Towards InnoGraph: A Knowledge Graph for AI Innovation

M.Besher Massri
Jozef Stefan Institute
Jozef Stefan International Postgraduate School
Ljbuljana, Slovenia
besher.massri@ijs.si

Blerina Spahiu
University of Milano-Bicocca
Milan, Italy
blerina.spahiu@unimib.it

Marko Grobelnik
Jozef Stefan Institute
Ljbuljana, Slovenia
marko.grobelnik@ijs.si

Vladimir Alexiev
Ontotext (Sirma AI)
Sofia, Bulgaria
vladimir.alexiev@ontotext.com

Matteo Palmonari
University of Milano-Bicocca
Milan, Italy
matteo.palmonari@unimib.it

Dumitru Roman
SINTEF AS
Oslo, Norway
dumitru.roman@sintef.no

ABSTRACT

Researchers seeking to comprehend the state-of-the-art innovations in a particular field of study must examine recent patents and scientific articles in that domain. Innovation ecosystems consist of interconnected information about entities such as researchers, institutions, projects, products, and technologies. However, representing such information in a machine-readable format is challenging because concepts like "knowledge" are not easily represented. Nonetheless, even a partial representation of innovation ecosystems provides valuable insights. Therefore, representing innovation ecosystems as knowledge graphs (KGs) would enable advanced data analysis and generate new insights. To this end, we propose InnoGraph, a framework that integrates multiple heterogeneous data sources to build a Knowledge Graph of the worldwide AI innovation ecosystem.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; *Information extraction*; **Artificial intelligence**; **Machine learning**; • **Information systems** → **Semantic web description languages**; • **Social and professional topics** → **Computing / technology policy**; • **Applied computing** → *Digital libraries and archives.*

KEYWORDS

artificial intelligence, innovation, innovation ecosystem, knowledge graph, science knowledge graph, economics knowledge graph

1 INTRODUCTION

Innovation is one of the most important assets when it comes to improving the competitiveness of any organization and a major driver of economic and societal transformations. Having the capability to access information about cutting-edge innovation and monitor its ongoing development is a tremendously valuable asset for different families of stakeholders, from policymakers and investors to individual companies and social scientists in different subjects. Artificial intelligence (AI) is a key driver of innovation, enabling the development of new technologies and applications [15]. AI has the ability to process large amounts of data and make decisions based on that data, leading to significant advancements in industries such as healthcare [30], finance [2], transportation [1], manufacturing [6], image generation [22], etc..

The potential of AI to drive innovation is vast and will continue to impact various industries and fields in the coming years. This potential is also associated with two essential features: it is evolving at an unprecedented pace, and it is associated with several ethical and societal concerns [9]. For these reasons, monitoring its evolution is both difficult and valuable.

Recent technological advancements and the abundance of open data sources related to innovation enable more quantitative analyses and tool-supported discovery; to identify skill discrepancy between education and jobs [4], mapping the evolution path of technology [20], and identify knowledge and innovation spillover [16], among other usages.

In the last decades, the research community has been proposing various solutions to make science-related work (e.g., articles and their metadata) machine-readable [8]. Knowledge graphs (KGs) have been proposed as valuable solutions

because of their simple and effective structure. Entities – represented as nodes, are classified, interlinked by specific relations, and possibly described by additional attributes [13]. As a consequence, several KGs representing scholarly data have been published [8, 14, 26, 29].

A similar approach can be applied to support the monitoring of AI innovation, to exploit several advantages of KGs: (i) *Interconnectivity*: a structured and interconnected representation supports applications for easy exploration and discovery of new relationships between entities and a comprehensive and interconnected view of data, yielding to a better understanding of the AI innovation landscape and more informed decisions, e.g., about investments and partnerships; (ii) *Better data analysis*: the mix of graph-based data and text support advanced

data analysis, providing valuable insights into the AI innovation process and supporting data-driven decision-making; (iii) *Enhanced search*: the rich context attached to the data supports improved search functionalities, helping users find relevant information.

Numerous datasets and KGs related to AI innovation exist (see Section 3), but they focus mainly on publications and patents and include simple explicit relations (e.g., collaboration and citation relations). The coverage and quality of topic-wise classification varies across sources and it is not specific enough to track AI innovations. The economic aspects of AI innovation (e.g. companies,

investments, acquisitions, coverage across media) and other work products (datasets, software, ML models) are not covered. All of this makes it difficult for policymakers, investors, companies and researchers to gain useful insights and have a better understanding of *how*, *when*, and *where* AI innovation happens, and what future directions it may take.

This position paper introduces InnoGraph, a knowledge graph of the worldwide AI innovation ecosystem. The main goal is to build and maintain an "AI Innovation Knowledge Graph" (InnoGraph) by semantic integration of heterogeneous data sources that cover the different phases of the AI innovation lifecycle. The aim is to interconnect, model, and understand the relevant aspects of the global AI innovation ecosystem from the inception of the ideas (e.g., articles and patents) to their potential realization in business (e.g., investments), up to the uptake (e.g., media coverage) and decision-making at the research policy level. This KG can be used to efficiently manage and analyze data within the AI innovation ecosystem, allowing for a better understanding of trends, relationships, and patterns in the data, leading to improved decision-making and innovation in AI.

The main contributions of this paper include: 1) introduce InnoGraph and the different aspects of AI innovation and the need to integrate heterogeneous sources to cover its evolution (Section 2); 2) collect and analyze state-of-the-art related to scientific and innovation KGs and identify the gaps that need to be fulfilled (Section 3); 3) provide an overview of the architecture of the proposed solution, the technical challenges that must be solved to build a valuable monitoring tool, and the use cases for downstream usage (Section 4).

