Mining Semantic Relations in Data References to Understand the Roles of Research Data in Academic Literature

Lizhou Fan
lizhouf@umich.edu
School of Information, University of Michigan
Ann Arbor, Michigan, USA

Sara Lafia
slafia@umich.edu
Inter-university Consortium for Political and Social Research, University of Michigan
Ann Arbor, Michigan, USA

Morgan Wofford
mwofford@umich.edu
School of Information, University of Michigan
Ann Arbor, Michigan, USA

Andrea Thomer
athomer@arizona.edu
School of Information, University of Arizona
Tucson, Arizona, USA

Elizabeth Yakel
yakel@umich.edu
School of Information, University of Michigan
Ann Arbor, Michigan, USA

Libby Hemphill[*]
libbyh@umich.edu
Inter-university Consortium for Political and Social Research, University of Michigan
Ann Arbor, Michigan, USA

ABSTRACT

Research data serves important roles in scientific discovery and academic innovation. To appropriately assign credit for data work and to measure the value of research data, it is essential to articulate how data are actually used in research. We leveraged a combination of computational methods and human analysis to characterize different types of data use by mining semantic relations from the phrases where data are referenced in academic literature. In particular, we investigated references to data in the bibliography of a large social science data archive, the Inter-university Consortium for Political and Social Research (ICPSR). After retrieving and extracting semantic relations as subject-relation-object triples, we used rule-based methods to classify them. We then annotated samples from 11 frequent classes of data reference triples and found that they vary primarily along two dimensions of data use: proximity and function. Proximity describes the distance between the author and the data they reference (e.g., direct or indirect engagement). Function describes the role that data plays in each reference (e.g., describing interaction or providing context). These semantic relationships between authors and data reveal the ways data are used in scientific publications. Evidence of the variety of ways data are used can help stakeholders in research data curation and stewardship – including data providers, data curators, and data users – recognize the myriad ways that their investments in data sharing are realized.

KEYWORDS

information extraction, knowledge discovery, research data management, semantic triples, text mining

1 INTRODUCTION

Data citations, like citations of scientific literature, are an important form of credit that acknowledge work done with data. "Data work" includes data collection, analysis, and curation – the transformation and processing of data for reuse by others [58, 72]. It also involves the maintenance of research infrastructures to support long-term data sharing and preservation [47]. Assigning credit for data work in the form of citations is important for creating sustainable knowledge infrastructures [21]. Formal credit for data also allows readers to discover and engage with the work of others by making the lineage and impact of a given work clear [9].

Studying the discourse surrounding data citations is critical for understanding scholarly communication [13]. Despite calls to acknowledge research data as scholarly objects and efforts to develop data citation guidelines [32, 74], data citation practices are not consistent [17]. Authors often mention data (e.g., by using the name of a dataset) rather than formally cite them (e.g., using a unique identifier and style-guide appropriate reference) [49, 50]. Thus, capturing only formal, machine-actionable data citations paints an incomplete picture of overall data use[1]. Incomplete records of data use result in citation networks that can mislead data providers, curators, and users when planning for and evaluating their data work.

Informal data references in academic articles offer a window into data that are often discussed in research and how those data are used in practice; in the aggregate, data references also offer insights into the discourse surrounding data use [67]. In this sense, data references can inform stakeholders about research data use. Fortunately, understanding the contribution of many kinds of scholarly products, including research data, is increasingly possible given recent bibliometric advances in large-scale citation retrieval and mining [13]. Computational reference mining pipelines increase recall for both formal and informal data references. For example, a recent expansion of literature citing data from a large social science data archive – the Inter-university Consortium for Political and Social Research (ICPSR)[2] – was used to analyze patterns of data use in citation networks [42, 43]. Data references collected across academic literature sources offer a more representative depiction of the impact of scholarly data work.

[*] Also with School of Information, University of Michigan.

[1] We use the term "data references" to encompass both formal data citations and informal data mentions. Formal citations include the use of persistent identifiers as well as additional metadata elements such as creators' names, titles, dates, dataset versions, and data providers. On the other hand, data mentions typically lack many of these metadata elements and are, therefore, likely to be missed in citation analyses.

[2] The ICPSR Bibliography of Data-Related Literature is available at https://www.icpsr.umich.edu/web/pages/ICPSR/citations/.

In this project, we analyzed scholarly interactions with research data by focusing on sentence-level data references. Rather than treating data references as a binary relation between a publication and data (i.e., referenced or not), we used natural language processing (NLP) to capture a wider array of authors' statements mentioning data in their publications. Data references reveal many ways researchers rely on data to produce new knowledge and scholarly work; their uses are not limited to data analysis. To gather detailed evidence of researchers' interactions with data, we mined semantic relations in 7,486 sentences referencing research data, which we extracted from 1,128 data-related publications. We found two main structures distinguishing subject-relation-object triples in data references: the *proximity* between the author and the data – derived from the structure of subjects and objects in the sentences – and the *functions* of the data references – based on the verbs describing the actions that authors and data perform. Our analysis reveals users' interactions with data throughout the research process and provides a foundation for studying the impact of archived data on corresponding research and data management communities.

2 RELATED WORK

2.1 Tracing the Impact of Research Data in Scholarly Communication

Many prior studies of data reuse have focused on data reusers' attitudes as shown through interview studies, surveys, and analyses of data requests [19, 20, 27, 57]. These studies offer insights into researchers' considerations and motivations for seeking data. However, behavioral studies rely on users' accounts of their own data use. From these studies alone, we cannot know how users interact with data throughout the research process, let alone differentiate the types of work data supports.

Citations to other papers reveal the purpose the referenced paper plays in the work that cites it (e.g., background, compare/contrast, motivation) [60]. Similarly, data references reveal how researchers interact with secondary data (i.e., data produced by someone other than the creator) [7]. Researchers rely on both research literature and datasets to support their analysis and writing [39]. Like references to scientific literature, data references "establish evidentiary sources, give credit, and facilitate the discovery and retrieval of materials on which the citing publication is based" [75]. For example, authors might mention features of a well-regarded survey – such as its sampling frame or questions – without analyzing its data. This kind of attribution does not indicate data use and, therefore, can be challenging to detect. The survey and its producers however, deserve recognition for supporting scientific inquiry.

