Mining the literature for ethics statements: A step towards standardizing research ethics

Shweata N. Hegde[‡], Ayush Garg[§], Peter Murray-Rust[|], Daniel Mietchen[¶,#]

‡ Regional Institute of Education, Mysuru, India
§ University of Richmond, Richmond, United States of America
| University of Cambridge, Cambridge, United Kingdom
¶ Ronin Institute, Montclair, United States of America
# Institute for Globally Distributed Open Research and Education (IGDORE), Jena, Germany

Corresponding author: Daniel Mietchen (daniel.mietchen@ronininstitute.org)

Academic editor: Lyubomir Penev

Abstract

Ethical aspects of research continue to gain attention, be that in the process of proposing and planning research or performing, documenting or publishing it. One of the ways in which this trend manifests itself is the increasingly common addition of ethics statements to publications in fields like biomedicine, psychology or ethnography. Such ethics statements in publications provide the reader with a window into some of the practical yet typically hidden aspects of research ethics. As more and more publications are becoming available in full text and in machine readable formats through repositories like Europe PubMed Central, we propose to mine the literature for ethics statements and to extract information about the various aspects of research ethics that they address. The more standardized these statements are, the better the mined materials can be converted into structured and queryable information that can in turn be used to inform efforts towards higher levels of standardization in research ethics. This paper sketches out the motivation for such mining and outlines some methodological approaches that could be leveraged towards this end.

Keywords

text mining, Wikidata, ethics committees, ethics process, ethical approval

Introduction

Ethics is a key component of the way humans interact with each other and with their environments, including in research contexts. Research ethics provides a framework and guidance for making and evaluating decisions touching upon intellectual, social, legal,

practical, cross-cultural and other dimensions of research and the context in which it is situated (e.g. Bonde et al. 2015).

In some research fields - particularly those involving human subjects, animal experimentation, biodiversity or cultural heritage - the formalization of ethical norms and expectations has many decades of history (e.g. see Nelson 1967). This has led to detailed policies and guidelines that provide a framework for handling ethical issues and assisting compliance with applicable regulations (e.g. Yip et al. 2016, Childress and Thomas 2018). However, the norms may not be sufficiently standardized (e.g. Satalkar and Shaw 2013) in that they might lack clear practical implications like consistent incorporation into applicable workflows and cultural contexts, or simply uptake or proper communication (e.g. Murphy et al. 2015, Chiumento et al. 2020).

As formalization progresses, it tends to raise attention to ethical matters related to individual steps of research workflows, ranging from requesting ethical approval to documenting informed consent and providing ethics statements in funding applications or publications (e.g. Borovecki et al. 2018) .

Much of the process behind ethical review of research remains hidden (e.g. Vardigans et al. 2019) - for instance, it is rare to find public documentation of ethical approval (for an example, see Rothschild (2021)). This hampers the establishment of common standards and makes it difficult to teach or otherwise share practical aspects of research ethics across institutions, let alone jurisdictions. Furthermore, there is no mechanism through which ethics information can be looked up - let alone in a standardized fashion - for a given set of parameters, e.g. approval numbers (cf. Vardigans et al. 2019).

Overview of ethics statements

As illustrated in Fig. 1, ethics statements might contain information on a number of related matters (see JATS4R 2020 for best practice recommendations).

These frequently include

- the legal or policy basis for handling these issues on an international level (e.g. the Declaration of Helsinki) and/ or within a given jurisdiction or institution;
- the procedures followed to conform with these legal requirements, along with information about the role of key stakeholders in the process (e.g. approval by an ethics committee, or informed consent by donors and participants, or protocols for anonymization, or (parts of) organizations where the research was performed);
- the aspects of the research - if any - that pose ethical issues (e.g. acquisition of personally identifiable information, or animal experimentation or involvement of minors or prisoners).

This kind of information may assist others in engaging with the research that was performed, with the underlying methodology or the resulting data, with research projects of a similar nature or with education about matters related to said research.

While the majority of ethics statements refer directly to the research described in the respective publication, some such statements - particularly in certain types of reviews - refer to ethical aspects of cited publications, often summarizing the information for several of them using more generic phrases than in individual-article ethics statements. An example is given in Fig. 2 .

To ensure that ethics statements are present in publications when required by applicable policy or legislation, it is important that ethics-related information is available in a structured format to both humans and machine. This aim is in line with the FAIR principles (Wilkinson et al. 2016), whose application to ethical contexts (Mietchen et al. 2019) would imply that ethics-related information is

- **F** indable by everyone involved in the publishing process - authors and their co-authors as well as editors, reviewers, publishers and readers, along with any tooling that assists them in matching features of the reported research to relevant policy elements;
- **A** ccessible to the above stakeholders and their tool chains;
- **I** nteroperable across studies, institutions, journals, funders and others involved in research ethics workflows;
- **R** eusable in another context (e.g. a different clinical, geographic or demographic focus).

