of metadata can occasionally make automated extraction and parsing more challenging.

Accessibility: 10/10. Semantic Scholar is freely accessible and provides an intuitive user interface, making it easy for researchers, students, and professionals to search for and access academic papers. It also offers features like citation graphs and paper recommendations, enhancing the overall user experience without cost barriers.

Overall: 8.5/10. Semantic Scholar is a valuable resource for academic research, offering extensive coverage and high-quality data across various disciplines. While it excels in accessibility and breadth, its structured data and coverage of niche fields could be improved to match specialized academic databases.

3. Data Filling, Quality Assurance and Accuracy

AI Governance Policymakers			
Datapoint	**Source**	**Data Quality**	**Data Filling**
name	Bing Answer + NER, research	100%	100%
Unique Industry	classificator (GTP promt respons)	85%	90%
Unique Sector	classificator (GTP promt respons)	85%	90%
Super_Industrie	classificator (GTP promt respons)	85%	88%
description	Bing Answer + NER	90%	89%
email	Bing Answer + NER	90%	73%
website	Bing Answer + NER	90%	83%
logo	Logo Parser, Logo Validator	90%	98%
Domain	Bing Answer + NER	90%	79%
headquarters	Bing Answer + NER	100%	72%
company invested in	Bing Answer + NER	90%	22%
founders	Bing Answer + NER	90%	24%
employees_number	Bing Answer + NER	85%	57%
revenue	Bing Answer + NER	90%	15%

founded_date	Bing Answer + NER	85%	56%
founded_year	Bing Answer + NER	85%	48%
linkedin	Bing Answer + NER	95%	68%
facebook	Bing Answer + NER	90%	17%
x	Bing Answer + NER	90%	58%
company type internal	Bing Answer + NER	90%	45%

AI Governance Hubs

Datapoint	Source	Data Quality	Data Filling
name	Bing Answer + NER	100%	100%
Type	classificator (GTP promt respons)	90%	100%
description	Bing Answer + NER	90%	85%
logo	Logo Parser, Logo Validator	90%	98%
website	Bing Answer + NER	90%	71%
email	Bing Answer + NER	90%	49%
headquarters	Bing Answer + NER	100%	75%
employees_number	Bing Answer + NER	90%	24%
linkedin	Bing Answer + NER	95%	19%
facebook	Bing Answer + NER	90%	19%
x	Bing Answer + NER	90%	33%

AI Governance Organizations			
Datapoint	**Source**	**Data Quality**	**Data Filling**
name	Bing Answer + NER, research	100%	100%
type	classificator (GTP promt respons)	85%	100%
description	Bing Answer + NER, Description parser	90%	100%
email	Bing Answer + NER	90%	17%
website	Bing Answer + NER	90%	100%
domain	Bing Answer + NER	90%	100%
industry	classificator (GTP promt respons)	85%	98%
admin_tags	Bing Answer + NER	100%	98%
founded_date	Bing Answer + NER	90%	27%
headquarters	Bing Answer + NER	100%	100%
employees_number	Bing Answer + NER	90%	50%
logo	Bing Answer + NER	90%	99%
estimated_revenue	Bing Answer + NER	85%	60%
x	Bing Answer + NER	90%	20%

AI Governance Leaders			
Datapoint	**Source**	**Data Quality**	**Data Filling**
name	Bing Answer + NER	100%	100%
logo	Photo Parser, Photo Validator	90%	100%
type	classificator (GTP promt respons)	85%	100%
email	Bing Answer + NER	90%	72%

linkedin	Bing Answer + NER	95%	70%
x	Bing Answer + NER, Apollo Parser, Linkedin Parser	90%	51%
country	Bing Answer + NER, Apollo Parser, Linkedin Parser	85%	99%
company	Bing Answer + NER, Apollo Parser, Linkedin Parser	100%	98%
description	Bing Answer + NER, Apollo Parser, Linkedin Parser	90%	100%

4. Overall Data Filling, Quality Assurance and Accuracy

4.1. **AI Governance Organizations**. Average Data Quality Assurance and Accuracy is 89%. Average Data Filling Assurance and Accuracy is 64%.