Organizations, such as data repositories, emphasize the importance of properly citing data to give them credit as scholarly research objects. Many data providers assign unique persistent identifiers, such as DOIs, to datasets [8, 56]. Organizations, such as FORCE11, have also convened task forces to propose formal data citation principles [2]. These principles encourage authors to provide full data descriptions, with unique persistent identifiers, in their papers. Recent studies of data archives, such as ICPSR, have found that DOIs are frequently missing from data references [48–50]. Inconsistent data citation practices lead to many missed data references when quantifying impact and assigning credit.

Effective dataset retrieval relies upon the quality and quantity of metadata, which provides prospective data reusers with valuable context [38]. Data references offer insights into how researchers make sense of data produced by others (e.g., through re-analysis), and strategies for contextualizing or justifying data choices when communicating to a scientific audience [19, 77]. These insights and strategies provide information that may help science funders, data producers, and data curators, understand how their data are used. For instance, direct or indirect data mentions indicate the proximity between the author and data [36] as well as the type of data reuse (e.g., integrative, comparative) or non-reuse (e.g., discussing an implication, explaining the source of a linked variable) [26, 57].

Prior investigations of data references have restricted their analysis to specific elements of literature (e.g., abstracts) or types of works (e.g., data papers) [36, 46]. Manual approaches to study data references are time and resource-intensive, and do not easily scale [47]. By contrast, automated processes enable tracking and analysis of data references across scientific literature. Thus, mining data references can help capture data's broader impacts and improve the prospect of assigning credit to data producers and providers.

2.2 Analyzing the Semantic Relations in Data References as Scholarly Discourse

Citation analysis provides insights into knowledge production and related scholarly practices [12, 45, 68]. Authors' motivations for citing can be inferred and classified by analyzing source text [31, 60]. Citation contexts, or the portions of text citing target references, can be interpreted at a syntactic and semantic level [16]. Citation sentiment can also be inferred from analysis; for example, citations in a publication can criticize, compare, or substantiate [1]. Relationships between the form and function of citations also indicate citation intent [44]. The structural properties, such as sections of papers where citations occur, support classification tasks and can be used to predict the function of a citation (e.g., providing background information or describing a method) [10].

Computational discourse analysis is often applied to study the function of individual narratives (e.g., in social media) or in formal, written text (e.g., academic articles). These methods extract semantic information from written language using machine learning and natural language processing [37, 70]. Discourse analysis is useful for discovering large-scale citation trends, summarizing citation contexts, and inferring the content of documents [14]. Automatic approaches for identifying informal references to datasets in publications have used strategies, such as pattern induction and named entity recognition, to achieve moderate levels of recall [6].

General information extraction frameworks, such as OpenIE [3], provide a way to structure and assess the dominant linguistic features of corpora at scale. Extracted semantic triples provide an analytical base for inferring authors' intent and constructing new knowledge. Prior research has leveraged theoretical frameworks to classify agency as a principal relationship in textual content. Agency between a subject and an object entity relates the action taker to the receiver. For instance, Labov and Waletsky [41] classified agency in the narratives of conversational discourse by assessing temporal organization (i.e., the order in which the subject narrates events and actions in the story), evaluative description (i.e., subjective assessments of objects and events), and contextual orientation (i.e., supplementary information intended to help the audience process information in the narrative).

These criteria for classifying agency have been implemented in computational tools to label roles in discourse. Swanson et al. [69] developed a method for detecting narrative clause types and proposed a labeling algorithm for analyzing personal stories. Saldias and Roy [65] created a Natural Language Processing (NLP) model for understanding similar aspects of personal narratives. Fan and Presner [18] conducted an algorithmic close reading of Holocaust testimonies using a computational system to detect agency, evaluation, and orientation from interview transcripts. These implementations of semantic mining, based on Labov and Waltsky's model [41], provide a foundation for our analysis of data references, which we essentially interpret as researchers' narratives detailing how they interacted with data.

Recent research has also prototyped infrastructures and workflows to support computational discourse analysis on research materials. For example, Hanson et al. [29] designed a mapping system that used scholarly knowledge graphs to capture and preserve maps of relationships in distributed scholarly artifacts as Distributed Scholarly Complex Objects (DiSCOs), demonstrating the interactions among types of research components. Oelen et al. [52] developed and evaluated an infrastructure that used NLP-extracted scholarly knowledge statements for constructing a paper-centric knowledge graph. These infrastructures and workflows enable the visualization and analysis of complex relations among research objects, including datasets and publications.

3 DATA AND METHODS

Our approach depends on the full text of research publications, which we accessed through Dimensions, a bibliometric database with a full-text search index of over 69 million journal articles and other scholarly works [34]. We searched Dimensions for the names, study numbers, and digital object identifiers (DOIs) for all 10,491 publicly-available social science studies archived at ICPSR at the time of analysis (July 2022). We first retrieved 1,074 available full-text publications that included a reference to one or more ICPSR studies. We then tokenized each publication into sentences and applied a Named Entity Recognition (NER) model that we trained to identify informal dataset references as part of a previous project [42]. The model was trained on social science literature included in the current ICPSR Bibliography. We applied the NER model and extracted dataset entities from 7,486 sentences. Finally, we used OpenIE [3] to extract semantic triples (i.e., subject, relation, object) from each sentence. We manually categorized subjects, relations, and objects in each sentence using the notions of agent and agency [41]. Within these semantic triples, we identified 11 frequent combinations (i.e., classes) and labeled the roles of *proximity* and *function* in the sample triples to further characterize author-data interaction. We summarize our analysis steps in Figure 1.

3.1 Retrieve and Extract: From Data References to Semantic Triples

We used OpenIE [3] to extract general semantic information instead of limiting the extraction to a range of pre-defined entities and relations. Our prototype used this common setting because in data archives and bibliographies, there is little prior knowledge about data reuse. In particular, we extracted triples and retrieved relational semantic information from the data reference sentences. We input the full text of data references into OpenIE, which extracted triples by chunking and shortening sentences into independent clauses based on their semantic parsing tree structure. OpenIE then generated groups of Subject, Relation, and Object phrases, which are maximally compact semantic units that retain the key meaning of the original independent clauses. This process is shown in an example data reference sentence [25] in Figure 2.

