Enabling Social Demography Research using Semantic Technologies

Lise Stork[1 [0000−0002−2146−4803]], Richard L. Zijdeman[2 [0000−0003−3902−3720]], Ilaria Tiddi[1 [0000−0001−7116−9338]], and Annette ten Teije[1 [0000−0002−9771−8822]]

[1] Vrije Universiteit Amsterdam
[2] International Institute of Social History

Abstract. A shift in scientific publishing from paper-based to knowledge-based practices promotes reproducibility, machine actionability and knowledge discovery. This is important for disciplines like social demography, where study indicators are often social constructs such as race or education, hypothesis tests are challenging to compare due to their limited temporal and spatial coverage, and research output is presented in natural language, which can be ambiguous and imprecise. In this work, we present the MIRA resource, to aid researchers in their research workflow, and publish FAIR findings. MIRA consists of: (1) an ontology for social demography research, (2) a method for automated ontology population by prompting Large Language Models, and (3) a knowledge graph populated in terms of the ontology by annotating a set of research papers on health inequality. The resource allows researchers to formally represent their social demography research hypotheses, discovering research biases and novel research questions.

Keywords: Scientific Knowledge Graphs · Social Demography · Hypothesis Representation · Health Inequality · Information Extraction

Resource type: Ontology, Knowledge Graph

DOI: `https://doi.org/10.5281/zenodo.10286846`
URL: `http://w3id.org/mira`, `http://w3id.org/mira/ontology/`

1 Introduction

Research on social demography focuses on the statistical study of human populations, with the aim of understanding and predicting social, cultural and economic trends across populations. Since the first known census taken by the Babylonian Empire in 3500 BCE [25], demographers have involved themselves with the task of explaining aggregate statistics of a population [14]. Specifically, the research cycle of a social historian consists of analysing observational data about society to form novel hypotheses and theories about societal mechanisms (see left side of Figure 1). However, such studies tend to be restricted to specific time periods and regions, making comparison of hypothesis tests across the vast array of research

papers difficult. Moreover, demographers test research hypotheses in which variables are often social constructs such as intelligence, ownership, or nationality. Unclarity and ambiguity reporting research findings in natural language complicates their precise understanding. Due to these complexities, demographers often describe rather than explain demographic phenomena [14], whereas causal explanations could help shape better policies for the future. Below we describe a common motivating scenario.

Example 1 *A social historian aims to analyse a population census from the Netherlands between 1850-1922—consisting of certificates (births and marriages), occupational and survival data*[3] *—to better understand mechanisms of social inequality. For this, the historian has to carefully survey the literature for known theories and the datasets and statistics that support them.* What was the evidence for a specific hypothesis? How did they measure social stratification? What were the outcomes? Which societal factors, that are (not) in my dataset, can moderate the effect between my study variables? *The historian may find it challenging to discover relevant papers or interpret research findings precisely.*

To address such challenges, the paradigm of scientific publishing is seeing a shift from document-oriented publishing to knowledge-based publishing [1,20], with the aim of making the infrastructure for scientific publication of research output more FAIR (Findable, Acessible, Interoperable, and Reusable). Semantic technologies have been successfully employed in various domains to accommodate such a shift. Notable examples are the Open Science Knowledge Graph [1,31], the Unified Medical Language System (UMLS) [2], or Biomedical knowledge graphs such as [17], but many other examples exist [32,8,10,4,33,7,8,22,21]. These sources promote reusability of research findings, as well as the machine-aided design of novel research questions. Moreover, by adopting formal representations of knowledge, such resources promote transparency and explainability.

In social demography, work has been done formally describing observational data, such as census data hosted as linked data at the International Institute of Social History (IISH)[4] [24]. To the best of our knowledge, no studies formalise knowledge such as *hypotheses* and *findings* on social demography–and very few from social sciences in general [32,22], whereas these are important in each of the steps of the scientific workflow of a social historian (see Figure 1). The research process, hypotheses and findings are mostly written up in scientific documents in natural language, which can be ambiguous and imprecise. Such fields can thus benefit from adopting the FAIR data principles, to reduce unclarity and ambiguity in the research workflow of a social demographer.

In summary, there is a need for the improvement of the digital infrastructure underlying scientific publication in social demography and social history research, to stimulate a deep understanding, and reuse of existing hypotheses, methods and findings. To address this need, this work provides:

[3] for example: `https://datasets.iisg.amsterdam/dataverse/HSNDB-HSN`
[4] `https://iisg.amsterdam/en`

1. The *MIRA ontology*, which includes a set of classes, properties and axioms for capturing research findings (*observations, comparisons and explanations*) on social demography, as well as SHACL shapes following data quality criteria of [5], for data validation.
2. A Knowledge Graph Construction (KGC) method, based on: (i) prompting a Large Language Model to annotate paper abstracts using the ontology, (ii) mapping concepts to terms from NCBO BioPortal ontologies and GeoNames, and (iii) refining the final graph by a set of SHACL constraints, developed according to data quality criteria.
3. The *MIRA-KG*, a knowledge graph of machine-annotated paper abstracts on social health inequality in terms of the MIRA ontology. Annotations are linked to Linked Open Data. The resource is published on the druid datalegend database infrastructure[5], maintained by the international institute of social history (IISH) and Triply[6].

In general, this work (i) supports the shift towards knowledge-based scientific publishing by contributing a novel method for KGC construction that can easily be adapted to accommodate other application domains, and (ii) contributes to a more FAIR infrastructure for social demography research.