At present, FAIR information about ethics is an exception rather than a rule, and we argue that this should change if ethical aspects of research are to receive proper attention.

Once the ethics statements are present where they should be, another set of considerations revolves around standardization of these statements: are all necessary pieces of information present, and are they expressed in a way that allows them to be compared, aggregated, assessed for compliance with applicable policy or otherwise used across studies?

Here, several factors come into play, for instance

- Policy elements - what information is required by what part of which policy that is applicable to what aspect of the respective research;
- Checklists with standardized “boilerplate” language for each policy element;
- Machine actionability of these policy elements and their corresponding textual representations in the checklists.

Core ideas

In order to assist in the standardization of research ethics and associated documentation, we propose to do the following:

1. mine ethics statements from full-text articles using dictionaries (cf. Fig. 3) of seed terms and phrases as starting points and Europe PMC (EPMC) as an example repository;

2. extract associated entities (e.g. subject areas, policies, authorities or research facilities) and vocabulary (e.g. terms and phrases related to handling informed consent or incidental findings);

3. assess the degree to which the language or other aspects of these statements - e.g. their location within a publication - are already standardized;

4. reconcile the extracted entities and vocabulary terms with Wikidata items and lexemes;

5. prototype and facilitate the creation of open infrastructure and automated workflows that allow to look up and query information about the research ethics landscape in general as well as ethics approvals in particular, along with the corresponding processes, standards, entities and vocabulary.

Below, we will outline some use cases and practical steps towards implementing these ideas.

Use cases

Ethics statements contain information about ethical aspects of the research reported in the respective manuscript. Having straightforward access to such information may assist readers in engaging with said research or with research projects of a similar nature. Possible scenarios here include researchers wanting to pool their own data with that of the reported study, or wishing to repeat the study under slightly different conditions (e.g. involving a different demographic, location, time of the year or medical procedure). Other scenarios include patients or members of their social circles trying to find out about clinical trials to potentially enrol in, funders or institutions that wish to monitor compliance with their policies (e.g. as per Rasberry and Mietchen 2021), or students or instructors engaged in education about research ethics.

If the relevant information in the ethics statements were available in a standardized fashion, this would allow for it to assist discovery in such scenarios. For instance, the terms used there or the relationships between them could be reused for parameterizing searches or for filtering search results. To achieve such standardization, communal language and ontologies or other forms of structured terminologies need to be created,

and the process of creating them in turn assists in forming, strengthening or otherwise engaging such communities.

Methods

To demonstrate the feasibility of implementing the core ideas presented here, this section provides some methodological background, focusing on workflows that we used for prototyping.

Databases

Europe PMC

The full text of many biomedical articles is available via the literature repository PubMed Central (PMC) and its partner sites like EPMC. The articles can be accessed in several formats, usually including HTML, XML and PDF. Particularly suitable for mining is the XML format, which follows the Journal Article Tag Suite (JATS) specifications. JATS formally supports a wide range of section types and includes provisions for ethics statements. Much of the PMC and EPMC content predates both the current JATS version 1.3 and the dedicated recommendations (JATS4R 2020) for the tagging of ethics statements, while even the newest publications do not always follow these standards. All of this leads to some variation in the XML structures encountered when mining (for details, see the section on Information retrieval). Similarly minable full-text archives exist for other disciplines or with a particular institutional or geographic focus (e.g. SciELO, which is focused on Latin America).

Wikidata

To ensure that key elements of ethics statements are discoverable at scale by interested people, organizations or their tools, these elements need to be integrated into a coherent environment that is aware of the communal conventions and that can be curated by relevant communities. One platform that meets these criteria is Wikidata - a sister project to Wikipedia that can be considered the edit button for the semantic web. Wikidata hosts public domain data from across multiple domains of knowledge about a wide range of entities (referred to as items, of which there currently are about 100 million). These items are semantically annotated by a global community of thousands of curators using information extracted from reliable sources, including scholarly publications and thousands of databases. Due to their breadth of coverage, their granularity, ease of use and the broad integration with other resources, Wikidata items have great potential to assist in the identification of entities encountered in text mining.

Besides items - which are defined in a largely language-agnostic way, Wikidata has begun to build a similarly annotated collection of terms and phrases (referred to as lexemes, of which there currently are about half a million) that the World's languages use

to describe the underlying concepts, and it keeps track of semantic relationships between the items and lexemes. We thus propose to make the information mined from ethics statements available via Wikidata by curating the Wikidata entries for the respective items and lexemes and named entities.