OpenIE first makes a prediction of universal dependencies based on its pre-trained NLP model and assigns dependency parsing tags to each token in the sentence; the system then extracts all possible clauses, which append all possible subject, verb, and object parts, and disregards positional distances between tokens; triples are then extracted from each clause. More than one triple may be extracted from a clause due to the complexity of different combinations of language units in noun phrases.

To simplify our analysis of data references, we applied the following rules to select the single most informative triple among all triples extracted by OpenIE: we retained all triples with the same subject and relation and collected them into a subgroup; we then compared the objects of all triples in a subgroup and kept the one object with the greatest number of tokens; if two objects had the same number of tokens, we compared the total length of object strings and kept the longer one. We obtained 10,339 data reference triples through this retrieval, extraction, and selection process.

3.2 Classify: A Rule-based System for Agent-agency Characterization

We classified the extracted triples using a two-step rule-based system that characterized the triple with respect to its (a) agents in subjects and objects and (b) agency in relations[3]. We leveraged the notions of agent and agency, described in Section 2.2, to categorize subjects, relations, and objects. In a data reference, an agent is an entity that takes an active role or produces a specified effect. Here, agency is a relationship between an author and a dataset.

We detected the agents included in the data reference for subjects and objects. We developed rules to detect authors and datasets in the data reference triples. The rules, shown in Table 1, were inspired by existing rule-based methods [30] and indicative keywords for detecting dataset references in scientific literature [55]. They use regular expression patterns to determine if a triple element contains information about a dataset or an author. Subjects and objects that were neither datasets nor authors were assigned the label "other".

By combining the three possible labels for subjects and objects, the agent categorization step resulted in nine possible combinations of entity-role categories (for example, "author" as the subject and "data" as the object). We used a Tableau dashboard[4] to evaluate the

[3] The details of the algorithm in this system are provided in Appendix C Algorithm 1.
[4] The dashboard is available on Tableau Public https://tinyurl.com/DataRefDash

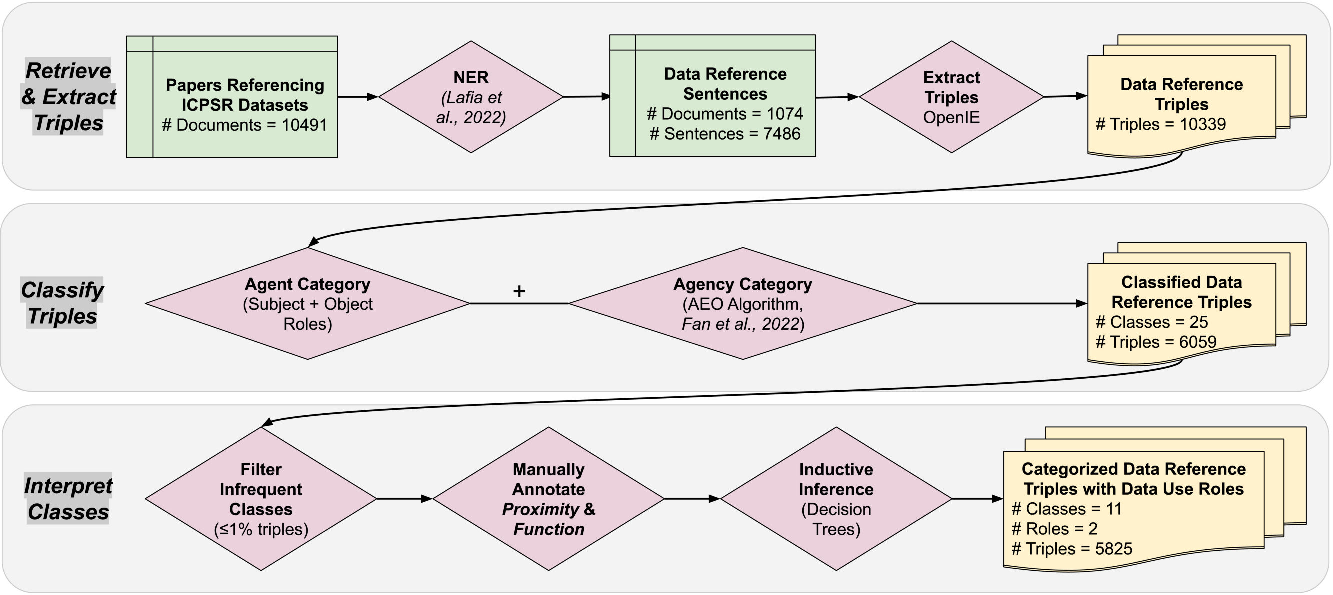

Figure 1: Our workflow for mining semantic relations in data references. The color and shape of each object represent the class of the sub-step – green squares (first row) are data references (text) extracted from papers that reference ICPSR data, yellow blocks (last column) are end-of-step data results (text), and red diamonds are analytical actions for interpretation.

Table 1: Rules and regular expression patterns for identifying dataset and author agents in data references

<table>
<tr><th>Agent</th><th>Rules</th><th>Regex Pattern</th></tr>
<tr><td>Dataset</td><td>Detect any token in names of a dataset</td><td>None</td></tr>
<tr><td></td><td>Keywords and their variations matched using regular expressions</td><td>‘(?:train|test|validation|testing|trainings?)\s*(?:set)’, ‘data’, ‘data\s*(?:set|base)s?’, ‘corp(us|ora)’, ‘samples?’, ‘tree\s*bank’, ‘collections?’, ‘benchmarks?’, ‘surveys?’, ‘stud(y|ies)’, ‘reports?’, ‘census(es)?’</td></tr>
<tr><td>Author</td><td>Detect any first-person pronoun of authors</td><td>‘I’, ‘(W|w)e’</td></tr>
</table>

nine agent category combinations. We decided to exclude four infrequent and irrelevant data reference categories. We also removed categories with fewer than 10 occurrences or incorrectly extracted triple relations. OpenIE’s internal limitation of searching long sentences resulted in some errors in triple extraction. Ultimately we kept 6,059 triples across five subject-object pairs: Author-Dataset, Author-Other, Dataset-Dataset, Dataset-Other, and Other-Dataset.