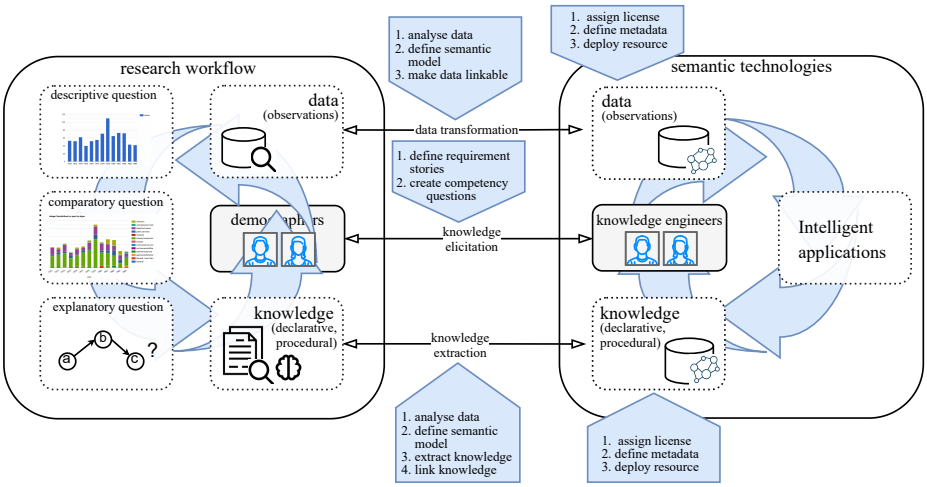

Fig. 1: The social demography research workflow. Blue boxes contain components of FAIR data management (e.g. for social history), which would enable social historians to make their research more FAIR.

2 Related Work

Ontologies and Vocabularies for Scientific Knowledge. Below we discuss examples of ontologies developed for various types of scientific knowledge:

[5] https://druid.datalegend.net/
[6] https://triply.cc/

Upper-level science ontologies. Examples of upper-level ontologies for scientific research are the Modern Science (ModSci) ontology [10], an upper ontology for modern science branches and related entities, and the Semantic-science Integrated Ontology (SIO) [9], a simple integrated ontology for rich description of scientific concepts and processes.

Hypothesis representations. Work has been done on formalising research hypotheses, and notable examples are the SuperPattern ontology [4], which allows researchers to write up the main claims of scientific articles as statements in formal logic; the nanopublication model [15], which aims at publishing "core scientific statements with associated context", or the DISK Hypothesis Evolution ontology [12], which captures the evolution of hypotheses over time. Many other representations exist, of which [12] provides an extensive comparative overview. An example of a domain-specific scientific ontology is the PICO ontology (for Population, Intervention, Comparator, Outcome) for synthesis and querying of clinical trial experiments [23].

Data representations. The RDF Data Cube vocabulary[7] allows researchers to publish their multi-dimensional data, and an example of a domain-specific model for publishing datasets is the the Data Scopes model [18,3], which captures how datasets are processed in in social history research.

Publication metadata. Lastly, vocabularies have been created to capture publication metadata. The most notable ones are the bibliographic ontology (BIBO)[8], the PRISM vocabulary[9], and the Dublin Core[10].

In this work, we formalise research hypotheses and link them to data representations and publication metadata. We do so for the domain of social demography, which has not been done before, and make use of some of the ontologies and vocabularies enumerated above.

Scientific Knowledge Graphs. Knowledge Graphs that capture scientific knowledge have been created from unstructured texts such as research papers, in various domains like Computer Science and Artificial Intelligence [6,7,8], social science[29], biomedicine [17], plant sciences [21], social and behavioural sciences [22], or scientific articles in general [30,1,31]. Resources such as these support a variety of tasks that aid researchers in the development of their research questions and methods. Our work is most related to the work of [29] and [22], but instead proposes a semantic model of social demography hypotheses and findings, and a knowledge graph construction (KGC) method that does not rely on the knowledge-intensive task of expert semantic annotation and links extracted knowledge to other linked open data (LOD).

[7] `https://www.w3.org/TR/vocab-data-cube/`
[8] `https://www.dublincore.org/specifications/bibo/bibo/bibo.rdf.xml`
[9] `https://www.w3.org/submissions/2020/SUBM-prism-20200910/psv-over.html`
[10] `https://www.dublincore.org/`

3 Knowledge Graph Construction

This section describes the knowledge graph construction pipeline (see Figure 2 for the entire pipeline). *Ontology creation* is discussed in Section 3.1, the resulting ontology in 3.2, *ontology population* in Section 3.3, and the resulting KG in Section 3.4.

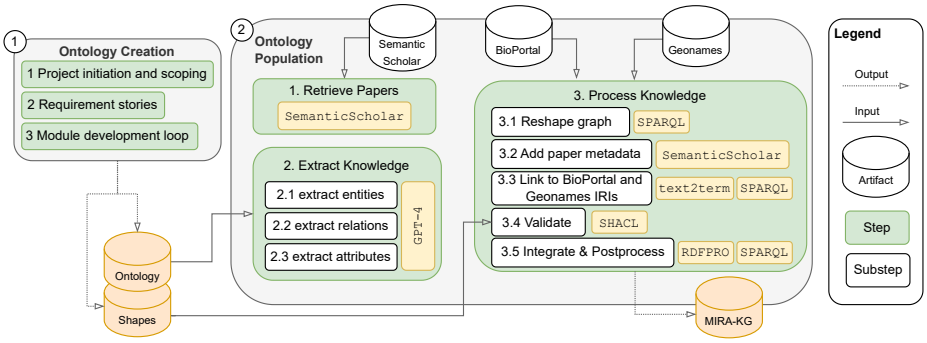

Fig. 2: **The knowledge graph construction** pipeline. Coloured artifacts are novel contributions; non-coloured artifacts represent external repositories.

3.1 Ontology Creation

Below, we describe ontology creation process (see ① of Figure 2). We follow the eXtreme Design (XD) method for ontology development [27], a collaborative, incremental, iterative method for pattern-based ontology design. We followed the main steps of the XD methodology: (i) project scoping, (ii) creation of requirement stories, (iii) a module development loop.