Software pipeline

Software for accessing Europe PMC and similar repositories exists in several programming languages. We chose here to develop a Python-based pipeline that builds on a software suite originally implemented in Java a few years back and currently being developed as a tool called *docanalysis* (Hegde et al. 2022) to mine the literature for specific content like ethics statements. Fig. 4 provides an overview of the pipeline.

We will discuss this process on the basis of the example use case of extracting information about ethics committees. However, the approach can be generalized to extracting other information, be that related to ethics - e.g. approval numbers, consent types, applicable policies and guidelines - or beyond, e.g. data availability (cf. Colavizza et al. 2020) or conflict of interest statements.

Scraping (Step 1)

First, we use *pygetpapers* (Garg et al. 2022) to identify suitable articles and to download their XML. It is available from GitHub under the Apache-2.0 License and can also be installed via the *pip* or *conda* package managers for Python. *pygetpapers* is a tool to query scientific repositories - specifically Europe PMC - and many pre-print platforms such as bioRxiv, medRxiv, Rxivist and others. It returns metadata, and if available, fulltext and other data. In our project, we use *pygetpapers* to query EPMC to download papers in JATS XML.

Sectioning and Information retrieval (Step 2 and Step 3)

Next, we use *docanalysis* to decompose each article's XML into sections that can be analyzed independently. We can split the downloaded papers into sections based on the JATS tagging. Some of the section headings are predetermined (e.g. `abstract`) but most others (such as subsections and paragraphs) are determined by the author, journal or publisher.

Ethics statements are normally contained within a single paragraph (some with only one or two sentences. There are two main methods of retrieving these:

- **Context:** The statement is surrounded by metadata such as a subsection title <sec title="ethics"> (the JATS4R recommendation (JATS4R 2020) is to use <sec sec-type="ethics-statement">). Papers have different levels of nesting and we must use "globbing" and wildcarding, such as **glob("**/sections/**/sec[contains(@title,'ethics')]**

- **Content:** The language of the statement is clearly related to some of the entries in our dictionary, e.g. “The project was approved by the IRB of X University”. This requires natural language processing (NLP) and/or machine learning (ML) for text classification. The dictionaries contain phrases like “approved by … IRB" and so on, which can be used to filter the relevant sentences.

Sometimes, both methods are required; context to find the relevant paragraphs and content to find the relevant sentence(s).

Information extraction (Step 4 and Step 5)

To extract information on ethics committees from the sentences/sections we previously retrieved, *docanalysis* is using libraries like spaCy that provide techniques like unsupervised Named-Entity Recognition (NER - see recent review by Sharma et al. 2021). We then create a dictionary (cf. Fig. 3) of the recognized entities, which can be either used for curation or annotation of the relevant literature. Such dictionaries can be created in several ways, e.g. based on sample text and/ or based on Wikidata queries.

Sentences with phrases present in the ethics dictionary are selected, while other sentences are filtered out. The retained sentences are then parsed through spaCy, allowing to extract strings pertaining to ethics committees. These entities can then be added back into the ethics dictionary for more refined searches (cf. the section “Creating iterative feedback loops between the mining, curation and annotation of ethics statements” below).

Cataloguing the extracted information in Wikidata

After extracting the ethics committee information through NER, we can convert it to structured data. These data can then, for instance, be overlaid to the original text (e.g. as per Frei et al. 2022) or fed into Wikidata, where it can be curated and integrated further, particularly through initiatives like WikiProject Ethics, WikiProject Medicine or WikiProject Clinical Trials (see Rasberry et al. 2022 for an overview of the latter). On the basis of such community-curated structured data, queries can be written that expose this information, as illustrated in Table 1

Entity extraction using Wikidata can be further enhanced by incorporating information from corresponding Wikipedia entries (cf. Möller et al. 2022).

Creating iterative feedback loops between the mining, curation and annotation of ethics statements

An ontology of ethics committees and Institutional Review Boards (IRBs) can be created via Wikidata and used via the Wikidata SPARQL service. This ever-updating resource can then be used to aggregate and visualize ethics committee information extracted from the wider scientific literature. For instance, one could ask questions like which ethics

committees have approved a particular study, or studies on particular subjects, involving specific demographics, using particular interventions or funding sources.

Large search engines are usually optimised for terms and synonyms, not higher levels of concepts like “ethics”, and they often rely largely or even solely on metadata, which might well contain no information about the ethics process. In order to find statements about ethical aspects of a publication, it is hence necessary to analyze its full text.

In subsequent rounds of mining, information from Wikidata can be used to finetune the entity recognition, e.g. by providing terms to be included in the dictionaries used for mining, or by providing context for entity disambiguation. For instance, geoinformation can be used to distinguish between Calvin University in South Korea and Calvin University in the United States. Further synonyms can frequently be resolved in a straightforward fashion: “X University” often maps to “University of X”, though for a small group of X (Wikidata knows 7 examples), both might exist as separate entities, either in close proximity (as is the case for Hyogo or Shizuoka), at different places within the same country (e.g. Rochester, Jinan, Miami), in neighbouring countries (Ottawa) or continents apart (York).