2 INNOGRAPH AND AI INNOVATION

Due to the breadth of the global innovation ecosystem, we plan to focus on the narrower field of AI spreading horizontally across many fields of research and technology.

Papers, patents, and grant applications have become less innovative compared to earlier work and are less likely to link different areas of knowledge, which are both factors that contribute to innovation [27]. Additionally, determining the most effective actions for an innovation ecosystem is challenging. Identifying the crucial factors within the ecosystem and predicting the effectiveness of specific measures are complex tasks and cannot be predicted easily.

The idea of InnoGraph is born around the concept of supporting the appropriate stakeholders in various stages of the ecosystem by delivering insights from the AI innovation lifecycle. Papers and patents play an increasingly important role in innovation and economic performance [24]. However, recent papers and patents do less to push science and technology in new directions [28]. The time between scientific discovery and its recognition with a Nobel Prize has grown, indicating a decline in the quality and impact of recent innovations compared to those of the past [23]. Thus, InnoGraph aims to fill this gap by (i) connecting more datasets rather than only patents and research papers (e.g., StackExchange and GitHub); (ii) providing insights at different stages of AI innovation (e.g., by modeling adequately the AI innovation ecosystem and using analytical methods that provide relevant insights to managers of different stages of the innovation) and (iii) to support different actors (e.g., policymakers for the preparation of new legislation, new funding programs, etc.).

The purpose of InnoGraph is to provide data-driven support to each stage of AI innovation which could be understood as a "journey of an AI innovation", from inception to implementation. The journey of innovation can be seen as a composition non strictly in the presented order of the following stages: (1) an innovation typically appears in the academic world; (2) projects are started around the innovation; (3) the innovation gets possibly patented; (4) companies are established around the innovation; (5) companies get investments, possibly in several rounds; (6) investments influence the job market; (7) market reacts to the quality and possible impact of the innovation; (8) public and expert perception gets formed; (9) media starts publishing about the innovation and companies; (10) educational institutions integrate innovation in their curricula; (11) policymakers regulate the innovation; and (12) to close the cycle, funding agencies create new funding opportunities to create space for follow-up innovations.

Each stage of the "journey of innovation" has its own set of stakeholders, including scientists and policymakers, who play their own unique roles. The "journey" can last from a few years, up to decades. Innovations can sometimes fail to materialize due to various reasons (like lack of potential or even politics). Currently, the innovation ecosystem is generally studied and analyzed only at a local and fragmented level, with each stage being considered individually rather than taking a holistic view of the entire lifecycle.

Therefore, the aim is to create a holistic model (in a form of an evolving knowledge graph) of the AI innovation ecosystem, identify interdependencies and influences between the stages, and deliver an operational prototype serving the stakeholders involved in the innovation process. In particular, the vision is to help decision-makers at various stages to properly manage innovations with potential and to avoid mistakes in the process.

3 RELATED WORK ON INNOVATION KNOWLEDGE GRAPHS

The aim of innovation knowledge graphs is to map out the relationships between different actors, resources, and processes involved in the innovation ecosystem. Knowledge graphs transform data into knowledge, whether they are general in nature, like Wikidata[1], focused on geographic information, like Geonames[2], or more specialized, like the Springer SciGraph [40] in the field of research.

Work on building innovation KGs can be divided considering different aspects: (i) data input, (ii) scope, and (iii) production automation.

Regarding data input, innovative KGs can be divided into (i) KGs that focus on a single data source, e.g., research papers like OpenAlex [29] or (ii) KGs that combines multiple sources of input, e.g., [38] that combines data from Github, Stack Overflow, and Wikidata.

Considering the scope of the innovative KGs, several applications are identified: scientific knowledge discovery (MAG [26], OpenAlex [29], ORKG [14], etc.), technology monitoring (SciGraph for publication trends and collaboration networks [40], SoftKG for software

[1]https://www.wikidata.org
[2]https://www.geonames.org

development monitoring [38]), personalized recommendations (proprietary KG, such as Facebook or Google KG), etc.

With respect to the automation process, innovative KGs are built by (i) manual or semi-automatic approaches (e.g., ORKG [14]) and (ii) fully automated (CS-KG [8], PKG [41]) that implement a fully automatic pipeline for the extraction of entities and predicates to populate their respective KGs.

In addition, there is a recent study focused on building a KG for Innovation Ecosystems [35]. Such KG leverages data from four domains: patents; projects; articles, talent, and funding; and finally organizations. Despite the fact that the above project is the closest to the one introduced in this paper, to the best of our knowledge such KG has remained bound in a limited number of domains and data sources. Moreover, there are no other papers or repositories describing the progress status of such a KG rather than the latest paper in 2020.

In the state-of-the-art, InnoGraph is positioned in the union of the above categories. In fact, as described in Section 4.2, InnoGraph leverages data from different sources, including the most prominent KGs of both categories. However, it is not limited to the above-cited works as it leverages data ranging from research into the industry with coverage of funding and investments, social media, and technical forums. To the best of our knowledge, no KG has been proposed so far that focuses on AI innovation ecosystems and has similar coverage of innovation stages.

4 THE INNOGRAPH KNOWLEDGE GRAPH

4.1 Technical architecture

As shown in Figure 1, the InnoGraph KG consists of five parts:

- *Data sources*: a set of temporal corpora providing covering multiple stages of the AI innovation cycle. These documents are represented in a shared representation to facilitate further analysis (see details in Sections 4.2 and 6),
- *Ontologies and KGs*: existing ontologies, KGs and classifications providing the necessary information about the topics (see details in Section 6).
- *Analytical modules*: modular components to handle various data enrichment tasks, including data cleaning, data linking, and relation extraction.
- *The knowledge graph*: embedding the AI ecosystem.
- *Use cases and applications*: built on top of the KG, to serve multiple stakeholders (See details in Section 4.4).