1. Project initiation and scoping. A first interview was set up with a domain expert on social history and humanities research from the International Institute of Social History (IISH)[11]. The outcome of the interview was used to define the task context, get a better understanding of the challenges, and write out a first list of ontology requirements (inspired by how a hypotheses is represented in DISK [12,13]): *Article metadata* (**RQ1**), *Hypothesis classification* (**RQ2**), *Hypothesis statement* (**RQ3**), *Hypothesis context* (**RQ4**), *Hypothesis qualifier* (**RQ5**), and *Hypothesis evidence* (**RQ6**). For **RQ2-6**, we collected requirement stories.

2. Requirement stories. Within this project, requirement stories were collected based on insights retrieved during the project scoping phase about hypothesis types and their elements. We show one requirement story below (Example 2), which shows an explanatory question (asking whether economic developments can explain differences in socioeconomic inequality).

[11] `https://iisg.amsterdam/en`

Example 2 *"Is the influence of ascription on education larger than the influence of achievement, and is the effect moderated by economic developments?"*

After a first elicitation round, academic papers were used as requirement stories. An example annotation, based on the paper "*Industrialization and inequality revisited: Mortality differentials and vulnerability to economic stress in Stockholm, 1878–1926*" [26] is used as an example in the visual schematic of the final ontology (indicated by ■), see figs. 3 to 5.

3. Module development loop. Ontological requirements were elicited by generalising terms from requirement stories and deriving competency questions (CQs), resulting in a categorisation of main entity types and relations derived from the requirement stories, see Table 1.

Table 1: Question Type (QT), Question Level (QL), Observation Type (OT), and Trend Type (TT) categorisations, based on requirement stories.

QT	QL	OT	TT
Descriptive Describing phenomena, e.g. *temperature is rising over time*	**Micro** *Individuals* *Human populations*	**Longitudinal** *Observations over time*	**Temporal trend** *Comparison of time intervals*
Comparative Comparisons between sample means, trends, or effects, e.g. *did temperature rise more after industrialisation?*	**Meso** *Social groups,* *Organisations*	**Intergenerational** *Observations over generations*	**Spatial trend** *Comparison of regions*
Explanatory Defining a potential cause of the outcome of a comparison, e.g. *did human emission cause the rise of temperature?*	**Macro** *Geographical regions*	**Intersectional** *Observations from a single point in time*	**Socioeconomic trend** *Comparison of distribution across social classes*

Moreover, contextual statements (CSs), and reasoning requirements (RRs) were derived. CSs informed the creation of SHACL constraints for validation, and RRs informed the creation of OWL[12] axioms. A collaborative environment in the form of a Wiki was set up to gather the CQs, CSs, RRs and other outcomes. A dump of the Wiki is published on Zenodo together with the dataset.

Ontology reuse. For representing article metadata (**RQ1**), we use BIBO, PRISM and Dublin Core. To classify hypotheses (**RQ2**), we use the classification (QT) from Table 1. Moreover, to write up hypothesis statements (**RQ3**) and their context (**RQ4**, based on (QL) from Table 1) we utilise the structure and properties of the SuperPattern Ontology [4]. However, we represent instances of the SuperPattern as individuals instead of classes. As in social demography, evidence often comes from specific time periods and regions, future work will explore the use of class axioms to collect all evidence belonging to a claim. We additionally reuse classes and properties from the Semanticscience Integrated Ontology (SIO) [33] (e.g. trend line `sio:SIO_000527` or human population `sio:SIO_001062`). For

[12] `https://www.w3.org/TR/owl-syntax/`

qualifiers (**RQ5**), we reuse qualifier categories from [16], as results can be translated to effect sizes. Lastly, to link hypotheses to their evidence (**RQ6**), we reuse the *RDF Data Cube Vocabulary*[13], furthering reproducibility and machine actionability.

Design Pattern (ODP) reuse. Scientific hypotheses and claims often consist of complex relations between variables, relations, and findings, and as such there are various ways of modeling them. One typical way is to reify the relation-holding context as a node with binary relations for the subject, object and property, and a fourth binary relation indicating the context [11]. A downside of such a representation is that the Web Ontology Language (OWL)[14] does not support reification. Some Ontology Design Patterns such as the Content Slices ODP[15] address the issue, but such modeling hampers easy integration with paper metadata. Therefore, we choose for *reification* with the idea that the reified hypotheses can later be transformed to binary relations, should OWL reasoning be required.

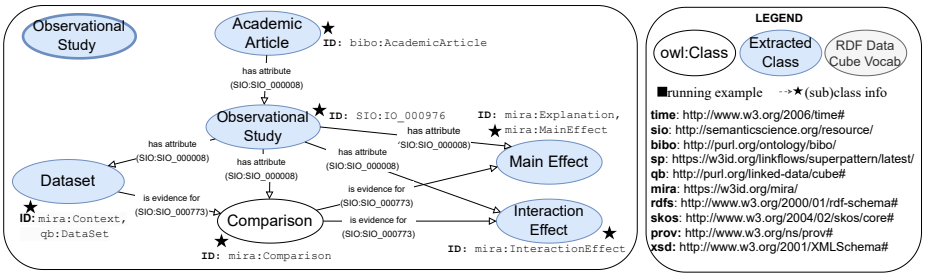

Fig. 3: Main classes of the MIRA ontology. A ★ points to the IRI of the class(es) or superclass. Elements retrieved from paper abstracts during the ontology population process (see Section 3.3) are indicated in blue. Commonly, paper abstracts from social demography do not contain statistical data and precise findings.