We categorized relations between these subject-object pairs by mining the agency of triples. Agency in these pairs corresponds to the dataset reference role. Our method classified the agency of data reference into five categories – active, passive, possible, evaluation, and orientation – derived from Labov and Waletzky’s [41] social-linguistic model of narratives and a recent computational adaption of this model [18]. Table 2 shows examples of agency classification in data reference sentences[5]. The *active*, *passive*, and *possible* categories, respectively, show the relationship between subjects and objects and what goals the authors met when referring to data. The *evaluation* category includes descriptions where the agent of subjects assesses objects, which usually use “to be” verbs. The *orientation* category focuses on contextual information that may help readers situate the narrative, such as the research areas that use a data reference.

3.3 Interpret: Explaining Proximity and Function of Data References

Fewer than 1% of the data reference triples in the papers retrieved from the ICPSR Bibliography accounted for 14 out of 25 data reference classes. We filtered out those infrequent classes, resulting in 5,825 triples (96.14%). The agent and agency classes of the 11 frequent classes are specified in Table 3.

We manually labeled a sample of semantic triples to identify the *proximity* of the author’s relation to data and the *function* that the data served in their paper. We iteratively refined the label definitions by independently labeling a subset of 220 triples. We reached

[5] The example triples are extracted from sentences in Appendix A.1

Table 2: Categories of data reference agency and examples of data references for each

Agency	Subject	Relation	Object
Active	We	obtain data from	Mannheim Eurobarometer Trend File
Passive	ADR data	are collected from	major depository bank websites
Possible	SETUPS data	can	be ordered for use with SPSS as card image
Evaluation	BRFSS	is	the largest, continuously conducted, tele-phone health survey
Orientation	our source	is	World Event Interaction Survey

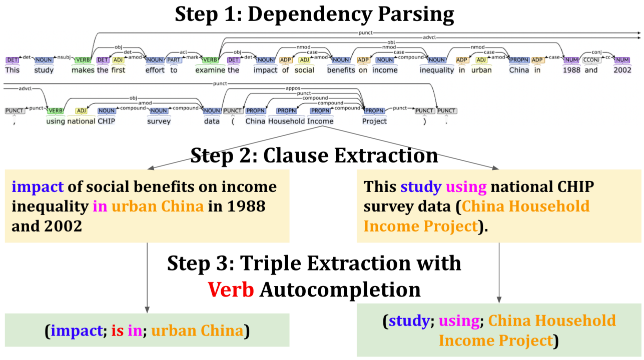
Step 1: Dependency Parsing
This study makes the first effort to examine the impact of social benefits on income inequality in urban China in 1988 and 2002
, using national CHIP survey data (China Household Income Project) .
Step 2: Clause Extraction
impact of social benefits on income inequality in urban China in 1988 and 2002
This study using national CHIP survey data (China Household Income Project).
Step 3: Triple Extraction with Verb Autocompletion
(impact; is in; urban China)
(study; using; China Household Income Project)

Figure 2: An example of the triple extraction process for a data reference using OpenIE. The sentence parsed in this example is "This study makes the first effort to examine the impact of social benefits on income inequality in urban China in 1988 and 2002, using national CHIP survey data (China Household Income Project)." We used Stanza [61] to visualize universal dependencies in the sentence.

Table 3: Combinations (i.e., "categories") of Subject agent, Agency of relation, and Object agent. Categories are listed here in alphanumeric order.

Category	Subject	Agency	Object	Percent
1a	author	Active	dataset	8.52
2a	author	Active	other	9.21
3a	dataset	Active	other	7.58
3b	dataset	Passive	other	3.63
3d	dataset	Orientation	other	5.94
4a	dataset	Active	other	14.49
4b	dataset	Passive	other	3.66
4d	dataset	Orientation	other	8.91
5a	other	Active	dataset	16.72
5b	other	Passive	dataset	5.69
5d	other	Orientation	dataset	11.77

agreement in the first round of annotation (Krippendorff's $\alpha = 0.92$ for proximity and $\alpha = 0.71$ for function) [40].

We considered *proximity* as either **direct** or **indirect**. Direct references indicate that the authors were in direct contact with the data. For instance, we labeled references as direct when authors said they had "used" or "viewed" the data. Indirect references indicate that something, usually other authors or analyses, stood between the authors and the data. For example, we labeled references as indirect when the author said that they "relied on" or "wrote about" what others had done with the data.

Similarly, we considered the *function* of a data reference as either providing **context** or describing an **interaction**. In context functions, authors provide background information about a dataset, such as historical context, to the reader. Interaction functions indicate that the object of the triple interacted with the data. Descriptions of data analyses or calculations are examples of interactions.

Using the triples that we annotated[6], we further interpreted each of the 11 data reference classes by assigning the triples *proximity* and *function* roles. We inferred the role labels based on classes of the sample triples using a decision tree, which is suitable for building knowledge-based systems [62]. We implemented the decision tree using the "rpart" R package and used the Gini coefficient as a metric to determine the best rule for splitting the data [71]. As Figure 3 demonstrates, if a triple is in category "1a" or "2a", it has a *proximity* role of **direct**. In contrast, the rest of the categories correspond to the role of **indirect**; similarly, if a triple is in category "3a", "3b", "3d", "4a", "4b", "4d", "5a", "5b", and "5d" it has a *function* role of **context**, while the rest of the categories correspond to the role of **interaction**. In other words, we used the *proximity* and *function* role labels to represent all of the triples in each class, including those that were not in the sample. This allowed us to infer and summarize author-data interactions for all data reference triples.

4 RESULTS

The structural features we identified in the data references indicated two main dimensions relating authors to data: *proximity* and *function. Proximity* describes the distance between authors and the data they reference (i.e., whether they interact directly or indirectly with the data). *Functions* distinguish data contexts (e.g., what previous users have done with data, how data are produced, or how data can be accessed) and data interactions (e.g., the operations[7] done with data, to data, or by data).

We observed a clear association between the *proximity* and *function* roles: **direct** data references usually indicate primary data-author **interaction**, such as manipulating data in an analysis (triple categories "1a" and "2a"); **indirect** data references usually provide **context** that explains and motivates data-related research (triple categories "3a", "3b", "3d", "4a", "4b", "4d", "5a", "5b", and "5d"). This

[6] As Appendix B Figure 6 shows, we used 202 triples that the annotators agreed on.