4.2. **AI Governance Policymakers**. Average Data Quality Assurance and Accuracy is 90%. Average Data Filling Assurance and Accuracy is 64%.

4.3. **AI Governance Leaders**. Average Data Quality Assurance and Accuracy is 89%. Average Data Filling Assurance and Accuracy is 68%.

4.4. **AI Governance Hubs** Average Data Quality Assurance and Accuracy is 85%. Average Data Filling Assurance and Accuracy is 75%.

5. Data Sources

5.1. Bing Answer + NER (Named Entity Recognition) is a structured process for gathering specific data entities from various online sources mainly from Google and Bing. This methodology combines search engine capabilities and advanced entity recognition techniques to extract and verify relevant information. Bing Answer + NER provides a robust framework for collecting and verifying a wide range of data entities. By integrating search engine capabilities with advanced NER methods, this approach ensures comprehensive and reliable data extraction across various domains.

5.2. The process of collecting logos takes place by making a request to Google and Bing search engines and obtaining the URL of the image according to the entered information, downloading the image of the logo from the appropriate source or resource using the URL of the image, and checking whether the image corresponds to the required formats (eg JPEG, PNG, etc.). Validation of process images using NNtool validation classification.

5.3. For Leaders Databases we also used Apollo API Parser and Linkedin Parser.

6. Data Classification

6.1. At DKG, we specialize in advanced text preprocessing solutions that optimize data analysis processes for businesses across industries. Our expertise lies in refining raw text data, extracting key objects, and ensuring data integrity through thorough cleaning and structuring. Text pre-processing is an important step in the data analysis pipeline, as it lays the foundation for accurate and insightful conclusions. By cleansing the data and extracting relevant objects, we reduce noise and improve the quality of information, thereby contributing to more effective decision-making and strategic planning for our clients.

6.2. Our text pre-processing services offer several advantages:

6.2.1. Improved data quality: We eliminate inconsistencies, errors, and redundancies in text, resulting in cleaner, more reliable datasets.

6.2.2. Improved object recognition: With advanced algorithms and natural language processing techniques, we excel at identifying and extracting key objects such as dates, names, locations, organizations, and more, enriching the dataset with valuable contextual information.

6.3. **Optimized analysis**: By organizing and structuring text according to predefined criteria, we simplify the data analysis process, saving time and resources while maximizing the depth of insights gained from the data.

6.4. **Increased accuracy**: Our rigorous pre-processing methods ensure that data is properly formatted and standardized, minimizing the risk of misinterpretation and errors during analysis.

6.5. **Entity recognition and extraction**: To extract named entities such as dates, names, locations, organizations, etc., we use models trained specifically for named entity recognition (NER). These models are typically based on deep learning architectures such as bidirectional LSTMs (long short-term memory networks) or transformer-based models such as BERTs (bidirectional encoder representations of transformers).

6.6. **GTP prompt classification**: Regarding the conditions of classification, it is assumed that the framework of the relation of classes, sub-classes, etc. is formed, the classification is one-way or multi-way. For each specific task of classification, the optimal GTP prompt is formed, and the corresponding filters are formed to form a structured answer.

7. Entity Recognition

7.1. Deep Knowledge Group (DKG) has spearheaded the development of advanced Entity Recognition algorithms, revolutionizing our ability to extract validated data from vast quantities of unstructured sources. These algorithms represent a breakthrough in data processing, enabling us to efficiently identify and validate entities within text and images, thereby enhancing the quality and reliability of our dataset.