3.2 The MIRA Ontology

The main concepts, relations and their domains and ranges of the MIRA Ontology are created in light of our requirements and categorisations (Table 1)[16]. These are shown in Figure 3. Figure 4 unfolds the *comparison* branch of the MIRA ontology. Descriptive statistics (such as ratios) can be compared, and are linked to their respective data cube slices. Figure 5 unfolds the *explanation* (main effect and interaction effect) branch of the MIRA ontology.

[13] `http://www.w3.org/TR/vocab-data-cube/`
[14] `https://www.w3.org/TR/owl-ref/`
[15] `http://ontologydesignpatterns.org/wiki/Submissions:Context_Slices`
[16] Some categorisations (such as OT and CT) are omitted as these can be derived by querying dataset information (the `qb:sliceStructure` of a `qb:Slice`).

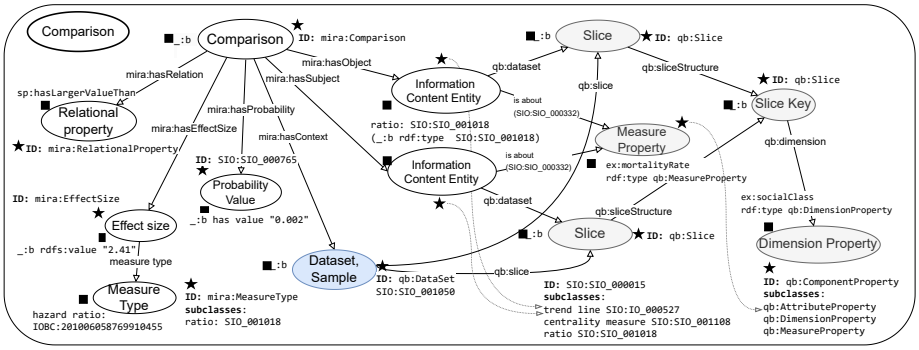

Fig. 4: The **Comparison** class and its links to the RDF Data Cube Vocabulary. Comparisons compare information content entities such as trend lines, centrality measures, and ratios, which are based on slices over multidimensional data cubes. Example instances and literals are based on the article *"Industrialization and inequality revisited: Mortality differentials and vulnerability to economic stress in Stockholm, 1878-1926"*, and indicated with ■ [26]

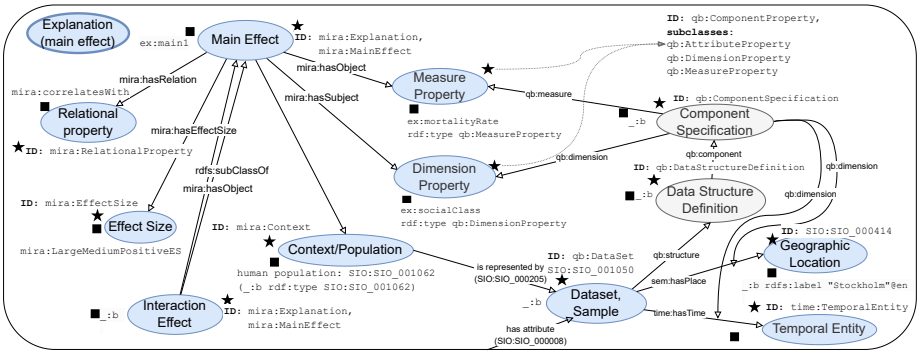

Fig. 5: The **Explanation (Main Effect)** class and its links to the RDF Data Cube Vocabulary. Explanations describe measured effects of dimensions in the dataset (such as `ex:socialClass`, instance of `qb:DimensionProperty`) on measurements in the dataset (instances of `qb:MeasureProperty` such as `ex:mortalityRate`). These properties can then be linked to concepts from ontologies using `qb:concept` (for example `ex:mortalityRate qb:concept STATO:STATO_0000414`). The `mira:correlatesWith` property is used as a subproperty of `mira:RelationalProperty`. No causal link is yet stated, since more evidence from various regions and periods is required to state a causal relationship. **Interaction Effects** are subclasses of main effects. They are modeled in the same way, with the following exceptions: the range of `mira:hasObject` is an instance of `mira:MainEffect`; the range of `mira:hasProperty` is `mira:moderates`.

From the CSs and RRs enumerated during the ontology design process, we developed a set of OWL axioms and SHACL constraints for

validation, which describe the semantic requirements of the hypotheses. An example of an OWL statement and axiom are: *is evidence for* is an `owl:transitiveProperty`, and `mira:InteractionEffect owl:subClassOf mira:Explanation` respectively. Examples of SHACL constraints are the use of `sh:minCount` and `sh:maxCount` to ensure that each paper has at least one study, each study has at least one explanation, each explanation has as exactly one context, subject, object, relation and qualifier.

3.3 Ontology Population

Below, we describe the ontology population process from scientific papers (See ② of Figure 2). Our method is related to other KGC methods such as [6], but explores prompt-based triple extraction as a single extraction step allows for fewer errors to be propagated to subsequent steps. Future work will compare such a method with other established KGC methods. As a use case, we focus on articles related to health inequality.

1. Retrieve papers. Papers included were retrieved from Semantic Scholar using the Semantic Scholar API, and selected by the following steps:

- retrieving papers between 2020-2023, based on the keywords: $\in$ {*social, inequality, disparities, mortality, population, socioeconomic, demographic, study*}, and the fields of study $\in$ {*Economics, History, Sociology*}. These keywords were terms encountered in the titles of the five papers used as requirement stories;
- filtering papers based on:
  - i *citation count:* removing articles with lower impact, with ≤ 10 citations;
  - ii *journal:* removing articles published in non peer-reviewed journals;
  - iii *abstract length:* including articles with abstract length $l = \mu - \sigma > l > \mu + \sigma$ (within plus-or-minus 1 standard deviation) as we found that it was easier to extract hypotheses from texts that were not too long nor too short due to conciseness and lack of information, respectively;
  - iv *doi:* whether or not articles included a Digital Object Identifier (DOI).