For common words, we may need stemming (“approved” => “approv~”) or more generally lexemes (“X is grateful” or “we are grateful”) => “X <be> grateful”. Modern NLP tools can now identify such phrases from their context with high confidence. Wikimedia has an active lexeme project which can resolve lexical forms and map them to concepts, e.g. the English terms “ethics committee” and “informed consent form” are represented by the Wikidata lexemes L497553 and L497589, respectively. These lexeme entries in turn link information about these English nouns, their grammar and meaning to information about the underlying concepts (e.g. Q59057226 for “ethics committee” as a subclass of committee) as well as equivalent terms in other languages, which can also occasionally be found in ethics statements.

For instance, Fig. 5 shows that the German noun for ethics committee, *Ethikkommission*, (known to Wikidata as L562403) is used in ethics statements, both in articles written in German - e.g. Nuessle et al. (2021), Gugatschka et al. (2021) - as well as in English - e.g. Krajka et al. (2021), Leuenberger et al. (2021).

Complementing these mono- and bilingual examples, Fig. 6 gives an example in which several ethics committees are listed using both a name in their original languages and an English-language equivalent.

Taking such cross-linguistic information into account can thus facilitate entity recognition in ethics statements even in English texts and help expand the methodology to mining articles in other languages as well, e.g. to identify or distill boilerplate phrases in a given language or cultural differences across languages in terms of how ethics-related information is handled. For any language with information about such boilerplate phrases, a score could be computed that could represent the similarity between boilerplate text and phrasing from a given article. Such scores could be used, for

instance, to guide community curation efforts - high similarity to known boilerplate means high potential for automation and less need for human oversight, while low similarity indicates a need for community review.

Ethics statements as a less explored use case for testing text mining approaches

The extraction of ethics statements is a special case of a more general requirement. Many such statements are formulaic, either because the discipline itself or the publication process requires it. Typically, these articles have paragraphs where the sentences are discrete and not part of a larger narrative flow. A simple test for this is whether the sentences

1. can have their order shuffled without much loss of meaning or
2. make little use of anaphoric pronouns like "it" for linking sentences (e.g. “X was converted to Y. It was then converted to Z.” - “It” is meaningless without its precedent).

Looking beyond ethics statements, we have explored the range of syntactically similar sentences – frequently including boilerplate, named entities and perhaps identifiers like ethical approval numbers – and created a non-exhaustive list of manuscript components where they can frequently be found:

- acknowledgements and thanks;
- methods sections;
- availability and location of data and software;
- roles of authors and their contributions;
- conflict of interest statements;
- copyright statements.

The pipeline and the tools we are developing can extract semantic information from all such syntactically constrained sections of the scientific literature – not just ethics statements.

Standardizing the ethics statements

Irrespective of the textual representation and of JATS-style document markup, we posit that the factual elements of all ethics statements can be arranged to fit a grammar that relates the entities and is decomposable to a set of semantic triples. If true, this means that ethics statements can be formally encoded by authors as a graph and captured in a

graph knowledge base. This graph would then be queryable by standard tools such as SPARQL. Typical examples might be:

- <proposal> <*was approved by*> <approving body>
- <approving body> <*is part of*> <institution>
- <proposal> <*about*> <research project>
- <proposal> <*uses*> <methodology>
- <research project> <*has participant*> <group of patients>
- <group of patients> <*has condition*> <condition>
- <group of patients> <*has age range*> <...>

The entities and the predicates linking them would be mapped to standard identifier systems, including Wikidata, which is integrated with many of the key resources in this space. For instance, ethics-related terms that have a MeSH Descriptor - e.g. ethical review, ethics committee, animal care committee, informed consent and consent form, or the Declaration of Helsinki - all have a Wikidata entry, as do related terms that do not have a MeSH Descriptor, e.g. ethical approval, ethical oversight, or the Nagoya Protocol. Good coverage of ethics-related terms can also be found in the Informed Consent Ontology.

In the future, an increased level of curation of such information could be used to enhance ethics mining efforts. Ideally, authors could, with help from an authoring tool, submit their ethics statement as a formal graph representation. One approach would be a public site which parses manuscript snippets and assists its users in mapping them to triple-based standardized statements about ethical aspects of one or more manuscripts. Assuming a user-friendly implementation, we hypothesize that authors would be prepared to accept a standard form of language that could also be machine-parsed.

The information curated this way could also be used to search more systematically for the context in which ethics-related information occurs (cf. Information Retrieval section), i.e. the more standardized language could be used as a lexical hook to fish for similar snippets elsewhere, then regularize them and ultimately collate and analyze the bulk information.