[7] We use the word "interaction" here to describe the interactive operation class of function, as opposed to the "context" class. We also use "author-data interaction" as an umbrella term for roles of data references.

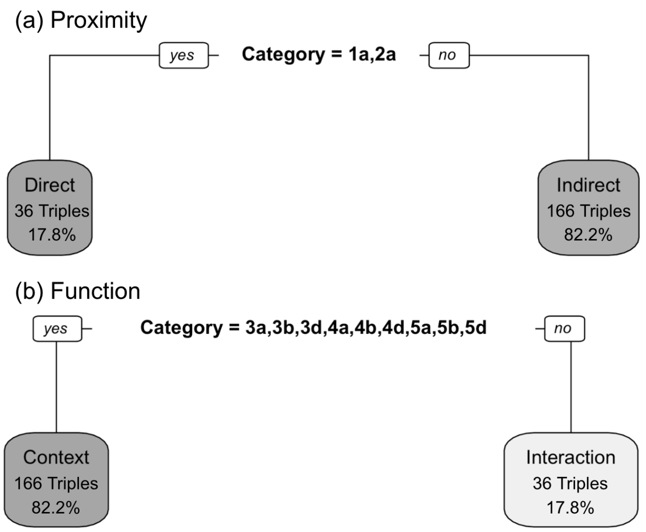

Figure 3: Decision trees for inferring the function from the agency:agent categories. The agent:agency categories produce non-overlapping distinctions between the proximity and function indicated in each triple.

association is demonstrated in the confusion matrix of the 202 triples in our sample (Table 4).[8] To illustrate the mappings between structural features, semantic relations, and types of author-data interactions, Figures 4 and 5 provide examples from the ICPSR Bibliography.

Table 4: Confusion matrix of the 202 triples in our manually-annotated sample. The row and column majority cells are in bold.

	context	interaction
direct	11	**31**
indirect	**143**	17

4.1 Authors' *Proximity* to Data: Direct and Indirect Data References Indicate Research Methods

Authors' proximity to the data they reference distinguishes the research method of their publication. The semantic relations network of the author, the dataset(s), and other related entities provide details about the direct and indirect data-author interactions and imply the research methods used in the scholarly work. For example, the triple “I”-“investigate”-“the effect of material status on well-being” indicates data analysis, while “access”-“is provided”-“by the ICPSR data archive” indicates a study related to a data archive.

We identified two different proximity relations: direct and indirect. Sentences where “author” was the subject, and either a “dataset” or something other than “author” was the object, signaled direct interactions with data. Sentences with “dataset” as the subject

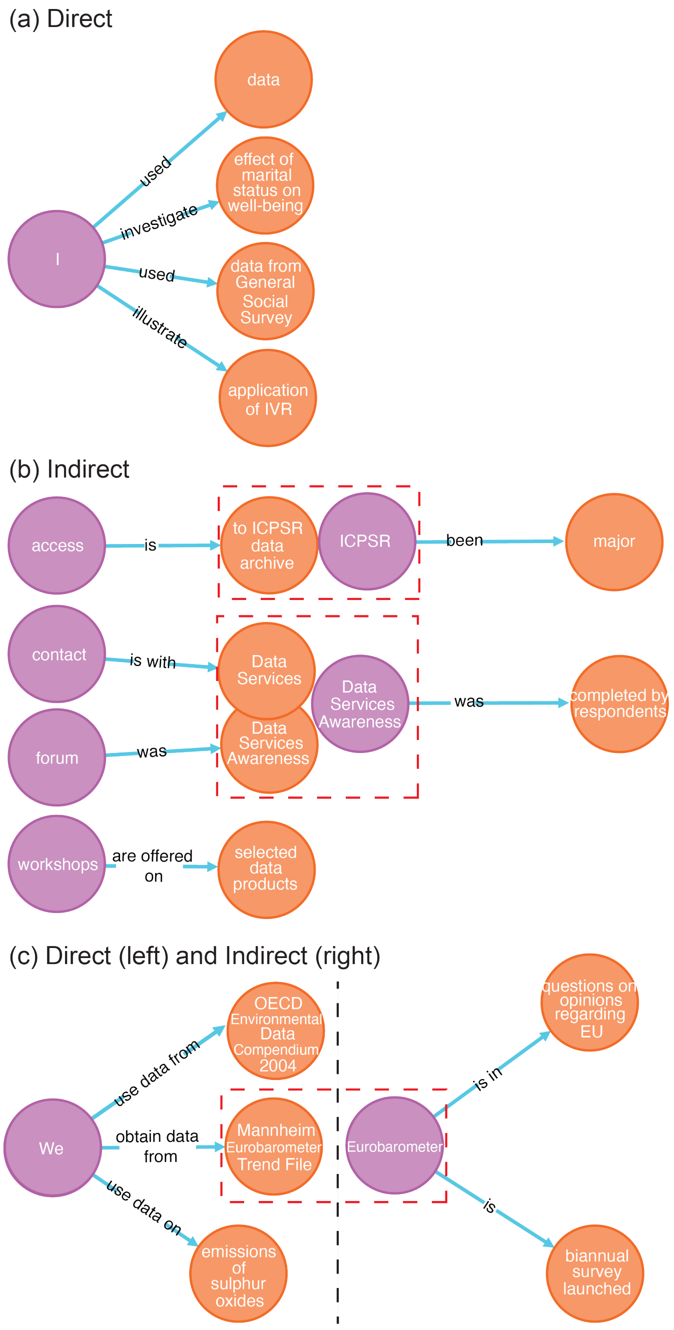

Figure 4: Examples of direct, indirect, and both direct and indirect semantic relations. Each subgraph includes triples from a single publication ([15], [63], and [28], respectively). Purple nodes are subject agents, orange nodes are object agents, and blue edges describe the agency (i.e. relations between subjects and objects). Red-dashed lines surround nodes that appear as both subjects and objects in the paper.

[8] The results in the confusion matrix (Table 4) are different from the previous decision tree (Figure 3). The confusion matrix is based on the annotators' agreements, while the decision tree is a prediction result based on agent and agency features.

and sentences that did not have an “author” entity signaled indirect interactions with data. The difference between the two proximity classes is reflected in passive references that orient datasets to “other” entities, which are neither “authors” nor “datasets” (indirect), rather than active interactions between “author” and “dataset” entities (direct). By analyzing all triples in a publication, through the proportion of the two proximity categories, namely direct and indirect, we are able to determine if the author’s research methods focus closely on data.