2. Extract Knowledge. In social demography, the abstract of a paper commonly includes the explanatory question and the time period and location of the population sample. We thus annotate paper abstracts with these parts of the ontology (see the blue classes in figs. 3 to 5). In order to extract the relevant entities, relations and attributes in terms of the ontology, we prompt GPT-4 (OpenAI, 2023), see Table 2. Large Language Models can learn multiple tasks without any explicit supervision [28,19] and can therefore take on the task normally performed through human annotators via crowdsourcing, without any training. Retrieved annotations were extracted directly as triples of the form $< subject, predicate, object >$ using the Resource Description Framework[17].

[17] `https://www.w3.org/RDF/`

Table 2: Prompt used to retrieve structured abstract annotations using GPT-4. Each prompt consisted of the paper abstract, the ontology (domains and ranges), and some instructions for formatting of individuals.

Prompt template	
< *abstract* > Describe the claim of the abstract above only using RDF (the turtle syntax), and using the following ontology: < *property* >< *domain* >< *range* > < *instructionnumber* >< *instruction* >< *example* > Only return proper RDF, no free text comments	
Instruction example	Answer
Instruction: use rdfs:label to describe all blank nodes, also the geographic region. Use short descriptions, pieces of text verbatim from the abstract, and add language tags to all labels. *Example*: [] :hasSubject [rdfs:label "social class"@en].	[ ] :hasSubject [rdfs:label "residential patterns and nutrition"@en]
Instruction: for instances of the class gn:Feature, find the URI for the place name in GeoNames (uri = https://www.geonames.org/<code>) *Example*: [] gn:locatedIn <uri>	[] gn:locatedIn <http://sws.geonames.org/2673730/>

3. Process Knowledge. After extracting a set of triples (a subgraph) from a paper abstract, we: (i) **reshape** the subgraph, as we prompted GPT-4 to return annotations in a somewhat simpler structure to avoid structural errors, (iii) **add paper metadata** by transforming paper metadata retrieved from SemanticScholar (abstract, DOI, citations, publication date, authors) to RDF, (ii) **link** study variables in the subgraph to Linked Open Data from BioPortal and GeoNames, (iii) **validate** the subgraph using the SHACL shapes, and (iv) **integrate** the subgraph with the MIRA-KG, and **postprocess** the merged graph by removing redundant entities.

Specifically, in the **link** step, instances of `qb:ComponentProperty` are linked to concepts from BioPortal[18] using `qb:concept`. Prior to linking, a manual mapping study was performed (output can be found in the Zenodo repository) to see which BioPortal ontologies contained most of the useful study variables. Selected ontologies were HHEAR[19], SIO[20], IOBC[21] and MESH[22]). Subsequently, we retrieve location metadata from GeoNames[23] through FactForge[24], as the

[18] `https://bioportal.bioontology.org/`
[19] `https://bioportal.bioontology.org/ontologies/HHEAR/`
[20] `https://www.ebi.ac.uk/ols/ontologies/sio`
[21] `https://purl.bioontology.org/ontology/IOBC`
[22] `https://purl.bioontology.org/ontology/MESH`
[23] `https://www.geonames.org/`
[24] `http://factforge.net`

knowledge extraction step already retrieves GeoNames IRIs. Location coordinates, for instance, can aid researchers in discovering regional biases, or temporal trends. For the **validate** step, SHACL shapes are used to check for schema and ontological correctness. One of the SHACL shapes, for instance, checks for syntactic validity of dates, by ensuring `sh:datatype xsd:date`, and by checking for ontological correctness (begin before end-date, see SHACL shapes). For the **integrate** and **postprocess** steps, the subgraph is merged with the MIRA-KG. After processing all paper subgraphs, we remove duplicate BioPortal concepts and their labels via `owl:sameAs` smushing using RDFpro[25].

The time it took to process a single paper abstract was 1 minute and 11 seconds (MacBook M1 Chip, 8 cores, 8gb RAM). A fully annotated paper abstract as machine-readable data can be found in the Github repository.

3.4 The MIRA-KG

Table 3 shows statistics of the final knowledge graph, formalising 398 paper abstracts. We remove paper metadata functional properties from this view (for example `dcterms:created`), as these counts are equal to the instance count for `bibo:AcademicArticle`. We present the average degree with and without types, as these are naturally densely connected hubs. More interesting are the densely connected papers via citations or BioPortal concepts and locations with high in-degrees. Nodes with the highest in-degree that were not classes of the MIRA ontology were: economics/economy (`MESH:D004467`, `IOBC:200906008393315870`, 222), mortality rate (`MESH:D009026`, 188), hygiene (`MESH:D006920`, 91), the United States (`geonames:6252001`, 90) health (`MESH:D006262`, 90), age (`MESH:D006262`, 83) and income (`MESH:D007182`, 47).

Table 3: Statistics of the MIRA-KG: Class instance count, property count, number of nodes and edges, and average degree.