Mapping the relevant terms creates a valuable positive feedback process between miners, corpora and open resources like the Wikimedia platforms. In some cases, Wikidata is well equipped with synonyms but at present, the entries are often stubs with very little information. The snowballing process will generate possible synonyms which can be collected together and offered in tools like Mix'n'Match for human editors to submit to Wikidata, or in tools like Drnote (Frei et al. 2022) to overlay Wikidata annotations on the original texts.

Discussion

In this work, we outlined a set of core ideas for mining the literature, extracting ethics-related entities and relationships, reconciling them with a controlled vocabulary, making

the information queryable and creating a positive feedback loop between the structured information and the mining workflows by iteratively using one to improve the other.

Much like in other areas of data mining, initial challenges for the mining of ethics statements include handling inconsistent approaches to the naming of relevant entities (e.g. institutions, ethics committees, laws and other relevant policy frameworks). This is compounded by inconsistency as to where in a document the ethics statements are located (e.g. in a dedicated section, or as part of the Methods or in an Annex).

If these challenges can be addressed, the mining of ethics statements can provide significant value in terms of elucidating the research ethics landscape (highlighting relevant organizations, along with policies, guidelines and other standardization efforts) as well as documenting, improving, teaching and standardizing current practices in research ethics. A systematic analysis of the ethics statements will also highlight institutional, disciplinary and other contexts in which such statements are common or well-developed, uncommon or underdeveloped, or anywhere in between.

This can form the basis for studying ethical aspects of the research process - as well as ethics review - under specific conditions and for addressing ethical aspects of research both in practice as well as in teaching. For instance, key elements of contexts in which well-developed ethics statements are common - such as a clear policy, readily actionable community guidelines or scalable workflows - could serve as a starting point for exploring best practices or synthesizing recommendations, while other contexts could be explored in terms of their potential for improvements.

Another point to consider is that access to the ethics-related information contained in a publication currently requires access to the full text. However, the basic ethics data - such as whether the research reported in the publication received ethical approval, what the approving bodies were and what the relevant approval numbers are - should be considered metadata and in the public domain. Ideally, they would be incorporated into the filtering mechanisms provided by individual databases or scholarly search engines and visualization tools more generally. Some databases like stem cell registries (cf. Kurtz et al. 2022) already provide ethics-related information.

Outlook

We plan to work towards implementing the core ideas presented here, and we very much welcome collaborations in this regard.

In particular, we plan to extract information and phrasing pertaining to ethics committees and other entities commonly found in ethics statements (e.g. policies and guidelines) and to make this information available via suitably annotated Wikidata items and lexemes that can in turn be used by mining pipelines. Once the data models in this area have stabilized, it would be possible to scale up these workflows by increasing their automation and expanding the mining to auxiliary materials like approval letters, which are currently shared only very rarely , or to annotating ethical aspects of things other than

formal publications, e.g. clinical trials or their consent forms that are now increasingly being made public too.

Further, we plan to work on visualizations that present this structured information and that can be incorporated into suitable parts of the open knowledge ecosystem, particularly through Wikimedia platforms and associated visualization services like Scholia (Nielsen et al. 2017, Lemus-Rojas et al. 2022).

Beyond ethics statements, we plan to apply the ideas outlined here also to other non-traditional parts of research manuscripts, e.g. data availability or conflicts of interest. We also aim to explore how these approaches can assist with the enrichment of mining efforts targeted at less-mined aspects of manuscripts, e.g. the citation of data, software and material resources. In doing so, we will focus on resources that are openly available.

Funding program

Neither the work proposed nor the work presented here has so far received funding.

Conflicts of interest

The authors declare that they have no conflicts of interest pertaining to the research described here.

References

- Bonde S, Briant C, Firenze P, Hanavan J, Huang A, Li M, Narayanan NC, Parthasarathy D, Zhao H (2015) Making Choices: Ethical Decisions in a Global Context. Science and Engineering Ethics 22 (2): 343-366. https://doi.org/10.1007/s11948-015-9641-5
- Borovecki A, Mlinaric A, Horvat M, Supak Smolcic V (2018) Informed consent and ethics committee approval in laboratory medicine. Biochemia Medica 28 (3). https://doi.org/10.11613/bm.2018.030201
- Childress A, Thomas C (2018) Navigating the Perfect Storm: Ethical Guidance for Conducting Research Involving Participants with Multiple Vulnerabilities. Kennedy Institute of Ethics Journal 28 (4): 451-478. https://doi.org/10.1353/ken.2018.0025
- Chiumento A, Rahman A, Frith L (2020) Writing to template: Researchers' negotiation of procedural research ethics. Social Science & Medicine 255 https://doi.org/10.1016/j.socscimed.2020.112980
- Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B (2020) The citation advantage of linking publications to research data. PLOS ONE 15 (4). https://doi.org/10.1371/journal.pone.0230416
- Cui Z, Wang D, Wang W, Zhang Y, Jing B, Xu C, Chen Y, Qi M, Zhang L (2021) Occurrence and Multi-Locus Analysis of Giardia duodenalis in Coypus (Myocastor coypus) in China. Pathogens 10 (2). https://doi.org/10.3390/pathogens10020179