4.2 Data sources

In order to cover the whole lifecycle of AI innovation, InnoGraph considers using a large number of datasets. On a conceptual level, these data sources can be categorized into the following aspects:

- *R&D*: research, patents
- *Media*: both mainstream (news) and social media
- *Software development*: technical forums, collaboration platforms, etc.
- *Funding and investments*: both public and private
- *Companies*: representing the commercial aspect of the AI innovation cycle and tightly coupled with investments
- *Policy data*: initiatives, regulations and economical indicators

More details on datasets under each aspect can be found in Section 6.

In addition to structured data, we are looking into science popularization websites and blogs, such as the Stanford AI Index Report[3] and AI Vibrancy Tool[4], LifeArchitect[5], State of AI Report 2022[6], 2022 AI Tech Trends Report[7] by Future Today Institute, Natural Language Processing Progress[8], etc. Furthermore, we are collecting an extensive Zotero bibliography[9].

4.3 Technical Challenges

To build such a KG, certain challenges and requirements need to be handled, as further specified below.

Harmonizing and ingesting a huge amount of data. Developing infrastructure for obtaining, parsing, filtering/selecting, and storing the different data sources in a shared representation.

Building a holistic taxonomy. Our objective is to develop a comprehensive taxonomy of AI concepts encompassing a wide range of scientific, technological, business, and policy-related topics, including both established and emerging areas within the field of AI and its applications. To accomplish this, we plan to merge several taxonomies, tag sets, and subject thesauri. (see Section 6).

Linking/matching named entities across data sources. Linking especially organizations (institutions and companies) names, by leveraging information from existing databases about organizations (e.g., Wikidata, ROR) as well as organization metadata in each of the data sources.

Semantic tagging. Annotating works/documents (possibly multilingual) with topics and other entities (e.g., products, organizations) to facilitate search over the corpora.

Classification of works/documents with topics. Going beyond semantic annotation and keyword matching, by building classifiers to tag documents with topics when needed.

Construction and continuous update of the dynamic harmonized KG. Not only building but also updating the dynamic harmonized KG, mirroring the exponential growth of AI. In specific, handling the concept and entity life cycles and the emergence of new concepts/entities.

Implement analytic tasks for different use cases and applications. See Section 4.4 for details.

4.4 Use cases

InnoGraph holds vast potential and can be utilized to get insights about various steps of the "journey of an innovation" lifecycle. There are two categories where InnoGraph finds application: (i) based on the final scope and (ii) based on the audience.

Some of the most prominent applications based on the final scope include, but are not limited to: **InnoGraph Explorer:** a platform for

[3] https://aiindex.stanford.edu/report
[4] https://aiindex.stanford.edu/vibrancy
[5] https://lifearchitect.ai
[6] https://www.stateof.ai
[7] https://futuretodayinstitute.com/trends
[8] https://nlpprogress.com
[9] https://www.zotero.org/groups/4918562/innograph/library

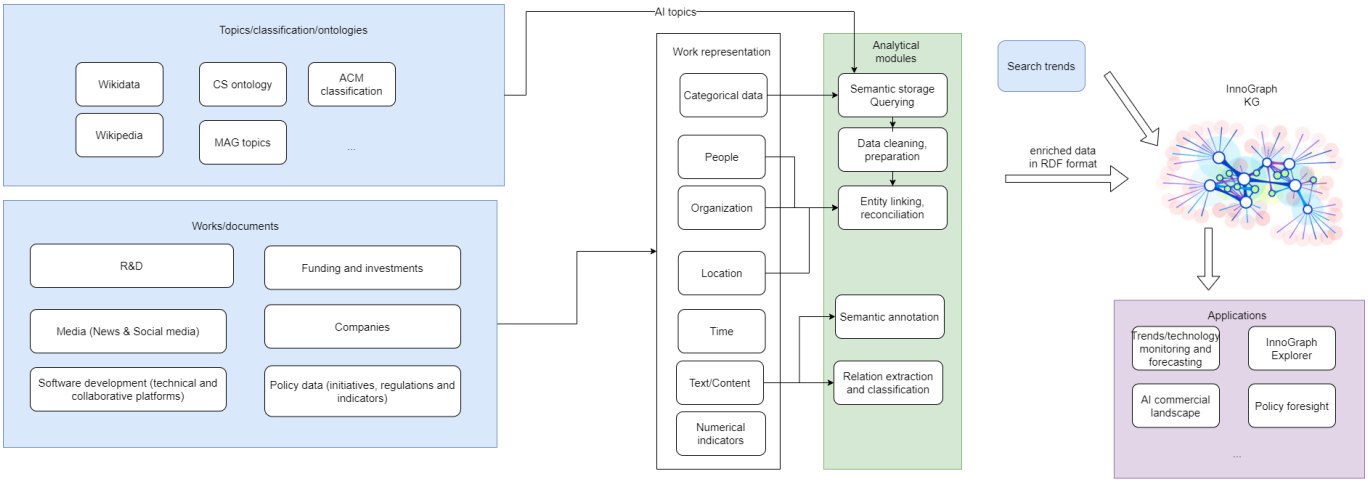

Figure 1: Proposed architecture of InnoGraph

visual exploration and SPARQL endpoint for navigation, exploring, and querying the InnoGraph KG.

Trends and Technology Forecasting: a tool for observing the status of the latest technologies in AI; monitoring their innovation lifecycle across multiple stages of development, identifying emerging technologies, and providing data-driven evidence.

AI Commercial Landscape: a tool for monitoring the commercial landscape of AI, categorizing and analyzing companies that develop or use artificial intelligence and machine learning technologies. Identify main application sectors and geographical regions that drive the development and implementation of AI.