Classes	**Instances**	**Properties**	**Count**
`time:Instant`	948	`bibo:cites`	13496
`qb:DimensionProperty`	701	`qb:concept`	4041
`skos:Concept`	587	`rdfs:label`	3133
`mira:Explanation`	582	`dcterms:contributor`	1176
`SIO:SIO_000414` (geographic region)	524	`SIO:SIO_000008` (has attribute)	1065
`SIO:SIO_001050` (sample)	477	`time:inXSDDate`	946
`time:TemporalEntity`	474	`mira:hasSubject`	607
`qb:MeasureProperty`	477	`mira:hasContext`	607
`SIO:SIO_000976` (observational study)	469	`SIO:SIO_000205` (is represented by)	607
`SIO:SIO_001062` (human population)	408	`mira:hasObject`	602
`bibo:AcademicArticle`	398	`mira:hasEffectSize`	589
`gn:Feature`	114	`mira:hasRelation`	583
`mira:InteractionEffect`	105	`sem:hasPlace`	479
`SIO:SIO_000012` (organisation)	12	`time:hasTime`	477
		`gn:locatedIn`	454
		`wgs84_pos:long`	172
Number of Nodes	24281	`wgs84_pos:lat`	172
Number of Edges	32359	`gn:name`	114
Av. Degree	1.59	`mira:hasMediator`	77
Av. Degree (no types)	1.33	`SIO:SIO_000061` (located in)	2

[25] `https://github.com/dkmfbk/rdfpro`

4 Evaluation

In this section, we first evaluate the ontology (Section 4.1), and then the whole knowledge graph via a use case and data quality measures (Section 4.2).

4.1 Ontology Evaluation

Here, we show one of the competency questions as a SPARQL query (for brevity we omit prefixes, but these can be found in Figure 3): The **CQ** retrieves a claim with a specific dependent variable and a potential mediator.

```
PREFIX ...
select distinct ?ind_lab ?ind_var ?qual
                    ?med_var ?med_lab where {
?exp mira:hasSubject/qb:concept/rdfs:label ?ind_lab;
     #what is the subject of the explanation
     mira:hasSubject/qb:concept ?ind_var;
     #the object is COVID-19 mortality
     mira:hasObject/qb:concept mesh:D000086382,
                               MESH:D00902;
     #what is the effectSize of the association
     mira:hasEffectSize ?effectSize .
     #which variables mediate the association
OPTIONAL{?exp mira:hasMediator ?med_var;
          mira:hasMediator/rdfs:label ?med_lab .}
}
```

(a) The SPARQL query for the **CQ**

	ind_lab	med_lab
1	Residential Segregation	
2	Economics	
3	Ethnicity	
4	Occupation	Ability
5	Occupation	Home

	effectSize
1	mira:largeMediumNegativeES
2	mira:largeMediumPositiveES
3	mira:largeMediumPositiveES
4	mira:smallPositiveES
5	mira:smallPositiveES

(b) The result for the **CQ** (first five rows)

Fig. 6: **CQ1:** *Have associations been found between a socioeconomic variable and COVID-19, and which explanations can be found for the inequality?*

Table 6b shows the output of the query for the **CQ** (first five rows). It indicates that residential segregation, economics and ethnicity were all correlated with COVID-19 mortality. The association between COVID-19 and occupation was specifically mediated by the ability to work from home (*ability*, *home*). The rest of the competency questions can be found as SPARQL queries in the Github repository. They could all be answered using SPARQL queries over the MIRA ontology, although some limitations were discovered during ontology creation, see Section 4.2 below.

Limitations. The sessions with the expert showed that within papers there are various reformulations of the same hypothesis, which are at times not semantically the same. Such variations cannot be covered entirely without overly complicating the ontology, and cannot always be fully understood without access to the data and experiments. Moreover, effect sizes are often not clearly reported in the paper abstracts, hampering the quality of the extracted effectsizes.

4.2 Knowledge Graph Evaluation

In this section, we evaluate the use of the MIRA-KG via a use case, and its quality via a set of data quality criteria.

Use Case: Geographic Biases of Hypotheses and Findings. A researcher is studying the effect of income on mortality rates. They aim at discovering whether there is a geographic study bias for the research question, and whether the strength of the association differs among geographic regions. The researcher queries the MIRA-KG for all studies researching the influence of income (`MESH:D007182`) on mortality (`MESH:D00902`).

```
PREFIX ...
select ?long ?lat ?locName ?es ?geoId where {
#the subject of the explanation is income
?exp mira:hasSubject/qb:concept MESH:D007182 ;
    #the object of the exp. is mortality
    mira:hasObject/qb:concept MESH:D00902 ;
    #what is the effect size of the explanation
    mira:hasEffectSize ?es ;
    #what is the context of the explanation
    mira:hasContext ?pop .
#which sample is used to represent the population.
?pop sio:SIO_000205 ?sample .
#where is the sample collected
?sample sem:hasPlace ?location .
#what is the GeoNamesId of the location
?location gn:locatedIn ?geoId .
?geoId rdf:type gn:Feature ;
    #what is the name of the location
     gn:name ?locName ;
     #what is the longitude
     wgs84_pos:long ?long ;
     #what is the latitude
     wgs84_pos:lat ?lat .  }
```

(a) The SPARQL query for the **use case**

	long	lat	locName
1	17.64	59.85	Uppsala
2	8.05	47.4	Aarau
3	24.93	60.17	Finland
4	-52.0	-13.84	Brazil

	geoId	es
1	gn:2666199	`largeMediumNegativeES`
2	gn:2661881	`largeMediumPositiveES`
3	gn:660013	`largeMediumPositiveES`
4	gn:3469034	`largeMediumPositiveES`

(b) The query result for the **use case** (first four rows)

Fig. 7: **Use case:** *A social historian queries the MIRA-KG for metadata of all studies researching the influence of income (`MESH:D007182`) on mortality (`MESH:D00902`)*

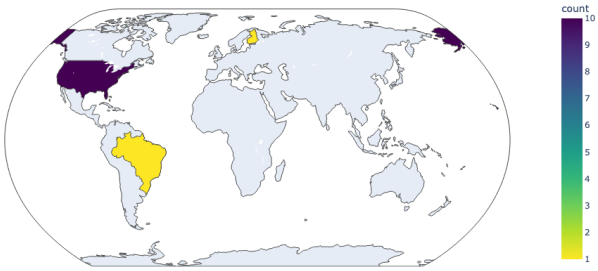

Fig. 8: Density plot for the **use case** query, showing a bias towards studies based on American population samples.