To demonstrate the direct and indirect classes of author-data proximity, and the implications of using different research methods, we provide three semantic relations networks where publications indicate (a) direct, (b) indirect, or (c) both direct and indirect author relationships to data (Figure 4). In publication (a) [15], there are only direct data references, where “I”, the author, is the subject and the relations are active operations on data. From the title of publication (a) (*Combating unmeasured confounding in cross-sectional studies: Evaluating instrumental-variable and Heckman selection models*), we can see that this is research uses data analysis as the research method. In publication (b) [63], there are only indirect data references, where elements of qualitative analysis, including the data service provider (ICPSR) and the research topic (data service awareness), are the important agents. Publication (b) (*Data services in academic libraries: Assessing needs and promoting services*) does not use data analysis as its research method; rather, it synthesizes literature and describes use cases. In publication (c) [28], there are both direct and indirect data references, where “we” indicates direct data references and “Eurobarometer”, a European public survey, indicates contextual sentences. Like publication (a), publication (c) (*Satisfaction with democracy and collective action problems: The case of the environment*) reports data analysis results and provides contextual information about the data used.

Semantic relations reveal author-data proximity in each sentence, which can then be aggregated into a network graph. The previous examples represent situations where the proximity of data references varies. When there are frequent uses of direct data references, the publication is likely to use data-driven research and analysis methods. If there are few or no direct data references, the publication is likely making contributions through other methods (e.g. synthesis of literature) rather than contributing new data analysis.

4.2 Data Reference *Functions*: Describing Interactions and Providing Context to Illustrate the Roles of Data in Research

Functions capture the roles that data references perform in the article. We identified two main functions of data references: to provide context or to indicate interactions with data. We found alignment between interaction and reuse purposes (e.g., performing a calculation) and between context and non-reuse purposes (e.g., giving credit to previous studies) previously defined in the literature we reviewed in Section 2.1.

An advantage of examining data reference functions from the perspective of datasets is that we can identify how different publications interact with or contextualize the same dataset. For example, 20 publications in our sample reference the National Health Interview Survey (NHIS) [54]. We observed four combinations of data reference roles in their semantic relations, (a) Direct + Interaction, (b) Indirect + Interaction, (c) Direct + Context, and (d) Indirect + Context. These example triples demonstrate the association between *proximity* and *functions*. We provide the semantic networks of this example dataset in Figure 5.

Most of the triples in the example are combinations (a) or (d). Combination (a) uses “we” as the subject, indicating that the author undertakes active operations on the dataset. These triples describe interactions with data, showing what authors did to the NHIS data or accomplished through interactions with other research materials (e.g., research software, research methods, or previous publications). Combination (d) contains information about data composition, origin, and collection methods, allowing readers to understand the background and provenance of the NHIS dataset. Some triples use “NHIS” as the subject, indicating how the “NHIS” dataset is useful for understanding data-related research. Other triples also provide background without using “NHIS” as the subject; for instance, "the survey included the variable age", used an instrument “containing questions derived from focus group findings”, and “collected information from the source of 402,154 respondents".

On the other hand, the smaller sets of triples with different role combinations are useful for surveying analytical methods and finding external resources. Combination (b) includes triples that show how people other than the authors use the data. For example, by referring to how “original samples" and “original estimates" are made using “NHIS” data, researchers can survey the analytical methods others have applied to the dataset. Combination (c) shows the external resources, including a checklist and a variable from another study, that have helped others use NHIS data. While these categories are not the core methods that a researcher should refer to or that a data repository should recommend, they add practical context about how others have used the data.

4.3 Classification and Interpretation of Semantic Relations in Data References

We found a correspondence between the *proximity* and the *function* of the data reference. Authors served as subject agents in only 17.7% of data references with **direct** and **interaction** roles. In other situations, authors served as object agents. This is partly due to the authors’ use of passive voice or because they were not mentioned in the sentences. These cases were challenging to extract using simple queries without performing additional semantic mining. We, therefore, omitted them. Nevertheless, the low proportion of author subjects shows that our semantic relation extraction method is unique; it extends the focus from authors’ active operations to broader author-data interactions. In general, the **indirect** and **context** data references (82.3%) either describe the source of the data or add details about a dataset’s composition. This insight aligns with prior findings from data reuse studies, which are reviewed in Section 2.1.

5 DISCUSSION

Many prior studies of data reuse have focused on data users’ motivations for seeking data [19, 20, 27, 57]. Our study focused on the interactions between authors and data expressed in published writing. We identified several ways that authors engage with research