Investment Recommendations: based on analysis of AI applications in various industries, industry convergence, strategic gaps, discovering promising startups, and recommending investments and acquisitions.

Policy Foresight Tool: a tool to assist policymakers in assessing the AI landscape in their countries, answering complex policy questions, comparing with other countries, and guiding research investment.

4.5 End users

The use cases can serve multiple types of users or target audiences:

AI Policymakers: responsible for developing and implementing policies and regulations related to AI in a given country or jurisdiction. Their main goal is to promote the responsible and ethical use of AI in society, and to ensure that AI technology is developed and applied in a way that benefits the community. They may work closely with other government agencies, AI experts, and stakeholders to develop and implement AI policies.

Investors: provide financial resources to organizations or projects that are involved in the development or application of AI technologies. Investors may be individuals, venture capital firms, or other financial institutions, and their main goal is to generate a return on their investment through the success of the AI projects or companies they invest in. Investors who are interested in AI may conduct due diligence and risk assessment to evaluate the potential of different AI projects or companies, and they may provide support and guidance to the teams or founders of these projects to help them achieve their goals.

Researchers: engaged in the process of studying and advancing the field of AI. They may work in a variety of fields and disciplines that are related to AI, such as computer science, engineering, mathematics, neuroscience, and cognitive psychology. Such researchers may work in an academic institution, government agency, nonprofit organization, or private company, and they often use a variety of methods and tools, such as machine learning algorithms, simulations, and experiments, to study and develop AI technologies.

Companies: startups and established companies seek to adopt and incorporate AI into their operations and offerings to gain a competitive advantage, improve efficiency, and increase revenue. Such organizations are interested in acquiring up-to-date knowledge, resources, tools, and techniques related to AI and monitoring AI-related policies and regulations.

5 SUMMARY AND OUTLOOK

Knowledge graphs simplify information access needed by different users on the "journey of an innovation" life cycle. They offer enriched data, thereby significantly enhancing the benefits gained in decision-making. In this paper, we introduced InnoGraph which represents innovation ecosystems as knowledge graphs (KGs) and enables the generation of new insights and would allow advanced data analysis. We presented the technical architecture and the set of data sources under consideration.

As a future work, we plan to implement the full pipeline for InnoGraph generation and test its application in different use cases, as part of the enRichMyData project.

ACKNOWLEDGMENTS

The work on InnoGraph is partially funded by the projects enRichMyData (HE 101070284), Graph-Massivizer (HE 101093202), DataCloud (H2020 101016835), and BigDataMine (NFR 309691). The original work is inspired by a partnership between OECD and JSI, on the OECD AI Policy Observatory.

REFERENCES

[1] Rusul Abduljabbar, Hussein Dia, Sohani Liyanage, and Saeed Asadi Bagloee. 2019. Applications of Artificial Intelligence in Transport: An Overview. *Sustainability* 11, 1 (Jan. 2019), 189. https://doi.org/10.3390/su11010189 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Arvind Ashta and Heinz Herrmann. 2021. Artificial intelligence and fintech: An overview of opportunities and risks for banking, investments, and microfinance. *Strategic Change* 30, 3 (2021), 211–222. https://doi.org/10.1002/jsc.2404 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jsc.2404.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[4] Katy Börner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans. 2018. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences* 115, 50 (Dec. 2018), 12630–12637. https://doi.org/10.1073/pnas.1804247115 Publisher: Proceedings of the National Academy of Sciences.

[5] Arie Cattan, Sophie Johnson, Daniel Weld, Ido Dagan, Iz Beltagy, Doug Downey, and Tom Hope. 2021. SciCo: Hierarchical Cross-Document Coreference for Scientific Concepts. https://doi.org/10.48550/arXiv.2104.08809 arXiv:2104.08809 [cs].

[6] George Chryssolouris, Kosmas Alexopoulos, and Zoi Arkouli. 2023. *A Perspective on Artificial Intelligence in Manufacturing*. Studies in Systems, Decision and Control, Vol. 436. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-21828-6

[7] Martin Czygan, Helge Holzmann, and Bryan Newbold. 2021. Refcat: The Internet Archive Scholar Citation Graph. http://arxiv.org/abs/2110.06595 arXiv:2110.06595 [cs].

[8] Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, Enrico Motta, and Harald Sack. 2020. AI-KG: An Automatically Generated Knowledge Graph of Artificial Intelligence. In *The Semantic Web – ISWC 2020 (Lecture Notes in Computer Science)*, Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal (Eds.). Springer International Publishing, Cham, 127–143. https://doi.org/10.1007/978-3-030-62466-8_9

[9] Amitai Etzioni and Oren Etzioni. 2017. Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* 21, 4 (Dec. 2017), 403–418. https://doi.org/10.1007/s10892-017-9252-2

[10] Michael Färber. 2019. The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In *The Semantic Web – ISWC 2019 (Lecture Notes in Computer Science)*, Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon (Eds.). Springer International Publishing, Cham, 113–129. https://doi.org/10.1007/978-3-030-30796-7_8

[11] Georgios Gousios and D. Spinellis. 2012. GHTorrent: Github's data from a firehose. In *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*. IEEE, Zurich, 12–21. https://doi.org/10.1109/MSR.2012.6224294

[12] Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. 2020. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies* 1, 1 (Feb. 2020), 414–427. https://doi.org/10.1162/qss_a_00022

[13] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *Comput. Surveys* 54, 4 (July 2021), 71:1–71:37. https://doi.org/10.1145/3447772

[14] Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19)*. Association for Computing Machinery, New York, NY, USA, 243–246. https://doi.org/10.1145/3360901.3364435