7.2. Our Entity Recognition algorithms are designed to tackle the inherent challenges posed by unstructured data, such as news articles, reports, social media posts, and other textual sources. Leveraging state-of-the-art natural language processing (NLP) techniques, machine learning models, and semantic analysis, these algorithms can accurately identify and extract entities, including entities such as organizations, people, locations, events, and more. By discerning context, relationships, and relevance, our algorithms ensure that extracted entities are not only identified but also validated for accuracy and reliability.

7.3. In addition to text, our Entity Recognition algorithms extend to image data, enabling us to validate and extract entities from visual content with precision and efficiency. Through advanced image recognition and pattern recognition techniques, we can identify objects, logos, landmarks, and other visual elements, enriching our dataset with valuable insights derived from multimedia sources.

7.4. Furthermore, our data enrichment and classification validation processes are augmented by querying search engines, leveraging their vast repositories of indexed information to corroborate and validate extracted entities. By cross-referencing data obtained from multiple sources, we enhance the accuracy and completeness of our dataset, ensuring that only validated information is incorporated into our database.

7.5. A key aspect of our Entity Recognition framework is the integration of a data validator, which provides continuous evaluation and feedback to the system. This feedback loop enables our algorithms to learn and adapt over time, incorporating new patterns, insights, and corrections to improve their accuracy and performance. By leveraging machine learning techniques, the system can dynamically adjust its algorithms based on real-world feedback, ensuring that the database remains clean, updated, and reflective of the most current information available.

8. Advantages of Utilising APIs, NLP and Web Scraping Technologies

8.1. **Efficiency**: By employing APIs, NLP, and web scraping, we optimise the data collection process, reducing manual efforts and enhancing efficiency. These technologies automate data retrieval tasks, allowing our team to focus on analysis and decision-making rather than data acquisition.

8.2. **Accuracy**: The use of advanced parsing techniques and structured data access through APIs ensures the accuracy and reliability of the collected data. NLP algorithms enhance data understanding and interpretation, enabling us to extract meaningful insights with precision.

8.3. **Comprehensive Data Coverage**: By combining APIs, NLP, and web scraping technologies, we achieve comprehensive data coverage across various sources and domains. This approach enables us to gather data from diverse sources, including structured databases and unstructured web content, facilitating holistic analysis and informed decision-making.

8.4. **Real-Time Access**: APIs provide real-time access to data, ensuring that our analyses are based on the latest information available. This real-time data access enhances the timeliness and relevance of our insights, enabling agile responses to changing market conditions and emerging trends.

8.5. **Flexibility**: The flexibility of web scraping allows us to extract data from any website, regardless of whether an API is available. This flexibility enables us to adapt to evolving data sources and requirements, ensuring continuity and adaptability in our data collection efforts.

8.6. **Innovation**: Deep Knowledge Group's commitment to innovation drives continuous improvement in our algorithms and technologies. By staying at the forefront of advancements in APIs, NLP, and web scraping, we continually enhance our capabilities and maintain a competitive edge in data analytics and decision support.

9. Data Quality Control, Accuracy and Lawfulness

9.1. Data quality control and accuracy are paramount in Deep Knowledge Group's (DKG) analytical endeavours, particularly given the diverse array of data sources we rely on. Our commitment to excellence begins with the careful selection of source material, drawn from a multitude of reputable outlets, including news articles, press releases, organisation websites, scientific publications, patent databases, government websites, and various open sources encompassing a wide spectrum of information.

9.2. Recognising the inherent variability and potential biases across different data sources, we prioritise independent data from external sources outside our organisation. This approach not only enriches our dataset with diverse perspectives but also serves as a foundational principle for ensuring the reliability and credibility of our analyses. By cross-validating each data point across multiple sources whenever possible, we mitigate the risk of inaccuracies and inconsistencies, thereby bolstering the overall quality and trustworthiness of our database.

9.3. The process of data validation and quality control at DKG is meticulous and thorough. We employ a multifaceted approach that involves rigorous scrutiny of each data point, leveraging advanced algorithms, machine learning techniques, and human expertise to assess its validity, relevance, and reliability. Through automated checks and manual verification procedures, we identify and weed out low-quality, outdated, or questionable data, ensuring that only the most accurate and up-to-date information is retained in our database.