- Frei J, Soto-Rey I, Kramer F (2022) DrNote: An open medical annotation service. PLOS Digital Health 1 (8). https://doi.org/10.1371/journal.pdig.0000086
- Garg A, Smith-Unna RD, Murray-Rust P (2022) pygetpapers: a Python library for automated retrievalof scientific literature. Journal of Open Source Software 7 (75). https://doi.org/10.21105/joss.04451
- Gugatschka M, Grossmann T, Hortobagyi D (2021) Molekulare Laryngologie. HNO 69 (9): 695-704. https://doi.org/10.1007/s00106-021-01016-1
- Hegde S, Garg A, Sharma C, Murray-Rust P, Singha A, Faria E (2022) docanalysis. Zenodo 0.1.9 https://doi.org/10.5281/zenodo.7063887
- JATS4R (2020) NISO JATS4R Ethics Statements v1.0. NISO https://doi.org/10.3789/niso-rp-33-2020
- Krajka V, Naujock M, Pauly M, Stengel F, Meier B, Stanslowsky N, Klein C, Seibler P, Wegner F, Capetian P (2021) Ventral Telencephalic Patterning Protocols for Induced Pluripotent Stem Cells. Frontiers in Cell and Developmental Biology 9 https://doi.org/10.3389/fcell.2021.716249
- Kurtz A, Mah N, Chen Y, Fuhr A, Kobold S, Seltmann S, Müller S (2022) Human pluripotent stem cell registry: Operations, role and current directions. Cell Proliferation 55 (8). https://doi.org/10.1111/cpr.13238
- Lemus-Rojas M, Odell J, Brys LF, Ramirez Rojas M (2022) Leveraging Wikidata to Build Scholarly Profiles as Service. KULA: Knowledge Creation, Dissemination, and Preservation Studies 6 (3): 1-14. https://doi.org/10.18357/kula.171
- Leuenberger A, Winkler M, Cambaco O, Cossa H, Kihwele F, Lyatuu I, Zabré H, Farnham A, Macete E, Munguambe K (2021) Health impacts of industrial mining on surrounding communities: Local perspectives from three sub-Saharan African countries. PLOS ONE 16 (6). https://doi.org/10.1371/journal.pone.0252433
- Mietchen D, Schwaiger V, Beyan O (2019) FAIR Ethics: Making Ethical Review Processes more Machine Actionable. Zenodo https://doi.org/10.5281/zenodo.2559997
- Möller C, Lehmann J, Usbeck R (2022) Survey on English Entity Linking on Wikidata: Datasets and approaches. Semantic Web1-42. https://doi.org/10.3233/sw-212865
- Murphy S, Nolan C, O'Rourke C, Fenton JE (2015) The reporting of research ethics committee approval and informed consent in otolaryngology journals. Clinical Otolaryngology 40 (1): 36-40. https://doi.org/10.1111/coa.12320
- Nelson B (1967) Anthropologists Overwhelmingly Approve Research Ethics Statement. Science 156 (3773): 365-365. https://doi.org/10.1126/science.156.3773.365
- Nielsen FÅ, Mietchen D, Willighagen E (2017) Scholia, Scientometrics and Wikidata. Lecture Notes in Computer Science237-259. https://doi.org/10.1007/978-3-319-70407-4_36
- Nuessle S, Luebke J, Boehringer D, Reinhard T, Anton A (2021) Akuter Winkelblock. Medizinische Klinik - Intensivmedizin und Notfallmedizin 117 (2): 137-143. https://doi.org/10.1007/s00063-021-00790-8
- Rasberry L, Mietchen D (2021) Monitoring policy compliance with Wikidata and Scholia. Zenodo https://doi.org/10.5281/zenodo.5594662
- Rasberry L, Tibbs S, Hoos W, Westermann A, Keefer J, Baskauf SJ, Anderson C, Walker P, Kwok C, Mietchen D (2022) WikiProject Clinical Trials for Wikidata. medRxiv https://doi.org/10.1101/2022.04.01.22273328
- Rothschild D (2021) Data Quality of Platforms and Panels for Behavioral Research. Open Science Framework https://doi.org/10.17605/osf.io/342dp

