Computational Critical Theory: Illuminating implicit power relations in AI research through a Knowledge Graph

Anna Natalia Wróbel[1]

Vrije Universiteit Amsterdam
a.n.wrobel@student.vu.nl
Artificial Intelligence Bachelor Final Project
supervised by Dr Stefan Schlobach

Abstract. With the rapid advancement of Artificial Intelligence (AI), scholars in critical theory have increasingly engaged with the rise of AI technologies, offering diverse perspectives that highlight the implicit power relations embedded in contemporary AI research and industry practices. This work presents a tool designed to support such critical analyses in an automated way through the development of a Research-Power-Knowledge Graph (RP-KG), constructed using a custom-designed Research-Power Ontology (RPO). This solution models the scholarly domain with a particular focus on power dynamics, drawing on methodologies from the field of linked open data and focus areas forwarded by critical scholars. It integrates existing ontologies and online knowledge graphs, and enriches them with domain-specific properties and instances using standard natural language processing (NLP) techniques. The resulting tool is evaluated in terms of scalability, usability, and its capacity to respond to SPARQL queries that reflect concerns articulated within critical theory. Although the tool remains limited in scope and exhibits constraints in precision, it addresses a significant research gap by combining the strengths of multiple existing resources while enabling the investigation of critical issues related to power dynamics in AI research.

Keywords: semantic web · knowledge graph · scholarly metadata · critical theory · power relations · AI research · AI industry

1 Introduction

Focus on artificial intelligence (AI) has been growing rapidly in academic circles. We can observe a magnitude of research papers published every year, developing new methods and software solutions which get incorporated into our daily lives. Apart from successful AI technologies which enraved the public, and inspired scientists, there are multiple cases which led to drastic consequences. For instance, the ability of Generative AI (GAI) to produce Deepfakes facilitates disseminating disinformation, posing a threat to personal rights as well as democratic processes[3].

While AI scholars constantly put effort into understanding why algorithmic structures reproduce such biases and contribute to other possible societal risks, they rarely play a role in directing the way and scale of implementation of such technologies, especially in the private sector[24]. This is because research production and implementation is often guided by the power relations between research institutions and private enterprises, whose functioning is driven by a preoccupation with profit over the wellbeing of the users [23]. Companies invest in gathering materials, tools, and funds, and can then exert power over research institutions through allocation the above-mentioned resources in ways that accelerate scientific developments which are the most profitable for the private sector[1].

Power relations in knowledge production have been conceptualized in many ways by Critical Theory scholars, and a particular focus on economic incentives consistently emerges as a useful lens through which we can analyse how financial power structures influence academic dynamics[20]. This framework can be applied to examine the production, dissemination and application of AI research. One methodological approach for applying the framework involves using linked open data to systematically model scholarly metadata. Such tools can capture key features of academic influences, including authorship networks, thematic trends, publication and funding relationships, thereby offering a structured means of analysing some aspects of power dynamics within the field.

Cataloguing existing knowledge so that it becomes machine-readable and submittable to statistical analysis is often done by developing knowledge graphs (KGs) with scholarly data such as OpenAlex[33]. These resources allow us to get a broader overview of academic works, but critical analyses require qualitative investigations which go beyond the surface-level metadata. There is a need for creating new ways of organising AI knowledge in order to produce socially beneficial research. Without novel frameworks which take into account various critiques forwarded by critical studies on computation and knowledge production, there is a great risk of implementing software that amplifies existing biases and centralizes power in profit-oriented entities at the expense of focusing on preventing potential societal harms. Existing tools in the area of Semantic Web, which could be suitable for this task, include SPAR ontologies[30], which model publishing metadata such as authorship, citations and bibliography. Another prominent software is the Computer Science Knowledge Graph[11] (CS-KG), which looks beyond scholarly metadata by applying Natural Language Processing (NLP) to derive field-specific methods and tasks constituing the research contribution. Online tools such as Openalex[33] or Crossref[1] include both metadata as well as authors' affiliations and funding information, crucial for understanding power dynamics in academia.

How do these methods of knowledge organisation correspond to the analyses presented by Critical Theory and Critical Data researchers? Existing body of literature provides multiple perspectives on power relations in academia, which are often connected to maintaining economic advantage or ideological pursuit.

[1] https://www.crossref.org/

Foucault [15] argues that knowledge and power are deeply intertwined. Consequently, all production of knowledge is influenced by an underlying distribution of power, which is enforced by institutional structures [13]. These can be found in most, if not all entities contributing to the process of producing knowledge, e.g. universities, private research institutes, publishing agencies, funding entities [31]. Another popular claim is that research is inherently value-laden and under the influence of particular ideologies [14][7]. Simon Lingren [24] emphasizes this view by utilizing Deleuze and Guattari's [17] framework to present AI as an assemblage. He outlines how AI technologies are built upon input from multiple subjective counterparts - from humans and data, through political economy, to algorithms and artifacts. Academic structures are part of this assemblage too. CS and AI departments are exercising power through specific allocation of funds, establishing academic rankings, the ability to shape the curriculum, and leveraging prestige in order to receive funds from external, often profit-oriented entities. Economic capacities of departments decide whether scholars receive access to non-open-access publications, or State-of-the-Art tools for conducting research. Scholars, who have to rely solely on provided resources, without the opportunity for critically motivated selection of quality materials, risk using biased materials, as datasets in CS research often reflect and reinforce existing biases [6]. The distribution of research money can influence academic freedom by giving a financial incentive to pursue only particular research areas [18]. Existing research on this topic shows that funding entities, despite presenting themselves as neutral towards all research topics, deliberately shaped academic interests [1][21]. By selecting which research to fund, they diverged attention from the topics they deemed contrary to their goals.

These works put forward a demand to inquire the academic field of Artificial Intelligence by taking into consideration the multiple components of the AI assemblage. Funding bodies, affiliations and publishers are agents which have an influence on the content and success of the paper. Resources used - whether it's research articles, datasets, methods or metrics - need to be modelled as agents, because they carry and enforce values and tend to centralize power[14]. Geographical features of publishing platforms and affiliations need to be connected to both the methods and the topics and implications of research, as they showcase processess which can centralize power in western-centric perspectives.

In order to fill this research gap, I propose an Research-Power-Ontology (RPO) - an ontology about AI research and industry, inspired by various acknowledges semantic web tools, and populated with metadata from online resources in order to create a Research-Power Knowledge Graph (RP-KG). The graph is further enriched by information from the CS-KG and similar NLP approaches to track more corporate affiliations and economic links in the provenance of resources. I propose a few trial queries to investigate the structural composition of the ontology and populated KG in order to answer problems outlined by Critical Data studies. The goal of this project is to create a transparent architecture of scholarly data which can later be easily used by critical theorists to further analyse the field of AI from a critical perspective. Thus, this study addresses the following research question: Can scholarly ontology and knowledge graph enriched with NLP-mined features facilitate critical analyses of the field of Artificial Intelligence?

The Literature Review outlines how previous literature has up until now approached the task of organising and analysing scholarly data of AI research from a computational and critical perspective. The Methodology section outlines an integrated framework combining ontology design, knowledge graph construction based on a suitable dataset, and NLP enrichment. Experiments and Evaluation assesses the resulting knowledge graph through structured queries and validation methods. Finally, the Discussion evaluates the tool's strengths, limitations, and potential directions for further research.