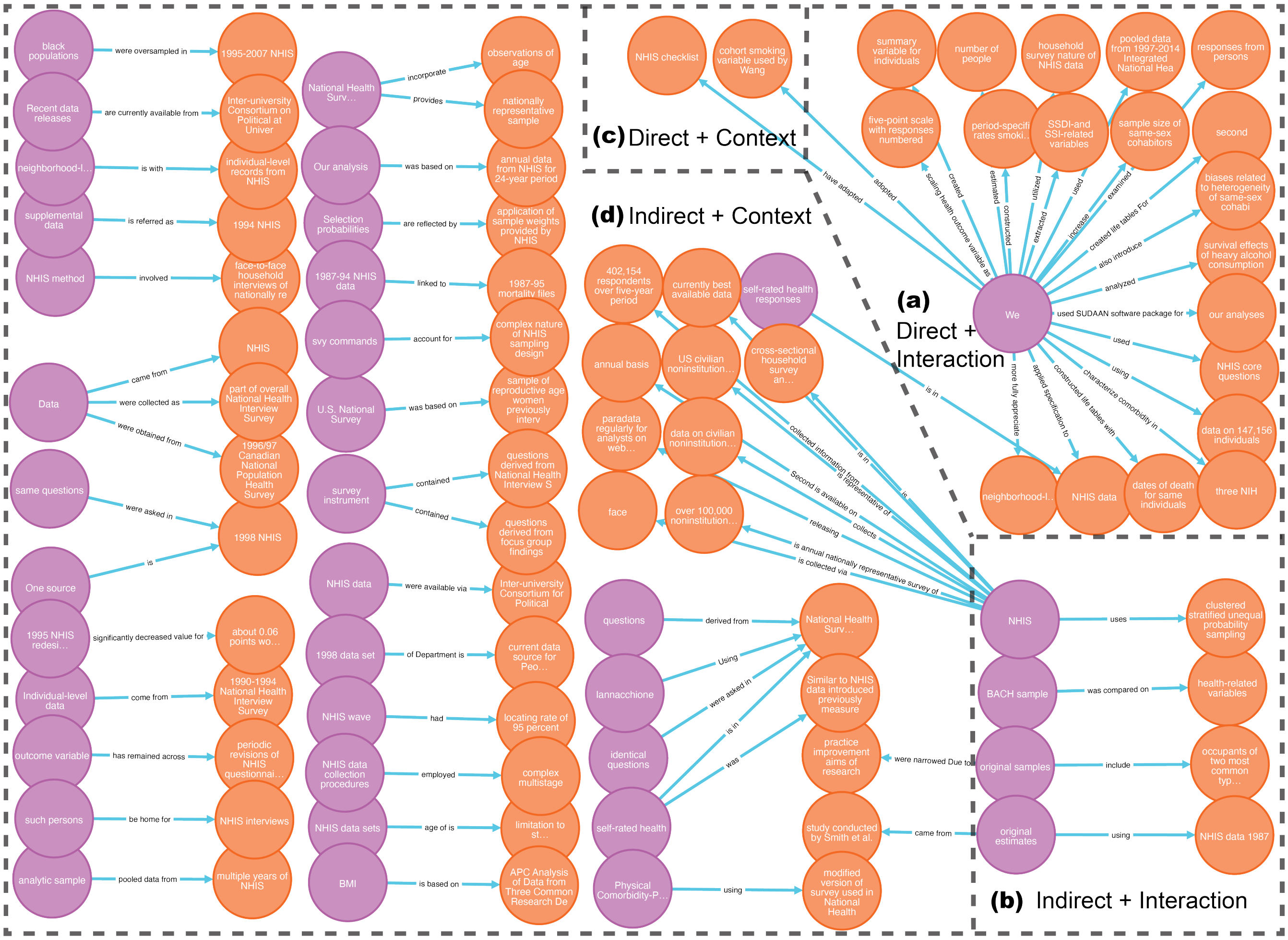
black populations
were oversampled in
1995-2007 NHIS
Recent data releases
are currently available from
Inter-university Consortium on Political at Univer
neighborhood-l…
is with
individual-level records from NHIS
supplemental data
is referred as
1994 NHIS
NHIS method
involved
face-to-face household interviews of nationally re
Data
came from
NHIS
were collected as
part of overall National Health Interview Survey
were obtained from
1996/97 Canadian National Population Health Survey
same questions
were asked in
1998 NHIS
One source
is
1995 NHIS redesi…
significantly decreased value for
about 0.06 points wo…
Individual-level data
come from
1990-1994 National Health Interview Survey
outcome variable
has remained across
periodic revisions of NHIS questionnai…
such persons
be home for
NHIS interviews
analytic sample
pooled data from
multiple years of NHIS
National Health Surv…
incorporate
observations of age
provides
nationally representative sample
Our analysis
was based on
annual data from NHIS for 24-year period
Selection probabilities
are reflected by
application of sample weights provided by NHIS
1987-94 NHIS data
linked to
1987-95 mortality files
svy commands
account for
complex nature of NHIS sampling design
U.S. National Survey
was based on
sample of reproductive age women previously interv
survey instrument
contained
questions derived from National Health Interview S
contained
questions derived from focus group findings
NHIS data
were available via
Inter-university Consortium for Political
1998 data set
of Department is
current data source for Peo…
NHIS wave
had
locating rate of 95 percent
NHIS data collection procedures
employed
complex multistage
NHIS data sets
age of is
limitation to st…
BMI
is based on
APC Analysis of Data from Three Common Research De
NHIS checklist
cohort smoking variable used by Wang
(c) Direct + Context
(d) Indirect + Context
402,154 respondents over five-year period
currently best available data
self-rated health responses
annual basis
US civilian noninstitution…
cross-sectional household survey an…
paradata regularly for analysts on web…
data on civilian noninstitution…
face
over 100,000 noninstitution…
collected information from
is representative of
is in
is
Second is available on
collects
releasing
is annual nationally representative survey of
is collected via
questions
derived from
National Health Surv…
Iannacchione
Using
were asked in
Similar to NHIS data introduced previously measure
identical questions
is in
was
practice improvement aims of research
self-rated health
study conducted by Smith et al.
Physical Comorbidity-P…
using
modified version of survey used in National Health
(a) Direct + Interaction
We
summary variable for individuals
number of people
household survey nature of NHIS data
pooled data from 1997-2014 Integrated National Hea
responses from persons
five-point scale with responses numbered
period-specifi rates smoki…
SSDI-and SSI-related variables
sample size of same-sex cohabitors
second
biases related to heterogeneity of same-sex cohabi
survival effects of heavy alcohol consumption
our analyses
NHIS core questions
data on 147,156 individuals
three NIH
dates of death for same individuals
NHIS data
neighborhood-l…
have adapted
adopted
scaling health outcome variable as
created
estimated
constructed
utilized
extracted
used
increase
examined
created life tables For
also introduce
analyzed
used SUDAAN software package for
used
using
characterize comorbidity in
constructed life tables with
applied specification to
more fully appreciate
is in
NHIS
uses
clustered stratified unequal probability sampling
BACH sample
was compared on
health-related variables
original samples
include
occupants of two most common typ…
were narrowed Due to
original estimates
using
NHIS data 1987
came from
(b) Indirect + Interaction

Figure 5: Semantic network of the NHIS dataset. Each pair of nodes and an arrow represent a triple extracted from a paper. NHIS is either a subject or object in each triple. Purple nodes are *subject agents*, orange nodes are *object agents*, blue edges indicate *agency*, and four combinations of *proximity* and *function* are placed in four dashed frames.

data. We leveraged semantic structures to distinguish subjects, objects, and their relationships. We decomposed data references into semantic units by mining semantic relations in data references. Categorizing these units provided a more comprehensive understanding of how data is used in scholarly work. Identifying the relationships between data and authors can help stakeholders assess the value of data in scientific knowledge production.