[15] Chinmay Kakatkar, Volker Bilgram, and Johann Füller. 2020. Innovation analytics: Leveraging artificial intelligence in the innovation process. *Business Horizons* 63, 2 (March 2020), 171–181. https://doi.org/10.1016/j.bushor.2019.10.006

[16] Pantelis Koutroumpis, Aija Leiponen, and Llewellyn D. W. Thomas. 2020. Digital instruments as invention machines. *Commun. ACM* 64, 1 (Dec. 2020), 70–78. https://doi.org/10.1145/3377476

[17] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. Association for Computing Machinery, New York, NY, USA, 107–110. https://doi.org/10.1145/2567948.2577024

[18] Michael Ley. 2002. The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *String Processing and Information Retrieval*, Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Alberto H. F. Laender, and Arlindo L. Oliveira (Eds.). Vol. 2476. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–10. https://doi.org/10.1007/3-540-45735-6_1 Series Title: Lecture Notes in Computer Science.

[19] Michael Ley. 2009. DBLP: some lessons learned. *Proceedings of the VLDB Endowment* 2, 2 (Aug. 2009), 1493–1500. https://doi.org/10.14778/1687553.1687577

[20] Huailan Liu, Zhiwang Chen, Jie Tang, Yuan Zhou, and Sheng Liu. 2020. Mapping the technology evolution path: a novel model for dynamic topic detection and tracking. *Scientometrics* 125, 3 (Dec. 2020), 2043–2090. https://doi.org/10.1007/s11192-020-03700-5

[21] Manghi, Paolo, Atzori, Claudio, Bardi, Alessia, Baglioni, Miriam, Schirrwagen, Jochen, Dimitropoulos, Harry, La Bruzzo, Sandro, Foufoulas, Ioannis, Mannocci, Andrea, Horst, Marek, Czerniak, Andreas, Iatropoulou, Katerina, Kokogiannaki, Argiro, De Bonis, Michele, Artini, Michele, Lempesis, Antonis, Ioannidis, Alexandros, Manola, Natalia, Principe, Pedro, Vergoulis, Thanasis, Chatzopoulos, Serafeim, and Pierrakos, Dimitris. 2022. OpenAIRE Research Graph Dump. https://doi.org/10.5281/ZENODO.3516917 Version Number: 5.0.0 Type: dataset.

[22] Marian Mazzone and Ahmed Elgammal. 2019. Art, Creativity, and the Potential of Artificial Intelligence. *Arts* 8, 1 (March 2019), 26. https://doi.org/10.3390/arts8010026 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[23] Patrick Collison Nielsen, Michael. 2018. Science Is Getting Less Bang for Its Buck. https://www.theatlantic.com/science/archive/2018/11/diminishing-returns-science/575665/ Section: Science.

[24] OECD. 2004. *Patents and Innovation: Trends and Policy Challenges*. Organisation for Economic Co-operation and Development, Paris. https://doi.org/10.1787/9789264026728-en

[25] José Luis Ortega. 2014. *Academic search engines: a quantitative outlook*. Chandos Publishing, Amsterdam ; Boston. OCLC: ocn879582408.

[26] José Luis Ortega. 2014. Microsoft Academic Search: the multi-object engine. In *Academic Search Engines*. Elsevier, 71–107. https://doi.org/10.1533/9781780634722.71

[27] Mikko Packalen and Jay Bhattacharya. 2020. NIH funding and the pursuit of edge science. *Proceedings of the National Academy of Sciences* 117, 22 (June 2020), 12011–12016. https://doi.org/10.1073/pnas.1910160117 Publisher: Proceedings of the National Academy of Sciences.

[28] Michael Park, Erin Leahey, and Russell J. Funk. 2023. Papers and patents are becoming less disruptive over time. *Nature* 613, 7942 (Jan. 2023), 138–144. https://doi.org/10.1038/s41586-022-05543-x Number: 7942 Publisher: Nature Publishing Group.

[29] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. https://doi.org/10.48550/arXiv.2205.01833 arXiv:2205.01833 [cs].

[30] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. 2022. AI in health and medicine. *Nature Medicine* 28, 1 (Jan. 2022), 31–38. https://doi.org/10.1038/s41591-021-01614-0 Number: 1 Publisher: Nature Publishing Group.

[31] Richard Sever, Ted Roeder, Samantha Hindle, Linda Sussman, Kevin-John Black, Janet Argentine, Wayne Manos, and John R. Inglis. 2019. *bioRxiv: the preprint server for biology*. preprint. Scientific Communication and Education. https://doi.org/10.1101/833400

[32] Andrey Tagarev, Laura Toloşi, and Vladimir Alexiev. 2017. Domain-Specific Modeling: A Food and Drink Gazetteer. In *Transactions on Computational Collective Intelligence XXVI*, Ngoc Thanh Nguyen, Ryszard Kowalczyk, Alexandre Miguel Pinto, and Jorge Cardoso (Eds.). Springer International Publishing, Cham, 186–209. https://doi.org/10.1007/978-3-319-59268-8_9

[33] Jie Tang, Duo Zhang, and Limin Yao. 2007. Social Network Extraction of Academic Researchers. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, Omaha, NE, USA, 292–301. https://doi.org/10.1109/ICDM.2007.30

[34] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*. Association for Computing Machinery, New York, NY, USA, 990–998. https://doi.org/10.1145/1401890.1402008

[35] Alberto Tejero, Victor Rodriguez-Doncel, and Ivan Pau. 2020. Knowledge Graphs for Innovation Ecosystems. https://doi.org/10.48550/arXiv.2001.08615 arXiv:2001.08615 [cs, econ, q-fin].