- Satalkar P, Shaw D (2013) Not Fit for Purpose: The Ethical Guidelines of the Indian Council of Medical Research. Developing World Bioethics 15 (1): 40-47. https://doi.org/10.1111/dewb.12036
- Sharma A, Amrita, Chakraborty S, Kumar S (2021) Named Entity Recognition in Natural Language Processing: A Systematic Review. Proceedings of Second Doctoral Symposium on Computational Intelligence817-828. https://doi.org/10.1007/978-981-16-3346-1_66
- Vardigans C, Malloy M, Meynell L (2019) Breaking barriers to ethical research: An analysis of the effectiveness of nonhuman animal research approval in Canada. Accountability in Research 26 (8): 473-497. https://doi.org/10.1080/08989621.2019.1684906
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18
- Yang W, Akhtar S, Franek E, Haluzík M, Hirose T, Kalyanam B, Kar S, Wu T, Gogas Yavuz D, Unnikrishnan AG (2022) Postprandial Glucose Excursions in Asian Versus Non-Asian Patients with Type 2 Diabetes: A Post Hoc Analysis of Baseline Data from Phase 3 Randomised Controlled Trials of IDegAsp. Diabetes Therapy 13 (2): 311-323. https://doi.org/10.1007/s13300-021-01196-7
- Yip C, Han N, Sng B (2016) Legal and ethical issues in research. Indian Journal of Anaesthesia 60 (9). https://doi.org/10.4103/0019-5049.190627

This study was performed with strict adherence to the recommendations of the <legal basis>Guide for the Care and Use of Laboratory Animals of the Ministry of Health, China</legal basis>. The research protocol was <boilerplate>reviewed and approved by</boilerplate> the <ethics committee>Research Ethics Committee of Tarim University</ethics committee>(approval no. <approval number>ECTU 2018-0026</approval number>). Farm owners <boilerplate>gave permission</boilerplate> before we commenced <ethics trigger>fecal sample collection</ethics trigger>.

Figure 1.

Ethics Statement from Cui et al. (2021) with putative markup of some key elements. Colors indicate the legal basis (pink), some boilerplate language pertaining to ethical review, approval and permissions (purple), oversight body (yellow) and approval number (green) as well as the aspect of the research that triggered the need for ethical oversight (grey).

The individual trials considered for this post hoc analysis were approved by health authorities according to the corresponding local regulations and by the local independent ethics committees. These trials were conducted in accordance with the Declaration of Helsinki [32] and Good Clinical Practice Guidelines. All participants provided written informed consent prior to enrolment into the respective trials [19–31].

Figure 2.

Ethics statement from Yang et al. (2022), a meta-analysis, with putative markup of some key elements. Color code as above. The language is more generic overall than in Fig. 1.

```
<?xml version="1.0" encoding="UTF-8"?>
<dictionary title="ethics_committee_key_phrases">
 <entry term="animal care and use committee of " name="animal care and use committee of "
id="CM.ethics_committee_key_phrases.0"/>
 <entry term="ethical committee of " name="ethical committee of "
id="CM.ethics_committee_key_phrases.1"/>
 <entry term="ethics committee of " name="ethics committee of "
id="CM.ethics_committee_key_phrases.2"/>
 <entry term="iacuc of " name="iacuc of " id="CM.ethics_committee_key_phrases.3"/>
 <entry term="institutional animal care and use committee of " name="institutional animal
care and use committee of " id="CM.ethics_committee_key_phrases.4"/>
 <entry term="institutional review board of " name="institutional review board of "
id="CM.ethics_committee_key_phrases.5"/>
 <entry term="the irb of " name="the irb of " id="CM.ethics_committee_key_phrases.6"/>
…
</dictionary>
```

Figure 3.

An example dictionary for text mining, containing various seed terms in a structured format that can be easily expanded. Each entry consists of three parts:

- (left): a string found or to be found in the mined texts (this part is mandatory)
- (center): a human-readable name for the string (this is optional)
- (right): an identifier for the string (still optional, but highly recommended)

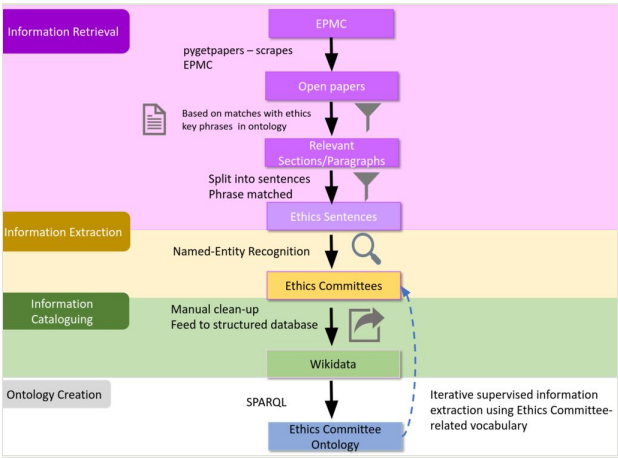

Figure 4.