The output of the SPARQL query (first four rows shown in Table 7b) is visualised on a world map (see Figure 8), and shows a bias of research studies

towards American samples for the specific correlation. Moreover, by aggregating the results, the historian notices that four studies that are based on American population samples measure a negative correlation, and four a positive.

Data Quality Evaluation. We use the RDF data quality criteria as defined by [5] to, on the one hand, validate the MIRA-KG as well as the KGC pipeline, and, on the other hand, refine the graph. We summarise the most important ones for our case below.

Accuracy. Accuracy in terms of Linked Data deals with *syntactic* and *semantic* validity, as well as *duplicate entries*. *Syntactic validity:* the `DatatypeConstraintComponent` catches data type violations, such as the use of `rdf:XMLLiteral` where `xsd:date` should have been used. Furthermore, Table 3 shows all retrieved GeoNames identifiers were correct identifiers, as metadata was retrieved for all of them from FactForge. *Semantic validity:* The `LessThanConstraintComponent` measures logical temporal violations (begin date before end date). *Duplicate entries:* from Table 4 we see that with eight violations, these were annotated with high logical accuracy (98.3%). Lastly, duplicate BioPortal entities were handled during the postprocessing step of the KGC pipeline, as described in Section 3.3.

Table 4: **Left**: SHACL constraint violations for the MIRA-KG. **Right**: LLM retrieval errors for 20 research articles.

Constraint violation	Count	LLM retrieval error (20 articles)	Count
`sh:LessThanConstraintComponent`	8	`Study variables`	3
`sh:DatatypeConstraintComponent`	23	`Location`	2
`sh:ClassConstraintComponent`	38	`Context`	1
`sh:MinCountConstraintComponent`	122	`Time period`	7
`sh:MaxCountConstraintComponent`	4	`Effect size`	1
`sh:XoneConstraintComponent`	27		

Trustworthiness. This dimension is defined as the degree to which the information is accepted to be correct, true, real and credible. As the MIRA-KG was created automatically using GPT-4, we analyse the first twenty hypotheses (annotations can be found on Zenodo). Errors with respect to the retrieval are shown in Table 4. *Study variables:* in case study variables were incorrect, GPT-4 used related terms (such as *time* instead of *age* as independent variable). *Time period:* most common errors were incorrectly retrieved dates. However it should be mentioned that, in most cases, dates were incorrectly extracted when no dates were mentioned in the abstract.

Even though errors appear, most terms appear to be retrieved correctly, even for quite complicated hypotheses such as: *In the context of a human population, "Social inequality", mediated through "Health-services structure", is correlated with "COVID-19 mortality" with a largeMediumPositiveES."Governmental response to COVID-19" moderates this correlation. Evidence comes from "São Paulo residents", from "São Paulo" between the years "2020-03-01" and "2020-09-30".*

Consistency. Semantic consistency is the extent to which a collection uses the same values for conveying the same meanings throughout. SHACL constraints were applied to enforce schema-correctness (such as a hypothesis has at least one independent variable). Constraint violations can be found in Table 4 (`sh:Class`, `sh:MinCount`, `sh:MaxCount` and `sh:Xone-ConstraintComponent`). By looking at the exact violations, we see that common errors were errors in structure, such as a context which was linked to a study instead of an explanation, resulting in violations of both the `Class` as well as the `MinCount` constraint components. The graph was queried to retrieve and dissolve these common errors to further consummate the entire graph.

Limitations Due to the limitations mentioned in Section 4.1 (underspecification, ambiguity, and unclarity), it is challenging to extract hypotheses from scientific papers in a precise manner. The ontology will, however, allow researchers to publish their findings in a machine-readable manner alongside paper publication. Second, representing complex concepts, such as 'the ability to work from home', is non-trivial, as these are not always defined in ontologies or vocabularies and can be described in many different ways. For querying, linking such a variable to the concepts 'ability' as well as 'home' would suffice, but if precise reasoning is required, other solutions need to be considered (e.g. OWL axioms).

5 Conclusions

Studies in (historical) social sciences, such as social demography, tend to be restricted to specific time periods and regions, making comparison of hypothesis tests across the vast array of research papers difficult. To assist social historians and demographers in the scientific process, this study describes MIRA, a method knowledge graph construction, and a resource consisting of a ontology and knowledge graph to capture hypotheses and findings in social demography.

This study shows that the MIRA ontology allows researchers to formulate their key questions, and research results in a structured and semantically sound way. Moreover, using over 400 abstracts from the field of social demography, our knowledge graph construction pipeline demonstrates that the knowledge-intensive task of semantic annotation can be (semi-)automated when employing a Large Language Model, even without any training, validating data quality after automated annotation.

Future work will focus on dealing with variations in natural language, and on expanding the MIRA-KG by application of the MIRA ontology on a larger set of studies in order for the field to reorganize itself, for instance by motivating researchers to publish their findings in a machine-readable manner.

6 Acknowledgements

This work was funded by the European MUHAI project (Horizon 2020 research and innovation program) under grant agreement number 951846. We thank Tobias Kuhn, Frank van Harmelen, and Inès Blin for the insightful discussions that contributed to this work.