2 Literature review

The computational approaches relevant to this project encompass existing ontologies suitable for modelling scholarly metadata, structures of academia, industry relations and provenance of resources and ideas. Furthermore, several online datasets and knowledge graphs such as Crossref of Openalex provide a practical framework for orgaising such characteristics. Open general KGs such as DBpedia and Wikidata model instances relevant to the AI industry with enough detail to link them to general problems highlighted by Critical Theory scholars. CS-KG employs an NLP pipeline to enrich scholarly metadata with information derived from paper abstracts. The theoretical approach used to support the design choices for this ontology and KG encompasses critical and sociological perspectives on academia which analyse the socioeconomic relations that create power relationships. More specific strands of this field, such as Critical Data Studies, Decolonial Theory and Science and Technology Studies (STS) provide critical perspectives on the emerging technologies such as AI and the consequences of their widespread production application.

Relevant ontologies. SPAR ontologies[30] are academically acknowledged and popular tools for modeling scholarly publishing, encompassing a range of scholarly data about bibliography, citations, publishing processes and standardized evaluation metrics. Bibliographic ontologies like FaBiO[29] classify published works, identifiers (e.g., ORCID), and article components. Citation ontologies, such as CiTo[29], categorize citation intent, enabling granular citation analysis and mapping methodological provenance. Publishing process ontologies (e.g., PRO[28], SCoRO[2], FRAPO[3]) define personal and institutional roles (authors, reviewers, funders), institutional affiliations, and temporal aspects of contributions. Complementary ontologies like FiveStars[35] and BiDO[27] quantify research impact. FiveStars evaluates publications using rubrics such as open-access compliance, peer-review transparency, and dataset availability or provenance.

[2] https://sparontologies.github.io/scoro/current/scoro.html
[3] https://sparontologies.github.io/frapo/current/frapo.html

These features allow for assessing compliance with open science standards, but can also be extended into tools to look for implicit power structures. This set of ontologies is used as an inspiration to outline the most relevant features of the scholarly publishing domain to be included in the RPO design. These features encompass modeling of: bibliographic metadata, authorship, institutional affiliations, publishing data, citations, materials used and accessibility.

PROV Ontology (PROV-O) is an ontology designed to model provenance, structured around three core classes: Agent, Entity, and Activity. It captures how Entities (e.g. charts, research tasks, or digital artifacts) are generated by Agents (individuals, institutions, or software) through Activities, with temporal specifications such as start and end times. It allows for modeling derivation of Entities from other Entities and attribution of Activities to Agents. An interesting feature is the SoftwareAgent subclass, which explicitly models automated creation of entities. Nowadays, we can see an example of this by looking at AI-generated content. These classes are used as a starting point for the RPO.

Relevant Knowledge graphs and data networks. The CS-KG leverages PROV-O to annotate research entities (e.g., Task, Method, Material) extracted from research papers via NLP. It leverages specific properties to link entities to the NLP methods that identified them, ensuring provenance transparency. For instance, a statement like <cskg:semantic_interoperability, cskgont:usesTask, cskg:ontology_matching>[11] is derived from the abstract of a research paper, and later verified via a Machine Learning checker module. The resulting KG integrates: machine-readable research data (e.g., tasks, methods), provenance traces via PROV-O, and sources of such statements via pointers to research papers. This allows scholarly metadata KGs to be enriched field-specific methods, metrics, tasks and materials used in research.

OpenAlex[33] organises metadata into key entities, including academic works, authors, related institutions, funding bodies, and publishers. It covers a wide range of metadata, including geographical information, and is said to be the largest open dataset of scholarly works [9]. Some studies have noted limitations in its coverage [39], but these claims don't contest the wide usage, popularity and overall user satisfaction with OpenAlex. Another similar resource is Crossref[4], which is based on the academic community's contributions of scholarly metadata and employs tagging mechanisms to reveal connections between publications based on topics, funding sources, authors, and other attributes.

DBpedia[22] and Wikidata[37] are large-scale knowledge graphs that integrate both machine-readable and human-curated data across diverse domains. Within this project, they serve primarily to augment the knowledge graph with metadata otherwise inaccessible through OpenAlex, Crossref, CSKG, or standard NLP extraction techniques. While much of this data, such as institutional structures, geographic context, and industry networks, are not directly related to the scholarly domain, they are necessary to address critical inquiries.

[4] https://www.crossref.org

NLP enrichment to scholarly knowledge graphs. CS-KG employs a specialized natural language processing pipeline called SCICERO[12], which processes abstracts to generate structured statements in the form of RDF triples. The extraction process involves several NLP tools:

- DyGIE++: extracts a set of entities of 6 pre-defined types (Method, Task, Material, Metric, Other-Scientific-Term, and Generic) and seven kinds of relationships between them (Used-for, Hyponym-Of, Compare, Part-of, Conjunction, Feature-of, Evaluate-for).
- CSO Classifier (CSO-C): used to extract CS-specific topics from the text.
- OpenIE: extracts relations based on the entities previously defined by DyGIE++ and CSO-C.
- Part of Speech Tagger (PoST): used to find verbs in the relations found by OpenIE so that triples of form (entity, verb, entity).

Following extraction, a handler modules and ML checker modules clean up the KG from entities which do not conform to the design. An evaluation of this pipeline demonstrates that the combined use of OpenIE, PoS tagging, and DyGIE++ yields complementary extraction capabilities, resulting in a knowledge graph containing over 41 million statements and 10 million entities derived from 6.7 million scientific papers. Furthermore, CS-KG integrates external knowledge by linking extracted entities to established knowledge bases such as DBpedia and Wikidata. The resulting KG is stored at http://w3id.org/cskg and is used in this project to extract domain-specific information from a selected set of papers.

Critical approach to power relations in academia is informed by sociology and various strands of critical theory, including critical data studies, science and technology studies (STS), decolonial theory and feminist studies, and applied to the modern context of academia, academic funding and knowledge production. Critical scholars have long argued that academia is governed by implicit power relations, masked by a veil of scientific neutrality [15]. As Dosi and Marengo [13] observe, knowledge is inherently intertwined with the power dynamics of institutional structures. Abdalla and Abdalla [1] draw from the example of tobacco industry and argue that corporate interests can significantly shape the direction and visibility of tech research, skewing attention away from findings that threaten the industry's bottom line. Similarly, Lagner and Knyphausen-Aufseß [21] emphasize that funding bodies, by virtue of controlling the allocation of financial resources, often prioritize projects that conform to dominant paradigms or respond to narrowly defined societal concerns. Danchev et al. [10] prove that tight-knit academic communities built around shared methodologies and exposure to similar prior knowledge, reinforced by performance-based reward systems, produce research of lower replicability. Birhane et al. [7] conclude that wealth accumulation (among other objectives) can significantly direct the field of AI research by underfunding research with commitment to social benefit and justice.

Scholars from the area of decolonial theory point out that claims of scientific objectivity often obscure global power imbalances, allowing western research norms to dominate internationally [25] [4]. These norms are frequently irrelevant to, or even harmful in non-western contexts and are seen as extensions of colonial-era dynamics, especially when embedded in academic funding and collaboration structures. Alvares [4] llustrates this through a European–Indian university partnership, where European input is framed as progressive while Indian knowledge is marginalised.