Ethics Statements mining pipeline. Works identified through a search query are being retrieved in full text, the text is then searched for key terms from the ethics dictionary to identify ethics-related article sections, which are then partitioned into sentences that are parsed to try to identify named entities. The results of the mining can be compared to entities and terms known from Wikidata and/ or the dictionaries, which can be continuously improved in an iterative process that can lead to a controlled vocabulary and eventually an ontology for ethics statements, ethics committees and related concepts.

Figure 5.

The German term for ethics committee - *Ethikkommission* - used in the context of documents otherwise written in English (top, from Krajka et al. 2021) or German (middle, from Nuessle et al. 2021, and bottom, from Gugatschka et al. 2021).

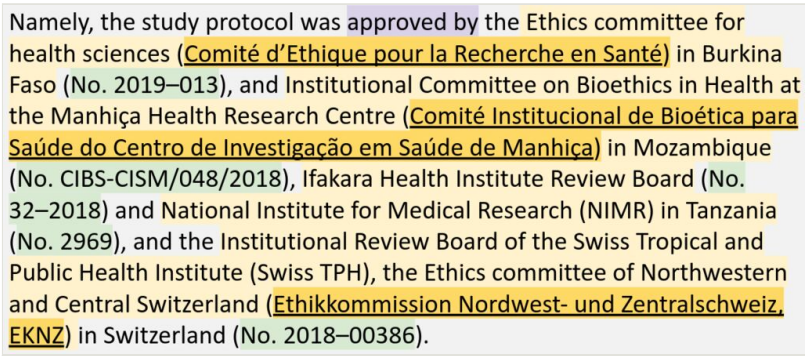

Figure 6.

Examples for entities in English text (highlighted in peach), with text elements in multiple languages (underlined, darker shade): French, Portuguese, German.

Table 1.

Part of the result of a Wikidata query for ethics committees. *committee* stands for the Wikidata entry for a given ethics body, and *committeeLabel* for the corresponding label in English. To access the live results, use https://w.wiki/4$GC. Such queries can be refined further, e.g. to enrich the above list with examples of research approved by these committees, to get a list of publications with information about the ethics bodies that have approved the underlying research or a list of topics for which publications have reported ethical approval. Most of the current entries in the list were the result of testing our pipeline, so the information associated with them is often minimal. However, once these entries exist and are linked to other entries (e.g. for the parent organization), they become part of the community curation workflows on Wikidata, which can in turn enrich the mining efforts over time.

committee	committeeLabel
http://www.wikidata.org/entity/Q94657657	Ethics Committee of the American Society for Reproductive Medicine
http://www.wikidata.org/entity/Q107345824	Ethics Committee of the University of Debrecen
http://www.wikidata.org/entity/Q107561623	Cambridge Local Research Ethics Committee
http://www.wikidata.org/entity/Q107561531	Institutional Review Board of Fujita Health University
http://www.wikidata.org/entity/Q107561540	Institutional Review Board of the Chulalongkorn University Faculty of Dentistry
http://www.wikidata.org/entity/Q107561629	Ethics Committee of University Hospital Hradec Kralove
http://www.wikidata.org/entity/Q107561627	People's Hospital Ethics Committee
http://www.wikidata.org/entity/Q106580821	Research Ethics Committee of Galway University Hospitals
http://www.wikidata.org/entity/Q107172445	Beaumont Hospital Ethics Committee
http://www.wikidata.org/entity/Q107306694	Hartford Hospital Ethics Committee
http://www.wikidata.org/entity/Q107306814	Scotland A Research Ethics Committee
http://www.wikidata.org/entity/Q107417881	Biobanks Ethics Committee of the University of the Witwatersrand
http://www.wikidata.org/entity/Q107561634	Committee for Ethics in Research of the University of São Paulo
http://www.wikidata.org/entity/Q105725690	National Ethics Committee of Senegal
http://www.wikidata.org/entity/Q107417865	Human Research Ethics Committee (Non-Medical) of the University of the Witwatersrand
http://www.wikidata.org/entity/Q107417840	Saint Barnabas Medical Center Institutional Review Board

http://www.wikidata.org/entity/Q107429572	Inrae-Cirad-Ifremer-Ird joint ethics advisory committee
http://www.wikidata.org/entity/Q107561539	Institutional Review Board of Sanyo-Onoda City University
http://www.wikidata.org/entity/Q107561635	Emirates Institutional Review Board for COVID-19 Research
http://www.wikidata.org/entity/Q107561625	Local Ethics Committee of Medical University of Silesia