[36] Kleanthis Vichos, Michele De Bonis, Ilias Kanellos, Serafeim Chatzopoulos, Claudio Atzori, Natalia Manola, Paolo Manghi, and Thanasis Vergoulis. 2022. A Preliminary Assessment of the Article Deduplication Algorithm Used for the OpenAIRE Research Graph. In *Proceedings of the 18th Italian Research Conference on Digital Libraries (CEUR Workshop Proceedings, Vol. 3160)*, Giorgio Maria Di Nunzio, Beatrice Portelli, Domenico Redavid, and Gianmaria Silvello (Eds.). CEUR, Padua, Italy. https://ceur-ws.org/Vol-3160/#short16 ISSN: 1613-0073.

[37] Alex D. Wade. 2022. The Semantic Scholar Academic Graph (S2AG). In *Companion Proceedings of the Web Conference 2022*. ACM, Virtual Event, Lyon France, 739–739. https://doi.org/10.1145/3487553.3527147

[38] Jihu Wang, Xueliang Shi, Lin Cheng, Kun Zhang, and Yuliang Shi. 2020. SoftKG: Building A Software Development Knowledge Graph through Wikipedia Taxonomy. In *2020 IEEE World Congress on Services (SERVICES)*. 151–156. https://doi.org/10.1109/SERVICES48979.2020.00042 ISSN: 2642-939X.

[39] Jian Xu, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F. Rousseau, Xin Li, Weijia Xu, Vetle I. Torvik, Yi Bu, Chongyan Chen, Islam Akef Ebeid, Daifeng Li, and Ying Ding. 2020. Building a PubMed knowledge graph. *Scientific Data* 7, 1 (June 2020), 205. https://doi.org/10.1038/s41597-020-0543-2 Number: 1 Publisher: Nature Publishing Group.

[40] Yuchen Yan and Chong Chen. 2022. SciGraph: A Knowledge Graph Constructed by Function and Topic Annotation of Scientific Papers. (June 2022).

[41] Yu Yang, Jiangxu Lin, Xiaolian Zhang, and Meng Wang. 2022. PKG: A Personal Knowledge Graph for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3334–3338. https://doi.org/10.1145/3477495.3531671

6 ANNEX: INNOGRAPH DATASETS

In order to cover the whole lifecycle of AI innovation, InnoGraph considers using a large number of datasets covering various kinds of entities as described below. In many cases we also mention important kinds of identifiers used with those entities.

Topics. Key to finding and filtering relevant items from all datasets. Therefore, at the beginning, we focus on developing a holistic taxonomy of AI topics (more than 10k), with identifiers and links to these datasets, aliases, descriptions, and hierarchy. We leverage topics from all datasets to be used by InnoGraph and more, e.g. (i) StackExchange "tag info" pages (e.g., Machine Learning on Stack Overflow[10] and on AI Exchange[11]) cover numerous relevant topics and provide names, descriptions, aliases, and related sub-topics. (ii) The Mathematics Subject Classification has only about 15 topics devoted to AI (code 68T[12]) but paper classification is curated and of high quality (so we can be certain that the 109671 AI papers listed in zbMATH[13] are indeed related to AI), and the topics have lateral (see also) links. (iii) arXiv (a prominent preprint archive) has about 20 relevant sub-archives, of which 10 are directly related to AI, e.g. CS: Artificial Intelligence, CS: Machine Learning, Stats: Machine Learning, Computer Vision and Pattern Recognition, etc.

We also collect then prune Wikipedia Categories (similar to the approaches in [32, 38]) and harvest pages (topics) and descriptions, to ensure comprehensive coverage, to organize topics hierarchically, and to reach out to Wikidata where we find aliases and additional external identifiers for these topics.

In addition to AI, we target other hot innovation topics such as Augmented Reality, 3D Printing, Blockchain.

Academic Datasets. Cover scientific publications, citation relations between publications, authors, institutions (affiliations), venues (journals, conferences), and fields of study (topics). In addition to commercial academic databases (e.g. Elsevier ScienceDirect, Clarivate Web of Science), open academic databases have become increasingly prominent in the last 10 years [25].

Google Scholar is perhaps the best well-known academic database, but it is closed in the sense that it doesn't offer a data dump or API.

CrossRef[14] [12] is the clearing-house for DOIs, the most important identifiers for academic works (although many works do not have DOIs). CrossRef offers various APIs and currently has 143M works (100M journal articles, 23M books and chapters, 7.7M conference papers, 455k standards, 2.7M datasets), 111k journals, 97k conferences. It also has a sizable number of links: 59M works with references, a total of 1.4B citation links, and 2.9M works with funder info. Thus it is a central resource that sets a "yard stick" for comparing the scale/completeness of academic data. However, it doesn't offer as many value-added services as some other datasets.

Microsoft Academic Graph (MAG) [26] was a very prominent resource, covering 280M works including papers and patents, and a machine-learned Fields of Science taxonomy of 750k entries. A KG (linked data) version is also available [10]. However, MAG was shut down at the end of 2021.

OpenAlex[15] [29] by OurResearch (makers of UnPaywall) started with MAG data and is being updated continuously from a variety of sources (CrossRef, arXiv, institutional repositories, etc). It includes 248M works (125M journal articles, 20M book chapters, 9M conference papers, 4.75M preprints, presentations, 2.87M datasets, unlike MAG it includes no patents), 227k venues (journals and conferences), 7.2k publishers 65k grants, 100.5M authors (non deduplicated), 108k institutions. It has a comprehensive and performant API, MySQL dumps, ranking (of works, topics, venues, authors, institutions), and connects to important identifiers and database: ORCID for researchers (only 3.5% have it), ROR for orgs (96% have it), DOAJ and ISSN for journals, Unpaywall for full texts, General Index for N-grams of works, Pubmed and Pubmed Central for medical publications. It has 65k topics inherited from MAG and fully linked to Wikidata, but the quality of paper classification is low. Other quality problems include gaps in some important venues, and that authors not deduplicated.