Brown [8] uses a marxist lens to critique the capitalist economy's influence on research production and argues that under current economic conditions, academic work is increasingly commodified, with scientific outputs treated as tradable goods. When economic efficiency emerges as the primary criterion for funding decisions, the focus shifts from knowledge advancement to profit generation. Arboledas-Lérida [5] further contends that becaue of competitive project-based funding (CPBF), the pursuit of funding becomes a structuring force that dictates research agendas. Knoche [20] extends the marxist critique to the domain of science communication and outlines that several actors are at play when publishing articles, each of them with their own economic incentives. Part of this network consists of private, profit-oriented academic publishers and political institutions with an interest in maintaining control over regulations and governance. Different interests are at interplay through a network of power relations, which do not prioritise ethical research practice.

Critical approach to big data and AI. Recently, scholars apply these critical lenses to technology, big data, and AI. Birhane et al. [7] state that there is a "foundational understanding in Science, Technology, and Society Studies (STS), Critical Theory, and Philosophy of Science that science and technologies are inherently value-laden, and these values are encoded in technological artifacts". They show that values such as novelty, performance and generalisation are pursued at the expense of societal benefit, fairness, environmental considerations and evaluations of possible negative consequences. Framed as objective, these values serve as instruments of exerting power. Similarly, Mohamed et al. [25] highlight that values significantly influence the choice of research and the context in which science is applied. Thus, agents who decide which values are reflected in academic works, have the power to shape academic discourse and its application. As Pitsoe et al. note [31], this dynamic showcases power relations in both the production and legitimisation of knowledge.

To understand the complex power relations and values embedded in AI, this project utilises Simon Lindgren's approach[24] and conceptualises AI as an assemblage to argue that AI emerges through the interplay of multiple heterogeneous counterparts. Human developers, data, political economy, algorithms and material artifacts constitute parts of this network. In this view, AI systems are not neutral or self-contained but are constituted through socio-technical networks shaped by power dynamics. It is necessary to look at all these parts separately and at the power relations between them in order to understand AI research, development and application. Iliadis and Russo[19] support this framework and encourage viewing AI and Big Data as assemblages that emerge across multiple scales (local to international) through structures that exert power. Data ownership, access, and usage are unevenly distributed, giving rise to data capital [38], which is leveraged by organisations to dominate markets, shape research discourse by selecting which areas (and topics) are lucrative, and influence society at large by distributing preferred algorithms.

The power asymmetry extends to methods and resources used in research. Birhane et al. [7] critique the widespread and uncritical reuse of benchmark datasets, even when they do not represent reality that resulting algorithms will be applied to [7]. Marginalized groups are especially vulnerable to such omissions, as data on their circumstances may be harder to find, more expensive to curate, and therefore omitted. Iliadis and Russo [19] refer to such exclusions as the "power not to look," highlighting how invisibilising certain social-related data is an exercise of power. Dotan and Milli [14] further argue that access to large datasets and the computational power required to process them, are concentrated in the hands of a few actors. This technical and infrastructural exclusivity centralizes power, even when resources appear affordable. Thus, technical constraints are deeply entwined with the political economy of AI.

Geography also emerges as an actor in AI assemblage through connection to decolonial theory. Mohamed et al. [25] situate the reapprasal of western-centric cultural perspectives within a broader framework of algorithmic oppression, calling for the integration of decolonial theory into critical AI discourse to better understand the transnational and racialised dimensions of power embedded in algorithmic systems. In addition to data and geography, attention must be paid to institutional actors and their influence on the AI research ecosystem, as strong corporate involvement shape not only research agendas but also notions of scientific value and prestige. Finally, at the level of public policy, Rieder and Simon[34] examine how data-driven logics increasingly inform governance structures, raising concerns about the uncritical adoption of decision-making models.

3 Methods

Among various perspectives which analyse or describe the state of AI and AI research, it seems that none encapsulates all the necessary aspects (table 1). Online tools like OpenAlex provide large, machine-redable dataset of papers and their characteristics, but do not consider the content of the paper and do not have a focus on power relations. CS-KG, in turn, focuses on the content of the paper to mine more metadata regarding research entities, but does not encompass funding information and any power dynamics. Lastly, critical studies look at funding, paper content, and power relations, but do not produce machine-readable outcomes and apart from having a broad perspective on the field, they do not operate by analysing a large dataset of papers, but rather support general claims with examples.

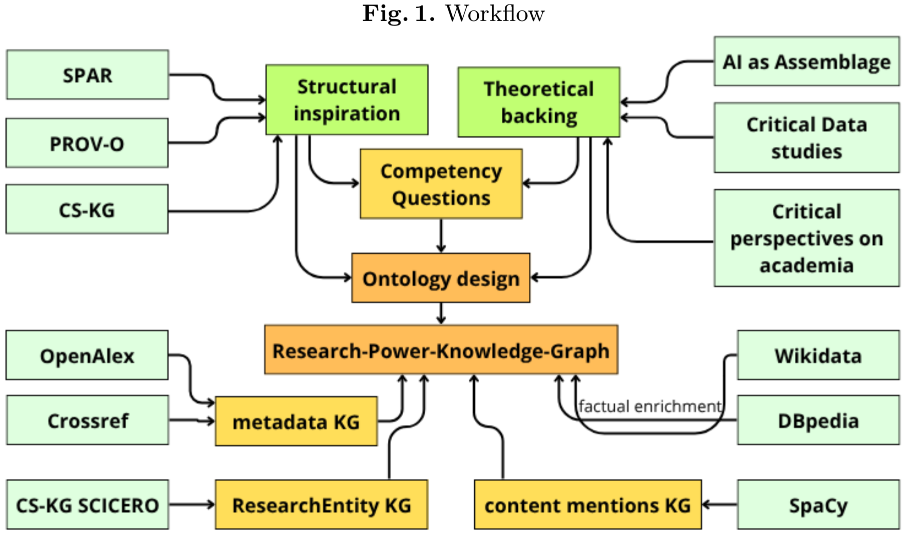
There is a need for a new solution encompassing all these characteristics. This project aims to create a Research Power Knowledge Graph, based on a Research Power Ontology designed to capture power relations, populated with scholarly metadata and enriched with text-mining methods to get more information from the content of the paper. This tool will then be used to answer the research question: Can scholarly ontology and knowledge graph enriched with NLP-mined features facilitate critical analyses of the field of Artificial Intelligence?

Table 1. Existing work and the research gap

feature / tool	OpenAlex	CS-KG	Critical studies	RP-KG
Large dataset	yes	yes	partial	?
Machine-readable	yes	yes	no	?
Scholarly metadata	yes	partial	no	?
Funding information	yes	no	yes	?
Analyses paper content	no	yes	yes	?
Focus on power relations	no	no	yes	?

The methodology encompasses: (1) designing an ontology, (2) selecting a set of papers and (3) populating the KG with its online-scraped metadata, (4) mining more metadata using custom NLP methods. The workflow is illustrated in figure 1.

Fig. 1. Workflow



The ontology design process followed established principles of ontology development created by Noy and McGuiness [26]. The initial phase involved defining the ontology's scope through two categories of competency questions. Basic competency questions (presented in table 2) were inspired by the relevant existing ontologies and KGs, and Lindgren's [24] framework of AI as an assemblage, which conceptualizes AI research components as elements capable of mutual influence through power relations. These questions informed the core ontological structure, with Knoche's [20] analysis of scholarly publishing agents providing additional grounding. Extended competency questions emerged from critical theory literature and served dual purposes: guiding ontology design and providing evaluation criteria. These questions represent a significant contribution, as the resulting software's structure should allow to answer questions emerging from the field of Critical Theory.