References

1. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.: Towards a knowledge graph for science. In: Proceedings of the 8th international conference on web intelligence, mining and semantics. pp. 1–6 (2018)
2. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
3. de Boer, V., Bonestroo, I., Koolen, M., Hoekstra, R.: A linked data model for data scopes. In: Metadata and Semantic Research: 14th International Conference, MTSR 2020, Madrid, Spain, December 2–4, 2020, Revised Selected Papers 14. pp. 345–351. Springer (2021)
4. Bucur, C.I., Kuhn, T., Ceolin, D., Van Ossenbruggen, J.: Expressing High-Level Scientific Claims with Formal Semantics. K-CAP 2021 - Proceedings of the 11th Knowledge Capture Conference pp. 233–240 (2021). https://doi.org/10.1145/3460210.3493561
5. Candela, G., Escobar, P., Carrasco, R.C., Marco-Such, M.: Evaluating the quality of linked open data in digital libraries. Journal of Information Science **48**(1), 21–43 (2022)
6. Dessí, D., Osborne, F., Recupero, D.R., 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
7. 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: The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings. pp. 678–696. Springer (2022)
8. Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E., Sack, H.: Ai-kg: an automatically generated knowledge graph of artificial intelligence. In: The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19. pp. 127–143. Springer (2020)
9. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., et al.: The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. Journal of biomedical semantics **5**, 1–11 (2014)
10. Fathalla, S., Lange, C., Auer, S.: An upper ontology for modern science branches and related entities. In: European Semantic Web Conference. pp. 436–453. Springer (2023)
11. Gangemi, A., Presutti, V.: A multi-dimensional comparison of ontology design patterns for representing n-ary relations. In: SOFSEM 2013: Theory and Practice of Computer Science: 39th International Conference on Current Trends in Theory and Practice of Computer Science, adfordCzech Republic, January 26-31, 2013. Proceedings 39. pp. 86–105. Springer (2013)
12. Garijo, D., Gil, Y., Ratnakar, V.: The disk hypothesis ontology: Capturing hypothesis evolution for automated discovery. In: K-CAP Workshops. pp. 40–46 (2017)
13. Gil, Y., Garijo, D., Ratnakar, V., Mayani, R., Adusumilli, R., Boyce, H., Srivastava, A., Mallick, P.: Towards continuous scientific data analysis and hypothesis evolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
14. Graham, E.: Theory and explanation in demography: The case of low fertility in europe. Population Studies **75**(sup1), 133–155 (2021)

15. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information services & use **30**(1-2), 51–56 (2010)
16. de Haan, R., Tiddi, I., Beek, W.: Discovering research hypotheses in social science using knowledge graph embeddings. In: The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18. pp. 477–494. Springer (2021)
17. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S.E.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife **6**, e26726 (2017)
18. Hoekstra, R., Koolen, M.: Data scopes for digital history research. Historical Methods: A Journal of Quantitative and Interdisciplinary History **52**(2), 79–94 (2019)
19. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
20. Kuhn, T., Chichester, C., Krauthammer, M., Queralt-Rosinach, N., Verborgh, R., Giannakopoulos, G., Ngomo, A.C.N., Viglianti, R., Dumontier, M.: Decentralized provenance-aware publishing with nanopublications. PeerJ Computer Science **2**, e78 (2016)
21. Larmande, P., Todorov, K.: AgroLD: A Knowledge Graph for the Plant Sciences. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12922 LNCS**, 496–510 (2021). https://doi.org/10.1007/978-3-030-88361-4_29
22. Magnusson, I.H., Friedman, S.E.: Extracting Fine-Grained Knowledge Graphs of Scientific Claims: Dataset and Transformer-Based Results. EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings pp. 4651–4658 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.381
23. Mavergames, C., Oliver, S., Becker, L.: Systematic reviews as an interface to the web of (trial) data: using pico as an ontology for knowledge synthesis in evidence-based healthcare research. SePublica **994**, 22–6 (2013)
24. Meroño-Peñuela, A., Ashkpour, A., Guéret, C., Schlobach, S.: Cedar: the dutch historical censuses as linked open data. Semantic Web **8**(2), 297–310 (2017)
25. Missiakoulis, S.: Cecrops, king of athens: the first (?) recorded population census in history. International Statistical Review **78**(3), 413–418 (2010)
26. Molitoris, J., Dribe, M.: Industrialization and inequality revisited: Mortality differentials and vulnerability to economic stress in stockholm, 1878–1926. European Review of Economic History **20**(2), 176–197 (2016)
27. Presutti, V., Daga, E., Gangemi, A., Blomqvist, E.: extreme design with content ontology design patterns. In: Proc. Workshop on Ontology Patterns. pp. 83–97 (2009)
28. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
29. Spadaro, G., Tiddi, I., Columbus, S., Jin, S., Ten Teije, A., Team, C., Balliet, D.: The cooperation databank: machine-readable science accelerates research synthesis. Perspectives on Psychological Science **17**(5), 1472–1489 (2022)
30. Stocker, M., Heger, T., Schweidtmann, A., Ćwiek-Kupczyńska, H., Penev, L., Dojchinovski, M., Willighagen, E., Vidal, M.E., Turki, H., Balliet, D., et al.: Skg4eosc-scholarly knowledge graphs for eosc: Establishing a backbone of knowledge graphs for fair scholarly information in eosc. Research Ideas and Outcomes **8**, e83789 (2022)

31. Stocker, M., Oelen, A., Jaradeh, M.Y., Haris, M., Oghli, O.A., Heidari, G., Hussein, H., Lorenz, A.L., Kabenamualu, S., Farfar, K.E., et al.: Fair scientific information with the open research knowledge graph. FAIR Connect **1**(1), 19–21 (2023)
32. Tiddi, I., Balliet, D., ten Teije, A.: Fostering scientific meta-analyses with knowledge graphs: a case-study. In: The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17. pp. 287–303. Springer (2020)
33. Viet, S.M., Falman, J.C., Merrill, L.S., Faustman, E.M., Savitz, D.A., Mervish, N., Barr, D.B., Peterson, L.A., Wright, R., Balshaw, D., et al.: Human health exposure analysis resource (hhear): A model for incorporating the exposome into health studies. International journal of hygiene and environmental health **235**, 113768 (2021)