Semantic Scholar[16] [37] by the Allen Institute for AI (AI2) is similar in size to OpenAlex and includes 205M publications, 121M authors, 2.5B citation edges. It has good semantic processing (classification, influential citations, TLDR summaries, author profiles). It offers APIs and dumps, but requires agreement/approval.

Arnet Miner (AMiner)[17] [33, 34] by China Knowledge Centre for Engineering Sciences and Technology, Tsinghua University, and other Chinese researchers is one of the largest academic KGs with 333M works, 85M researchers, 1.1B citation links. It includes a machine-learned taxonomy of 8.8M concepts and great semantic/ML processing and tools: classification, author profiles, find an expert/reviewer, talent migration, tracing tree, ranking, most influential papers in a field, etc. It includes over 50 open datasets and ML models for building science KGs. Further, it is the most important gateway to Chinese research, in particular AI, e.g. see the "AI development monthly reports" for China[18]. It offers dumps/APIs, but requires establishing a cooperation agreement.

OpenAIRE[19] [21] is the EU initiative to build a research graph. It includes 174M works: 149M publications (16M research datasets,

[10] https://stackoverflow.com/tags/machine-learning/info
[11] https://ai.stackexchange.com/tags/machine-learning/info
[12] https://zbmath.org/classification/?q=cc:68T
[13] https://zbmath.org/?q=cc:68T
[14] https://www.crossref.org
[15] https://openalex.org
[16] https://www.semanticscholar.org
[17] https://www.aminer.cn
[18] https://www.aminer.cn/research_report/articlelist
[19] https://graph.openaire.eu

333k software items); 2.9M grants/projects (1.85M US NIH, 520k US NSF, 140k UK, 83k CH, 74k PT, 66.4k EU including all EU Framework Programmes since at least FP5, etc); 182k organizations. It aggregates data from 123.6k data sources (3.7M Zenodo which is a self-publishing archive ran by Cern and funded by the EU), 3.7M Euro PubMed Central 2.4M Hal-Diderot French preprint archive, etc). Despite significant work on deduplicating works coming from multiple sources ([36], and further publications[20]) some quality problems still remain. Further, authors are not separate objects but are embedded in publications and are not deduplicated.

Archive Scholar[21] [7], also known as FatCat[22] is a newer academic dataset by archive.org, the "internet memory of humanity". It includes 100M works, strong considerations for open access and permanent archiving, and is in active development. It incorporates the General Index: N-grams of 107M papers that can be used in lieu of full-text, but includes no topics.

In addition to the above huge academic KGs, there are numerous niche datasets (per-sector, per-publisher, etc).

arXiv is an established preprint server and it seems that the newest impactful ML papers and models are first posted here; as stated in arxlive[23] "arXiv: a window on innovation... is full of ground-breaking research". On a given day, we saw 150 new ML publications posted here.

DBLP[24] [18, 19] is an academic dataset devoted to Computer Science that includes over 6.5M publications (520k added per year), 5.8k conferences/workshops, 1.8k journals, 2.9M authors (165k manually checked/deduplicated, 102k ORCID). It is the most comprehensive archive for CS and has high-quality metadata, but no topics.

PubMed[25] and the PubMed KG [39] are the most important academic KGs in life sciences, and bioRxiv[26] [31] is another important source in that domain.

Publisher KGs include Springer Nature SciGraph[27], IOS Press KG[28].

The reason we need to deal with multiple academic datasets is that coverage and quality for the domain of interest vary significantly between them. We checked the coverage of some venues of interest (journals and conferences mentioning "Semantic") across some datasets:

- DBLP: 101 venues (including former conference names)
- FatCat: 59 venues
- OpenAlex: 38 venues (the Semantic Scholar dataset is counted as 1 venue)

We also checked for CEUR-WS, which is an important series of open self-published workshops:

- CEUR-WS: 3329 volumes
- DBLP: 112 venues
- OpenAlex: 5 unmerged records
- FatCat: 1 venue, maybe confuses the series with one volume Vol-1516

We also counted works in the following venues:

- ISWC:
  - DBLP: 5170 works, 16 constituent workshops
  - FatCat: 3000 works
  - OpenAlex: 971 + OM@ISWC 111 + Posters, Demos, Industry Tracks 99 = 1081 works
- ESWC:
  - DBLP: 2565 works
  - FatCat: 574 works
  - OpenAlex: "Extended" (current name) 82 + "European" (former name) 49 = 131 works
- CEUR-WS
  - DBLP: 49778 works
  - OpenAlex: 1155 works in 5 venues
  - FatCat: 140 works

Patents. Patenting activity reflects a significant part of the progress of innovation. Google Patents[29] is the largest dataset of patents, accessible through Google BigQuery. It covers applications and grants, families (same invention protected in several jurisdictions), patents translated from many languages, inventors (people) and assignees (companies). Includes detailed patent data, CPC topics, and vector embeddings computed from CPC and significant abstract key-phrases.

The Collaborative Patent Classification (CPC)[30] is used by both EPO and USPTO and has about 200k highly detailed topics. It is available from EPO as linked data[31], and class Y10S706/00 is devoted to AI. Google finds 11.7k patents classified as AI[32], but there are many more ML-related patents that are not so classified (e.g. over 100k matching "convolutional neural network")

Lens.org is an Australian science KG that has 252M academic works, 127M patent records (applications, grants, etc), 700k topics (from MAG), 4.9M keywords, 31.2M orgs, 2.1M funders; and strong citation information: 2B academic citations, 386M patent citations, and 26M citations from patents to academic works (a unique Lens feature that allows tracing the transition from academic work to commercialization).