Table 2. Competency questions and resulting ontology design components

Competency Question	**Guided Ontology Components**
What is this academic work influenced by?	Class `AcademicWork` and subclasses, influence properties, influencing entities
Where does the funding come from?	Subclasses of funding entities (funders, grants, fellowships)
What institutions are involved?	Subclasses of influencing entities, production/publishing properties
Who are the authors?	Class `Author`, authorship properties
What influences the authors?	Author affiliation properties, subclasses of author-influencing entities
What social need does this address?	Subclasses of societal needs, relevant properties
What research entities are used?	Subclasses for methods/materials, usage properties
What topics are mentioned?	Subclasses for AI topics, keywords, subject areas
Which geographies are related?	Geospatial classes (countries, territories) and properties
What are the citations?	Citation properties connecting works
What is the accessibility?	Publishing cost/access properties

Extended competency questions, outlined in table 3, are concluded from the critical theory work described in literature review. Questions 1-3 are inspired by the work of Dotan and Milli [14], Yousif[38], Knoche[20], Iliadis and Russo[19], who focus on economic links hidden in institutional structures and networks between academics, companies and publishers. These questions look at economic power relations in AI research field. Questions 4-6 are inspired by the work of Alvares [4], Agarwal et al. [2], Prasad[32] and Mohamed et al.[25], who offer a decolonial research perspective. These questions can be answered if there are sufficient geographical and geopolitical relations exposed between research entities, authors, influencing institutions, publishers and funding bodies. Questions 7-10, inspired by the work of Tang et al. [36], Iliadis and Russo [19], and [1], are about relevant topics in AI research and their application in public sector.

These questions can be answered if there are sufficient relations between topics of research papers (on the intersection of technological areas and public sector areas), institutional ties of such research, and whether they are profit-oriented. Questions 11-15 are inspired by the work of Danchev et al.[10], Dotan and Milli [14], and Birhane et al. [7]. They look for homogenous research practices, popular materials and topics and their relation to corporate ties and fund allocation.

Table 3. Extended competency questions

#	**Question**
1	Are there overlapping affiliations in publisher sponsors, authorship sponsors, methodology sponsors, and overall sponsors?
2	Who is the sponsor of materials or methods used in the research?
3	Is the publisher sponsored by any entity?
4	Does research promote western-centric perspectives?
5	Are there biased western-centric materials or methodologies applied to non-western context?
6	Are there financial dependencies between agents involved in production and dissemination of knowledge that reinforce centralisation of power in the global west?
7	What topics in AI research are popular in applications to the public sector?
8	Are there significant corporate influences in research concerning these topics?
9	Do the influential entities prioritize social benefit?
10	Do the related organisations constitute governmental agents?
11	Are there particular methods and materials that are overused, suggesting homogenous research ideology?
12	Are these methods and materials known to be biased, controversial, or outdated?
13	Are the projects using these materials sponsored by particular institutions?
14	Are there significant differences between popularity between particular methods or research areas, academic departments?
15	Are these correlated with differences in funding allocated to these areas?

The Research-Power-Ontology design is primarily inspired by the PROV-O ontology and CS-KG. RPO retains `Entity` and `Agent`, but omits `Activity`, because Activity requires temporal properties, but power relations can escape rigid temporal frames (e.g. popularity gained from well-funded research extends further than the process of publishing the research). The `Entity` class is enriched following the structure of CS-KG, incorporating research-specific entities and their subclasses. In addition to `ResearchEntity`, the ontology includes subclasses such as `AcademicWork` and its subclasses (e.g. Article), as well as `ResearchPlatform` and subclasses (e.g. Journal). This structure allows for a more granular representation of research outputs, linking them to for-profit platforms or costly resources. For modeling agents, the ontology adopts subclasses such as `Person`, `Organisation`, and `AgentGroup`. `SoftwareAgent` is introduced to capture the notion that software can encode ideologies and values, and thereby implicitly influence AI research in which it is used. This reflects the argument that technological artifacts are not ideologically neutral. The `Organisation` class is further refined with domain-relevant subclasses, such as `University` or `Company`. `Author` is modeled as a subclass of `Academic` under `Person`. Financial relations are captured through a dedicated `Fund` class, which supports specific properties like grant amounts and contract types (e.g., fellowships). However, financial influence is not limited to formal funding structures: the property `funded_by` allows agents to be linked outside of grant or contract contexts, such as in cases where research institutions or software agents receive patronage. Topical relevance is addressed through a set of classes including `Topic`, `Keyword`, `Goal` and `ResearchArea`, which are used to support competency questions about the prevalence of specific topics or alignment of research with societal needs. These thematic elements can be evaluated qualitatively to determine whether a given topic constitutes a societal concern, or whether a keyword is associated with known social issues. Geospatial aspects are included to situate agents and research outputs within a global context. Relevant classes are `Geographical`, `Continent`, `Country`, and `City`, along with property `located_in`.

A central concern of this ontology is to model various forms of influence. To this end, several subproperties of `influenced_by` are introduced and connect the consituents of the AI assemblage. These include `funded_by` to express financial influence, `written_by` to model intellectual influence via authorship, `works_with` and `works_for` to represent the influence of collaboration and institutional affiliation. The property `owned_by` is used to denote that software agents may carry embedded values inherited from their owners. Citations are also recognized as forms of influence, hence the inclusion of `cites` to reflect how researchers build upon and are shaped by prior work. `key_person` highlights how people in high managerial positions influence their organisations' practice. Properties `located_in`, `affiliated_with`, `part_of`, `published_by` describe relations pertaining to geography, affiliation, ideological or material constituency (part of a whole follows ideas enforced by the whole) and publication influence. Further, the ontology incorporates properties derived from CS-KG to model dependencies among research entities, for example, to show how methods are influenced by the datasets they use. This is especially important for tracing how bias in datasets can propagate through to the methods and findings of research. The property `uses` is included to capture the tools and resources applied in research, acknowledging that these too can carry implicit biases. These modeling choices were informed by the set of competency questions aimed at uncovering mechanisms of financial, institutional, geopolitical and ideological influence within the research ecosystem.

The process of KG population started with a dataset consisting of the 100 most-cited papers in the field of artificial intelligence, originally compiled by Birhane et al.[7], who harvested papers from NeurIPS and ICML proceedings for the years 2008, 2009, 2018, and 2019. The current study aims to extend the earlier analyses by representing these papers as structured knowledge based on the developed ontology. The popularity of these works demonstrates influence within the field. If power dynamics are present within these articles, their wide citation means such influences may propagate into subsequent research. A notable limitation is that this dataset is relatively small and does not encompass most recent works.

To maximize metadata coverage, both the OpenAlex and Crossref APIs were queried. Although OpenAlex is generally considered the most comprehensive open bibliographic database, 14 of the 100 papers were not found in either OpenAlex or Crossref, and thus were excluded from automatic metadata retrieval. For the remaining papers, the retrieved metadata was used to populate the ontology with RDF triples describing a range of foundational properties, including bibliographic metadata, authorship and institutional affiliations, funders and associated grants, cited works, keywords, topics, field or research area classifications, and connections to sustainability goals. Additional metadata about involved organisations was obtained via queries to the DBpedia SPARQL endpoint. This included information on location, organisational type (for class assignment), and key figures (e.g., owners or presidents). DBpedia was selected for its reliability and recognition as a trusted source of linked open data. The result of this process created a subgraph of the general knowledge graph with the structure presented in figure 2.