9.4. Furthermore, we recognise that maintaining data accuracy is an ongoing endeavour that requires constant vigilance and proactive measures. As such, we have implemented robust protocols for continuous monitoring, validation, and updates, enabling us to promptly address any discrepancies, errors, or changes in the underlying data sources. Whether it's through regular audits, real-time alerts, or feedback mechanisms from users and stakeholders, we remain vigilant in our efforts to uphold the integrity and accuracy of our data.

9.5. In addition to stringent quality control measures, we also prioritise transparency and accountability in our data management practices. We provide clear documentation regarding the sources, methodologies, and assumptions underlying our analyses, allowing users to assess the reliability and credibility of the insights derived from our data. Moreover, we actively engage with stakeholders to solicit feedback, address concerns, and refine our data collection and validation processes, thereby fostering a culture of continuous improvement and trust.

9.6. From a **legal** standpoint, we rely on certain legal exclusions provided by various legislation in the UK, EU and US concerning data collection and mining. In particular, based on the jurisdiction and aims of data collection, its further use, we rely on temporary reproduction in the UK and EU, the text and data mining exception in the EU, and fair use in the US.

9.7. At each stage, specific legal criteria receive and filter each individual data source:

9.7.1. Access to the content: Initial access to the content is subject to stringent legal scrutiny to ensure compliance with relevant regulations and standards;

9.7.2. Extraction and copying of content: The process of extracting and copying content is conducted within the bounds of applicable laws, with careful attention paid to intellectual property rights and data protection considerations;

9.7.3. Text and data mining: Text and data mining activities are conducted in accordance with the provisions and exceptions afforded by relevant legislation, balancing the imperative of innovation with the imperative of safeguarding privacy and intellectual property rights;

9.7.4. Data storage: The storage of mined data is executed in compliance with established data protection regulations, encompassing measures aimed at safeguarding the confidentiality, integrity, and availability of the data.

9.8. Furthermore, our adherence to privacy regulations is unwavering, with a particular focus on the EU General Data Protection Regulation, the UK General Data Protection Regulation, and the Data Protection Act 2018. With each occurrence of personal data collection, we establish legal grounds and implement a comprehensive array of supplementary measures. For instance, we provide users with expedited mechanisms such as quick contact forms for the deletion or rectification of personal data, ensuring that individuals retain agency over their personal information within the parameters of the law.

9.9. We diligently observe and abide by AI regulations promulgated by regulatory bodies such as the European Commission and the UK government. Specifically, we adhere to the EU

Regulation on Artificial Intelligence (AI Act) and the UK's AI Regulatory Framework, which provide comprehensive frameworks for the ethical and lawful deployment of AI technologies, including those involved in data collection and mining endeavours.

9.10. Our AI regulatory compliance efforts encompass several key facets:

9.11. **Transparency and Accountability**: We prioritise transparency in AI-driven data collection processes, ensuring that individuals are informed about the nature, scope, and implications of data collection activities. Moreover, we uphold principles of accountability, assuming responsibility for the ethical and lawful utilisation of AI technologies in data mining endeavours.

9.12. **Fairness and Non-discrimination**: In accordance with regulatory mandates, we uphold principles of fairness and non-discrimination in AI-driven data collection practices. We employ measures to mitigate biases and ensure equitable treatment across diverse demographic groups, thereby fostering trust and inclusivity in our data collection initiatives.

9.13. **Data Protection and Privacy**: As indicated above we prioritise the safeguarding of individuals' privacy rights and personal data integrity throughout the data collection lifecycle.

9.14. **Ethical Considerations**: We remain attuned to ethical considerations inherent in AI-driven data collection and mining activities, striving to uphold principles of beneficence, autonomy, and justice in our practices. Our ethical framework guides decision-making processes, ensuring alignment with societal values and ethical norms.