Next we describe other work results that are no less important than publications and patents.

Datasets and Benchmarks. AI cannot exist without data, so datasets are an important ingredient of AI work. Datasets and associated benchmarks often drive the progress of AI state-of-the-art. Popular datasets (e.g. the MNIST handwritten digits dataset) are important topics for discovering AI works. Some rich sources of ML tasks, benchmarks and datasets are Kaggle[33] and OpenML[34]. The EU is investing heavily into Data Spaces, which are initiatives for sharing of commercial data, while providing access control, usage control and sovereignty guarantees. Dedicated initiatives like DataCite aim to increase the exposure and FAIR description of Datasets. Some of

[20] https://graph.openaire.eu/docs/publications
[21] https://scholar.archive.org
[22] https://fatcat.wiki
[23] https://arxlive.org
[24] https://dblp.org
[25] https://pubmed.ncbi.nlm.nih.gov
[26] https://www.biorxiv.org
[27] https://www.springernature.com/gp/researchers/scigraph
[28] http://ld.iospress.nl
[29] https://patents.google.com
[30] https://www.cooperativepatentclassification.org
[31] https://epo.org/linked-data
[32] https://patents.google.com/?q=Y10S706/00
[33] https://kaggle.com
[34] https://www.openml.org

the academic KGs described above include info about Datasets, and self-publishing archives like Zenodo and FigShare make it easy to persist a dataset and get a DOI.

ML Models. Models are amongst the most important results of ML. Some popular models (e.g. BERT and its variations, GPT-3) are important topics that can be used to select works from datasets. The largest sources of models is Hugging Face[35], "the github of ML models". It offers easy deployment of models, code and apps using the models ("Spaces"), related datasets. Another similar site is Replicate[36]. Some academic KGs offer models that can be especially useful for InnoGraph e.g. SciBERT [3] and SciCo [5] of Semantic Scholar.

Software. Github is the largest "social coding" site and As of the end-2022, it has 330M repositories, 94M developers, 4M "organizations" (which are groups of repositories, often related to a development project: not necessarily constituted organizations). Github has tags (topics) and a rich data model and APIs. But it does not offer a dump, there are API limits, and the identification of many Github developers is problematic since they don't add personal info. GHtorrent [11] offers bulk access, but is irregularly updated (latest update in May 2021) and since May 2018 does not provide personal data.

People: Researchers, Developers, Innovators. People make innovation happen, and it is important for InnoGraph to identify person records and track international collaborations and the transition of people from academia to industry (and back). Each of the large datasets include person information, but it varies in presence of identifiers, level and quality of deduplication, and richness of personal profile info. The most important researcher identifier is ORCID, but unfortunately it is still sparsely populated in publication datasets. Sources like Crunchbase include rich personal information of founders and other key people in companies, including education and employment.

Organizations, Companies. Innovation happens in groups, and the interactions of research and commercial organizations is especially important: collaborations, spinoffs/startups, joint ventures. ROR[37] is the most important global identifier of orgs who do research and has 102k orgs. EU CORDIS Participant ID is important for EU research, and has an estimated 70k orgs (an import of only Horizon 2020 participants to Wikidata[38] has 33k). Apart from academic orgs, there are over 300M companies in the world. There are multiple commercial providers of company data; we have experience with Dun and Bradstreet, Refinitiv, S&P Global CapitalIQ, OpenCorporates, national trade registers, etc. Crunchbase provides info about 2M companies, including a good selection of innovative companies and startups.

Technical Forums. Developer discussions on technical questions can help determine important topics/keywords, and some of the prominent developers in these domains. StackExchange[39] is the largest family of forums (Reddit is an alternative, but it doesn't have such large concentration of tech discussions). Stack Overflow is the most popular "exchange" (23M questions, of which e.g. 55k on ML), followed by "exchanges" like Cross Validated (statistics, machine learning, data analysis, data mining), Data Science, Artificial Intelligence, etc.

Research Funding. Grants are an important source of info for government high-tech funding. The resulting projects are important clusters that allow to discover relevant work (through paper Acknowledgements/Funding sections). Some of the academic KGs provide grant/funding info, but the coverage is uneven. EU CORDIS provides detailed structured data about all grants under the EU Framework Programmes. OpenAIRE has a concentration of grants, not only for EU but also US, UK, etc. Dimensions.ai has info about nearly 6M grants that is available through Springer Nature SciGraph[40].

Funding Rounds, Acquisitions, Initial Public Offerings. Commercial investment in startups and other companies are important indicators of innovation. Crunchbase provides info about all these 3 kinds of investments. Other providers include Prequin (used by OECD.ai), PitchBook, Netbase Quid (used by Stanford AI Index)

Policy Regulations and Initiatives. Policy makers in the AI field include EC, Council of Europe, OECD, UNESCO, Global Partnership for AI (GPAI) and national initiatives, Standardization bodies in the AI field include ISO and US NIST. A good source of policies and standards in this domain is the AI Standards Hub[41]. InnoGraph needs to track the new developments of policies and standards, and policy makers will be an important user audience for InnoGraph.

News and Social Media. When innovation progresses enough, news and posts start appearing in mass- and social media. Tracking the growth of such news is an important indicator of the growth and relative importance of AI innovations. We plan to use EventRegistry [17] for technology news (a company affiliated to JSI), and Twitter for social media postings on the topic.

Received 7 February 2023; revised 6 March 2023

[35] https://huggingface.co
[36] https://replicate.com
[37] https://ror.org
[38] https://mix-n-match.toolforge.org/#/catalog/3344
[39] https://stackexchange.com/sites
[40] https://sn-scigraph.figshare.com/articles/dataset/Dataset_Grants/7376474
[41] https://aistandardshub.org