NLP enrichment was carried out by involving results of the SCICERO[12] pipeline and applying a standard NLP method on papers' contents. Access to CS-KG was obtained via its SPARQL endpoint. To populate subclasses of `ResearchEntity`, the DOIs of selected papers were queried to retrieve all associated statements. Based on these, a new graph was constructed, consisting of: the extracted statements, triples identifying connected entities, and references to the papers. A limitation of CS-KG is that it only includes publications from 2010 onward. Furthermore, over 80% of the papers from 2018 and 2019 in the dataset were not included in CS-KG, which restricted the usable subset to only 8 papers. The resulting subgraph comprised 61 instances of academic works, 49 statements, and 66 `ResearchEntity` instances. The statements were then integrated into the RP-KG by linking the research entities back to their respective papers, e.g. `rpo:paper/17 rpo:mentions cskg:grover`.

Since CS-KG contains only abstracts, a separate method was employed to extract entities from the full text of papers (from introduction to references). The SCICERO pipeline, while effective for identifying domain-specific computer science entities, is not suitable for detecting broader entities such as individuals or organisations, nor for modeling power relations, which are difficult to capture reliably with existing NLP methods. As a result, the next step focused on extracting entities, rather than full triples. Each paper was linked to identified entities using the generic `rpo:mentions` property, with sentence-level provenance included. Although the ontology supports more specific subproperties of `rpo:mentions` such as `rpo:criticises` and `rpo:endorses`, these were not de-

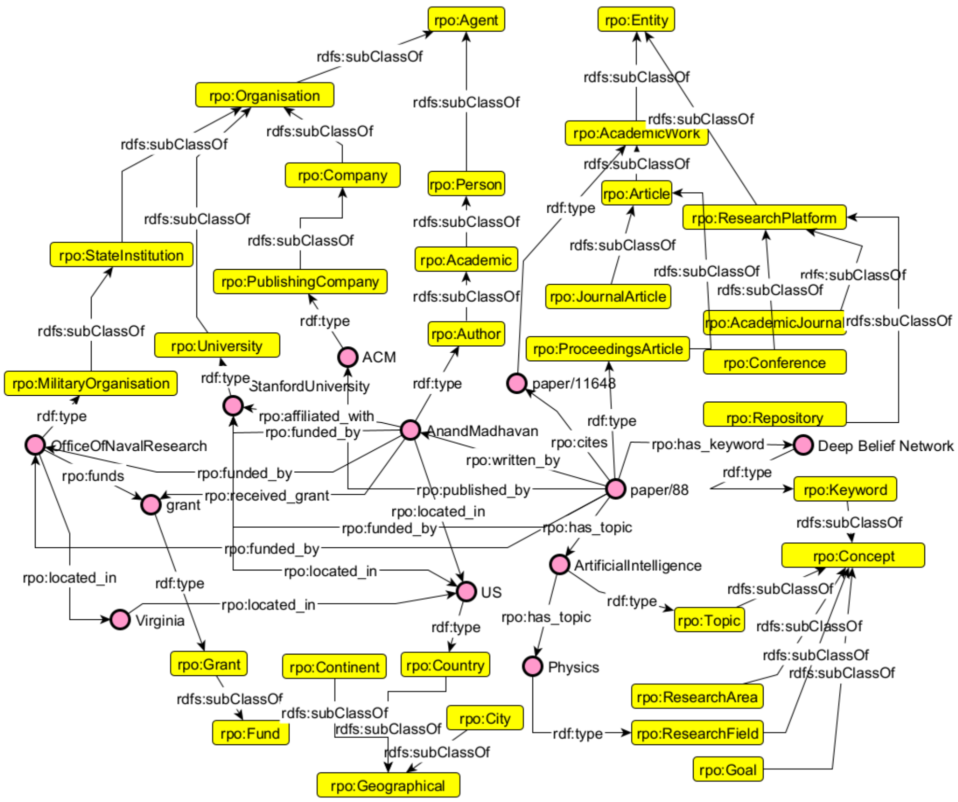


Fig. 2. Metadata subgraph of one author's one article

rived due to the limitations of current NLP techniques and the project's scope. This problem could direct future work on topic.

The SpaCy [16] pre-trained Named Entity Recognition model was used and applied to cleaned paper text. Other enhanced NLP methods designed to process scientific text were consiered, but their entity recognition ability was found to be overly fine-grained for the ontology's needs. Given the lack of a domain-specific training set, accuracy was expected to be limited. To improve reliability, a Wikidata verification module was implemented. This step ensured that only entities acknowledged by Wikidata are added to the KG. Entities' classes were also retrieved from Wikidata, and a heuristic mapping was applied to align it with the ontology's class structure. In cases of ambiguity, entities were conservatively assigned the generic type `owl:Thing`. As a result, 4493 named entities were initially extracted from the full-text corpus. Of these, 1,964 were excluded due to the absence of matching entries in Wikidata.

Subgraphs created by the three methods: OpenAlex and Crossref Retrieval, CS-KG Retrieval and SpaCy Retrieval were combined with the Research Power

Ontology to create the Research Power Knowledge Graph. The KG underwent final post-processing, which involved minor manual changes, as well as automatic duplicate removal process and automatic geospatial enrichment on the entities added with SpaCy Retrieval. The resulting software can be accessed at https://github.com/sawauua/RP-KG .

4 Evaluation

The most important quantitative properties of the resulting knowledge graph are described in table 4. All the methods used to construct the graph provided complementary contributions to populate specific classes or create triples with specific properties. As mentioned in section3, table 1, the key components to

Table 4. Resulting graph sizes

	OpenAlex, Crossref Retrieval	**CS-KG Retrieval**	**SpaCy Retrieval**	**RP-KG cleaned**
Individuals	2800	431	2612	5843
Authors	270+	0	0 (70 Persons)	270+
Organisations	80+	0	100+	200+
AcademicWorks	1800+(incl. cited papers)	60+ (incl. other papers supporting statements)	94	1960+
Geographical	60+	0	4	90+
Research Entities	0	60+	0	60+
influence triples	3900+	15	0	4000+
rpo:mentions triples	0	120+	760+	800+

the resulting software were: a large dataset, machine-readable format, scholarly metadata, funding information, data gathered from paper content and focus on power relations. RP-KG struggles to fulfill the first requirement, as the chosen dataset only encompasses the most cited papers from particular years, and the CS-KG failed to provide extensive data about Research Entities pertaining to this dataset. The requirement of analysis of paper content can be considered as partially fulfilled. A number of instances was created using SpaCy, yet there remains a possibility of noise, and incorrectly classified individuals due to lack of information in Wikidata. Analysis performed on the complete KG showed that over 60% of instances created wth NLP were conservatively classified as `owl:Thing`, which influences query capacities and requires more qualitative evaluations. The remaining criteria were met, and none of the existing solutions satisfy the criteria as much as RP-KG does, leading it to be a successful attempt of bridging the research gap.

Table 5. Existing work and RP-KG result comparison

feature / tool	OpenAlex	CS-KG	Critical studies	RP-KG
Large dataset	yes	yes	partial	no
Machine-readable	yes	yes	no	yes
Scholarly metadata	yes	partial	partial	yes
Funding information	yes	no	yes	yes
Analyses paper content	no	yes	yes	partial
Focus on power relations	no	no	yes	yes

Does this approach fit the desired critical perspective? To check if the project achieves the critical approach, the extended competency questions were evaluated with respect to the standalone structure of the ontology and the populated KG with inferred axioms to ensure completeness. SPARQL queries mirroring these questions are outlined in Appendix A.

Questions regarding funding: Extended Question (EQ)1 *Are there overlapping affiliations in publisher sponsors, authorship sponsors, methodology sponsors, and overall sponsors?* can be answered by comparing funders, and grant sponsors per paper. The answer can be extended with a suitable query tackling EQ2 *Who is the sponsor of materials or methods used in the research?*, which looks at triples `rpo:PaperX rpo:mentions rpo:AgentX`, types of mentioned entities and their owners or creators. Partial result of query 6 shows an instance where a paper which mentioned Google to describe data provenance, was funded by DeepMind, which is part of Google. EQ3 *Is the publisher sponsored by any entity?* can be answered structurally by searching for funders of Research Platforms (Journals, Conferences, Repositories). However, the RP-KG does not provide material answers to such queries, since the metadata mined from OpenAlex does not encompass funding of publishers.

	paper	mentions	funder	content
1	rpo:paper/4	rpo:Google	rpo:DeepmindUnitedKingdom	"The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer."

Fig. 3. Comparison of used entities and funding bodies in RP-KG

Questions inspired by decolonial theory can be answered by investigating papers addressing sustainability goals and whether they are influenced solely by agents rooted in the context of the global west (judged by geographical connections in the KG) or whether the entities mentioned or materials used emerge from a western context (EQ5 *Are there biased western-centric materials or methodologies applied to non-western context?*). EQ6 *Are there financial dependencies between agents involved in production and dissemination of knowledge that rein-*

force centralisation of power in the global west? can be answered by comparing whether western universities statistically receive more funding and citations, and from which types of organisations, along with their geospatial context. Sample queries 7 and 4 result in statistics which show that 23% of funders sponsoring papers with sustainability goals are corporate funders and they account for 27% of the sustainability-oriented papers. Such research originated in 11 countries, out of which 2 are not part of the global west, and these affiliations constitute only 3,5% of research with sustainability goals. 65% of papers are funded by organisations located in the USA.

To address questions 7-10, topics related to the public sector can be chosen manually. Alternatively, sustainability goals can be chosen as such concepts, and these papers should be investigated to find most popular scientific topics, and their corporate and state affiliations can be verified. EQ9 *Do the influental entities prioritise social benefit?* can only be answered by applying qualitative analysis of such corporate ties. Sample query 15 showed that in papers addressing global goals, the most popular mentioned entities were PyTorch, Grover, and algorithms in the area of machine learning and deep learning. Thanks to pointers to sentences, the RP-KG allows for analysis how such methods are applied.

Questions 11-15 which relate to biased or homogenous research environments can be answered by counting the most popular topics, methods used, or concepts mentioned in papers. Qualitative analysis is necessary to answer EQ12 *Are these methods and materials known to be biased, controversial, or outdated?*, as the Ontology does not structurally define any instance as controversial or biased, however, it enables easy verification by supporting `mentions` triples with sentences they are derived from. For answering EQ14 *Are there significant differences between popularity between particular methods or research areas, academic departments?* and EQ15 *Are these correlated with differences in funding allocated to these areas?*, comparisons can be made between popularity or funding received by papers mentioning popular topics and papers who stay outside of the "hot-topic" zone. Sample queries 12 13 showed that papers mentioning popular entities had on average more funders, but papers which included popular keywords had on average less funders.

	group	avgFunders
1	"not_mentioning_top"	"1.1477272727272727272727"^^xsd:decimal
2	"mentions_top"	"1.25"^^xsd:decimal

Fig. 4. Average funders for papers mentioning popular entities

In general, while the extended competency questions cannot be answered by simple properties, ontology design allows for constructing subgraphs which answer these questions. 5 out of 16 extended questions require qualitative analysis or more data in the KG. The above-mentioned examples do not encompass statistical analysis of the results, as the main goal of the project is to provide

	group	avgFunders	paperCount
1	"not_mentioning_top"	"1.2181818181818181818182"^^xsd:decimal	"55"^^xsd:integer
2	"mentions_top"	"1.0188679245283018867924 53"^^xsd:decimal	"39"^^xsd:integer

Fig. 5. Average funders for papers with popular keywords

a general infrastructure for evaluating different hypotheses. This goal is met because majority of queries can be supported by quantitative examples.

5 Discussion and conclusion

The evaluation indicates that the Research Power Ontology is well-structured to address key questions posed by scholars within critical theory and critical data studies. While certain inquiries, such as the addressing societal needs or defining exactly Western-centric perspectives, still require qualitative analysis, the ontology offers a coherent and sufficiently expressive class and property structure to represent the influence networks in which power relations may be embedded. The RPO constitutes a novel contribution, integrating established ontological standards (e.g., SPAR, PROV) with domain-specific innovations from sources like CS-KG, and adapting them to the particularities of the scholarly domain. The resulting Research Power Knowledge Graph connects bibliographic metadata with AI-specific concepts and enriches them with entities extracted directly from paper content. As illustrated in figure 5, most target components of the envisioned model have been successfully addressed, with the exception of a larger dataset. This leads to concluding that the answer to the Research Question is yes - a curated scholarly Ontology and KG can help critical analyses by providing a machine-readable way to analyse implicit power relations in AI research and industry.

The limited size of the dataset remains a key constraint. Although OpenAlex and Crossref provide broad metadata coverage, both proved incomplete; 14 out of the 100 targeted papers were missing. These sources also required significant validation and enrichment through external queries (e.g., to Wikidata and DBpedia), which themselves may not reflect the latest developments in AI research. For instance, certain entities identified by SpaCy were not found in DBpedia, highlighting the lag between emerging research and its representation in linked open data. Furthermore, current NLP tools struggle to detect complex properties such as provenance of fine-grained methods and ideas, or implicit power relations. A useful addition to this project would be a careful preparation of a suitable NLP model and statistical evaluation of its performance, using a manually annotated training and test dataset. Such model would also provide certainty about the correct classification of instances in the KG.

Additionally, the context of this study requires critical usage of widely adopted tools and datasets: tools that are popular and well-established (such as those outlined in Methodology and Literature Review), may themselves embed power structures. As such, while care was taken to mitigate subjective bias, it cannot be claimed that all contextual or structural biases were eliminated. The wide scope of theoretical grounding, which incorporates multiple strands of critical theory, enabled a comprehensive yet general ontology design. This choice limited the depth of analysis in specific areas. For example, the geopolitical dimensions of data, the social impacts of particular methods on marginalized communities, and systemic issues surrounding open access publishing, academic rankings, and research dissemination (e.g., peer review) were only touched upon in the ontology design. The work of Knoche [20] and others could inform more focused ontological modeling in future iterations.

These limitations point to the multiple directions of future work. First, rather than relying exclusively on external metadata sources, a more targeted corpus could be constructed through custom text mining methods or critically enhanced web-scraping tools. Secondly, more specialised ontologies could be developed for distinct strands of critical theory relevant to AI research and academia. Advancements in NLP, particularly critical models capable of identifying values, ideologies, or institutional biases, are also crucial. These could support wider use of more nuanced properties such as `criticises` or `endorses`. Such tools should connect great specificity when working with scientific text and data, and ability to highlight socioeconomic or political values uplifted by them.

The Research Power Ontology represents a foundational step toward bridging critical theory and AI research through semantic web technologies. Its machine-readable structure, alignment with established ontologies, and compatibility with linked data make it a valuable tool for conducting transparent critical analysis. The Research Power Knowledge Graph demonstrates the ontology's practical application, capturing instances of influence and institutional power structures in widely cited AI research. While the current implementation remains limited in scope, it provides a proof of concept and lays the groundwork for future, more comprehensive efforts in this area.

References

1. Abdalla, M., Abdalla, M.: The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 287–297. ACM, Virtual Event USA (2021). https://doi.org/10.1145/3461702.3462563
2. Agarwal, D., Naaman, M., Vashistha, A.: AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–21. ACM, Yokohama Japan (2025). https://doi.org/10.1145/3706598.3713564
3. Al-Khazraji, S.H., Saleh, H.H., Khalid, A.I., Mishkhal, I.A.: Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications. EPSTEM. 23, 429–441 (2023). https://doi.org/10.55549/epstem.1371792
4. Alvares, C.: A Critique of Eurocentric Social Science and the Question of Alternatives. (2025).

5. Arboledas-Lérida, L.: The Role of Competitive Project-Based Funding in the Commodification of Academic Research: A Marxist Analysis. Critical Sociology. 50, 845–862 (2024). https://doi.org/10.1177/08969205231209679.
6. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 610–623. ACM, Virtual Event Canada (2021). https://doi.org/10.1145/3442188.3445922
7. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M.: The Values Encoded in Machine Learning Research, http://arxiv.org/abs/2106.15590, (2022). https://doi.org/10.48550/arXiv.2106.15590
8. Brown, J.R.: One-Shot Science. In: Radder, H. (ed.) The Commodification of Academic Research. pp. 90–109. University of Pittsburgh Press (2010). https://doi.org/10.2307/j.ctt7zw87p.8.
9. Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., Mayr, P.: Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus. Scientometrics. 130, 2475–2492 (2025). https://doi.org/10.1007/s11192-025-05293-3
10. Danchev, V., Rzhetsky, A., Evans, J.A.: Centralized scientific communities are less likely to generate replicable results. eLife. 8, e43094 (2019). https://doi.org/10.7554/eLife.43094.
11. Dessí, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E.: CS-KG: A Large-Scale Knowledge Graph of Research Entities and Claims in Computer Science. In: Sattler, U., Hogan, A., Keet, M., Presutti, V., Almeida, J.P.A., Takeda, H., Monnin, P., Pirrò, G., and d'Amato, C. (eds.) The Semantic Web – ISWC 2022. pp. 678–696. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-19433-7_39
12. Dessí, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E.: SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain. Knowledge-Based Systems. 258, 109945 (2022). https://doi.org/10.1016/j.knosys.2022.109945
13. Dosi, G., Marengo, L.: The dynamics of organizational structures and performances under diverging distributions of knowledge and different power structures. Journal of Institutional Economics. 11, 535–559 (2015). https://doi.org/10.1017/S1744137414000204
14. Dotan, R., Milli, S.: Value-laden Disciplinary Shifts in Machine Learning, http://arxiv.org/abs/1912.01172, (2019). https://doi.org/10.48550/arXiv.1912.01172
15. Foucault, M., Gordon, C.: Power/knowledge: selected interviews and other writings, 1972-1977. Pantheon Books, New York (1980).
16.
17. Gilman, S.L., Deleuze, G., Guattari, F., Massumi, B.: A Thousand Plateaus: Capitalism and Schizophrenia. Journal of Interdisciplinary History. 19, 657 (1989). https://doi.org/10.2307/203963. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017).
18. Hooi, R., Wang, J.: Research funding and academic engagement: A Singapore case. Knowledge management research and practice 18.2 (2020): 162-174.
19. Iliadis, A., Russo, F.: Critical data studies: An introduction. Big Data & Society. 3, 2053951716674238 (2016). https://doi.org/10.1177/2053951716674238.
20. Knoche, M.: Kritik der politischen Ökonomie der Wissenschaftskommunikation als Ideologiekritik: Open Access. In: Ideologie, Kritik, Öffentlichkeit: Verhandlungen des

Netzwerks Kritische Kommunikationswissenschaft. pp. 140–174. Leipzig University (2019). https://doi.org/10.36730/ideologiekritik.2019.8.
21. Lagner, T., Knyphausen-Aufseß, D.: Rating Agencies as Gatekeepers to the Capital Market: Practical Implications of 40 Years of Research. Financial Market. 21, 157–202 (2012). https://doi.org/10.1111/j.1468-0416.2012.00173.x
22. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., Bizer, C.: DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web. 6, 167–195 (2015). https://doi.org/10.3233/sw-140134
23. Lindgren, S.: Introducing critical studies of artificial intelligence. In: Lindgren, S. (ed.) Handbook of Critical Studies of Artificial Intelligence. pp. 1–19. Edward Elgar Publishing (2023). https://doi.org/10.4337/9781803928562.00005.
24. Lindgren, S.: Critical Theory of AI. Polity Press (2023).
25. Mohamed, S., Png, M.-T., Isaac, W.: Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philos. Technol. 33, 659–684 (2020). https://doi.org/10.1007/s13347-020-00405-8.
26. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology.
27. Osborne, F., Peroni, S., Motta, E.: Clustering Citation Distributions for Semantic Categorization and Citation Prediction. In: Zhao, J., van Erp, M., Keßler, C., Kauppinen, T., van Ossenbruggen, J., van Hage, W. R. (Eds.), Proceedings of the 4th Workshop on Linked Science (LISC 2014), CEUR Workshop Proceedings 1282: 24–35. Aachen, Germany: CEUR-WS.org https://doi.org/10.25504/FAIRsharing.d7f0a9
28. Peroni, S., Shotton, D., Vitali, F.: Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents. In: Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12). Association for Computing Machinery, New York, NY, USA, 9–16 (2012). https://doi.org/10.1145/2362499.2362502
29. Peroni, S., Shotton, D.: FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Journal of Web Semantics. 17, 33–43 (2012). https://doi.org/10.1016/j.websem.2012.08.001
30. Peroni, S., Shotton, D.: The SPAR Ontologies. In: Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A., and Simperl, E. (eds.) The Semantic Web – ISWC 2018. pp. 119–136. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-00668-6_8
31. PITSOE, V., J., VLADUTESCU, S.: A CRITICAL ANALYSIS OF FOUCAULT'S POWER AND KNOWLEDGE IN HIGHER EDUCATION RESEARCH. (2024). https://doi.org/10.5281/ZENODO.15258133
32. Prasad, A.: Discursive Contextures of Science: Euro/West-Centrism and Science and Technology Studies. Engaging STS. 2, 193–207 (2016). https://doi.org/10.17351/ests2016.71
33. Priem, J., Piwowar, H., Orr, R.: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. (2022).
34. Rieder, G., Simon, J.: Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. Big Data Society. 3, (2016). https://doi.org/10.1177/2053951716649398.
35. Shotton, D.: The Five Stars of Online Journal Articles - a Framework for Article Evaluation. In: D-Lib Magazine, 18 (1/2). (2012) https://doi.org/10.1045/january2012-shotton

36. Tang, K.-Y., Hsiao, C.-H., Hwang, G.-J.: A scholarly network of AI research with an information science focus: Global North and Global South perspectives. PLoS ONE. 17, e0266565 (2022). https://doi.org/10.1371/journal.pone.0266565
37. Vrandečić, D.: Wikidata: a new platform for collaborative data collection. In: Proceedings of the 21st International Conference on World Wide Web. pp. 1063–1064. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2187980.2188242
38. Yousif, M.: The Rise of Data Capital. IEEE Cloud Comput. 2, 4–4 (2015). https://doi.org/0.1109/MCC.2015.39.
39. Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., Huang, Y.: Missing institutions in OpenAlex: possible reasons, implications, and solutions. Scientometrics. 129, 5869–5891 (2024). https://doi.org/10.1007/s11192-023-04923-y

A Appendix

Fig. 6. Query: Mentions and corporate influences

```
SELECT ?paper ?mentions ?funder ?content WHERE {
    ?paper rdf:type rpo:AcademicWork .
    ?paper rpo:mentions ?mentions .
    ?mentions rdf:type rpo:Organisation .
    ?mentions rpo:observed_in_sentence ?sent .
    ?sent rpo:has_content ?content .
    ?paper rpo:funded_by ?funder .
    {
    ?funder   rpo:influenced_by  ?mentions . }
    UNION {
    ?mentions rpo:influenced_by  ?funder .
  }
}
```

Fig. 7. Query: Funders in SDG-oriented papers

```
SELECT ?type (COUNT(DISTINCT ?agent) AS ?count)
WHERE {
    ?paper rdf:type       rpo:AcademicWork ;
           rpo:addresses  ?goal ;
           rpo:funded_by  ?agent .

    ?agent rpo:located_in ?country .
    ?country rdf:type       rpo:Country .

    BIND(
        IF(EXISTS { ?agent rdf:type rpo:Company }, "Company", "Other")
        AS ?type
    )
    }
GROUP BY ?type
```

Fig. 8. Query: Searching for popular mentions

```
# Top mentions
SELECT ?mention (COUNT(?paper3) AS ?mentionCount)
WHERE {
    ?mention rpo:observed_in_sentence ?sent .
    ?paper3 rdf:type rpo:AcademicWork ;
          rpo:mentions ?mention .
}
GROUP BY ?mention
ORDER BY DESC(?mentionCount)
LIMIT 10
```

Fig. 9. Query: Searching for most popular keywords

```
# Top Keywords
SELECT ?keyword (COUNT(?paper2) AS ?keywordCount)
WHERE {
  ?paper2 rdf:type rpo:AcademicWork ;
          rpo:has_keyword ?keyword .
}
GROUP BY ?keyword
ORDER BY DESC(?keywordCount)
LIMIT 10
```

Fig. 10. Most popular mentions

	mention	mentionCount
1	rpo:Imagenet	"68"^^xsd:integer
2	rpo:BERT	"58"^^xsd:integer
3	rpo:AI	"53"^^xsd:integer
4	rpo:Gaussian	"46"^^xsd:integer
5	rpo:SGD	"42"^^xsd:integer
6	rpo:Pytorch	"35"^^xsd:integer
7	rpo:NTK	"33"^^xsd:integer
8	rpo:GPU	"30"^^xsd:integer
9	rpo:GNN	"29"^^xsd:integer
10	rpo:RBM	"29"^^xsd:integer

Fig. 11. Most popular keywords

	keyword	keywordCount
1	rpo:Robustness	"10"^^xsd:integer
2	rpo:KernelAlgebra	"6"^^xsd:integer
3	rpo:Initialization	"6"^^xsd:integer
4	rpo:FeatureLinguistics	"5"^^xsd:integer
5	rpo:GenerativeModel	"5"^^xsd:integer
6	rpo:BoltzmannMachine	"5"^^xsd:integer
7	rpo:Smoothing	"4"^^xsd:integer
8	rpo:FeatureLearning	"4"^^xsd:integer
9	rpo:DeepBeliefNetwork	"4"^^xsd:integer
10	rpo:Regularization	"4"^^xsd:integer

Fig. 12. Query: Comparison of funders wrt. popular mentions

```
SELECT ?group (AVG(?funderCount) AS ?avgFunders)
WHERE {
  {
    SELECT ?paper (COUNT(DISTINCT ?funder) AS ?funderCount)
           (IF(BOUND(?topMention), "mentions_top", "not_mentioning_top") AS ?group)
    WHERE {
      ?paper rdf:type rpo:AcademicWork .
      ?paper rpo:written_by ?author .
      OPTIONAL { ?paper rpo:funded_by ?funder }

      # Check if the paper mentions a top entity
      OPTIONAL {
        ?paper rpo:mentions ?topMention .
        VALUES ?topMention {
          rpo:Imagenet
          rpo:BERT
          rpo:Al
          rpo:Gaussian
          rpo:SGD
          rpo:Pytorch
          rpo:NTK
          rpo:GPU
          rpo:GNN
          rpo:RBM
        }
      }
    }
    GROUP BY ?paper ?topMention
  }
}
GROUP BY ?group
```

Fig. 13. Query: Comparison of funders wrt. popular keywords

```
SELECT ?group
       (AVG(?funderCount) AS ?avgFunders)
       (COUNT(DISTINCT ?paper) AS ?paperCount)
WHERE {
  {
    SELECT ?paper (COUNT(DISTINCT ?funder) AS ?funderCount)
           (IF(BOUND(?topKeyword), "mentions_top", "not_mentioning_top") AS ?group)
    WHERE {
      ?paper rdf:type rpo:AcademicWork .
      ?paper rpo:written_by ?author .
      OPTIONAL { ?paper rpo:funded_by ?funder }

      # Check if the paper mentions a top entity
      OPTIONAL {
        ?paper rpo:has_keyword ?topKeyword .
        VALUES ?topKeyword {
                    rpo:Robustness
                    rpo:KernelAlgebra
                    rpo:Initialization
                    rpo:FeatureLinguistics
                    rpo:GenerativeModel
                    rpo:BoltzmannMachine
                    rpo:Smoothing
                    rpo:FeatureLearning
                    rpo:DeepBeliefNetwork
                    rpo:Regularization
        }
      }
    }
    GROUP BY ?paper ?topKeyword
  }
}
GROUP BY ?group
```

Fig. 14. Most popular mentions in papers with SDGs

	mentionedEntity	mentionCount
1	rpo:Pytorch	"35"^^xsd:integer
2	rpo:Grover	"26"^^xsd:integer
3	cskg:grover	"26"^^xsd:integer
4	rpo:GLUE	"25"^^xsd:integer
5	rpo:CPU	"25"^^xsd:integer
6	rpo:PGD	"17"^^xsd:integer
7	rpo:MLM	"16"^^xsd:integer
8	rpo:SVM	"16"^^xsd:integer
9	rpo:CNN	"15"^^xsd:integer
10	rpo:NAS	"14"^^xsd:integer

Fig. 15. Most popular mentions in papers with SDGs

	mentionedEntity	mentionCount
1	rpo:Pytorch	"35"^^xsd:integer
2	rpo:Grover	"26"^^xsd:integer
3	cskg:grover	"26"^^xsd:integer
4	rpo:GLUE	"25"^^xsd:integer
5	rpo:CPU	"25"^^xsd:integer
6	rpo:PGD	"17"^^xsd:integer
7	rpo:MLM	"16"^^xsd:integer
8	rpo:SVM	"16"^^xsd:integer
9	rpo:CNN	"15"^^xsd:integer
10	rpo:NAS	"14"^^xsd:integer