Data creators and funders, including government agencies (e.g. National Science Foundation (NSF) and National Institutes of Health (NIH) [24, 51, 53]) and private foundations (e.g., Bill & Melinda Gates Foundation and the Chan Zuckerberg Initiative [22, 35]), are invested in understanding the impacts of the science that they fund. Scalable, rule-based analysis of author-data interactions helped us identify types of data use that were previously hidden and, therefore, not recognized. Through the semantic analysis of subjects, objects, and their relations in data references, we contributed a novel workflow that objectively and comprehensively addresses the question of "what does it mean to reference data?". This workflow also automates the analysis of author-data interactions, increasing the potential for scaling and accelerating data-related analyses. Furthermore, the separate NER model we used to extract data references from publications was trained on a multidisciplinary corpus, ensuring the generalizability of our approach going forward.

5.1 Author as Data User: Crediting Data Work Described in Publications

Prior work has established a continuum of data reuse defined by users' goals, ranging from comparative (e.g., data used for groundtruthing) to integrative (e.g., data analyzed for correlations) [57]. Comparative data use involves indirect interactions with data, while integrative reuses are direct. This aligns with our findings from semantic mining. The proximity of data references also aligns with prior observations that differences in data acknowledgments often correspond with author intent (e.g., crediting data used in analysis versus data referenced to provide background information or make a claim) [19]. Data references often provide valuable background information or context used to inform the design of future studies; however, restricting credit to data analysis only means that other aspects of high-quality data provision, such as design and documentation, are not valued. Data creators and producers should receive credit for supporting scientific inquiry.

Our study treats data references and their semantic relations as "micro-narratives" containing explanations of research data's role in a given publication. When narrating data use, authors establish their relationship to data in direct and indirect ways. The distance between the author and the data (i.e., proximity) shows the authors' orientation to research data, indicated in part by the section of the article where the data is referenced. Since scholarly publications are typically awarded more credit than research data, variation in how datasets are used in publications has been largely ignored. The study of data references through semantic relations provides a common analytical unit for assigning credit to different kinds of data references. For instance, data use can be assessed through proximity. Comparative evaluation enables a more detailed estimation of who should receive credit for data work.

Moreover, the "value" of datasets can be qualified from the data user's perspective, providing insights into their needs, research methods, and the nature of their data interactions. Thus, descriptions of "data work" detailed in research publications can be better understood based on their data-related methods instead of their authors' academic affiliations or fields of research. In other words, the ways that authors actually use data, based on the kinds of research methods and analyses that the data support, can be used to characterize the scholarly impact of the data.

5.2 Dataset Utility Networks: Building on Prior Authors' Data Work

The semantic relations network aggregates the interactions between different data users and research materials, including the data itself, demonstrating their utility. Our analysis finds agreement with previous work by Gregory et al. [27], which assessed researchers' motivations when searching for data, and Jiao and Darch [36], which examined the role of data papers in scholarly communication. We found alignment between interactions and the reuse purposes that were previously defined (e.g., background information, calculation) as well as context and non-reuse purposes (e.g., giving credit to previous studies). Rather than assessing the specific purpose of each data reference sentence at a fine granularity, we used semantic mining to distinguish between interaction and context, capturing the general function of the reference.

The author-data interaction network is based on semantic relations and provides an overview of references made to a given dataset. Networks like these are "knowledge bases" of authors' statements about data, which can be decomposed and stored as facts. Large-scale, collaborative knowledge bases (KBs), such as Freebase and Wikidata support question-answering by traversing facts stored as relational triples (subject-relation-object) [73]. A knowledge base constructed from data references supports a closer look at the individual narratives that use the dataset and allows for a closer examination of data utility by switching perspectives between the author and the data. A user can understand how the data is typically used by analyzing clusters of relations (e.g., interaction or context). High-level views of data utility provide the history of data use, supporting impact analysis and inspiring future reuse.

Articulating the utility of research data is also valuable to various stakeholders involved in data work. Data providers, such as archives and repositories that curate data, analyze the scholarly impact of their collections by studying documented uses of their data. Similarly, researchers rely on data usage information to identify new data that can support their work. Funders have expressed interest in understanding how data from projects they fund are used in research [23]. Citation networks that include data references provide a high-level overview of data use. Such analyses show how well data fulfill their funders' goals based on their practical outcomes, such as community uptake.

5.3 Limitations and Outlook

Most machine learning and NLP models can handle semantics better than syntax. Our rule-based methods use both semantic and syntactic features, including part of speech tagging, dependency parsing, and rule-based subject and object classes for data citation. While our analysis can be scaled, it cannot achieve the same depth of detail uncovered in qualitative work. For example, different triples may use the same verb and semantic relations yet have different proximity and function: the triple "I"-"used"-"data from the General Social Survey" is an indirect data reference and provides a context function; however, the triple "we"-"use"-"survey-weighted inferences" is a direct data interaction reference[9]. At the same time, some triples are grammatically incorrect, including incorrect relation extractions and inaccurate tense and voice; this reduces the number of triples extracted from a data reference sentence and may lead to non-random information loss impacting the evaluation.

Going forward, machine learning models like SciBERT [5] and CiteBERT [76] that use transformer-enabled bi-directional deep learning can improve our current methods for making syntax-aware decisions and providing deeper insights into data reference functions. We will also simplify the classes of functions and consider what level or granularity of difference can be detected by machine learning models focused on semantics. Training a model using current rule-based automatic data annotation approaches helps improve the predictive accuracy of data reference functions. Furthermore, by implementing co-reference methods, we can automatically connect entities in data references (e.g., the nodes grouped with red dashed frames in Figure 4), which will enable a higher-level understanding of data-author interactions.

Given the challenges of distinguishing proximity and function in data references based solely on semantic relations, our future work will take a complementary, qualitative approach to study the rhetorical functions of data references. While we are limited in our ability to directly infer data reference purposes in detail, a closer reading of data references in context (e.g., the full-length academic articles in which they occur) will allow us to relate the syntactical structure of data references to their semantics[44]. This is important because it is difficult to identify and disambiguate "data-adjacent" language – such as variables or other descriptors that imply data use – that omit proper names or direct identifiers corresponding

[9] The corresponding sentences for these examples are in Appendix A.2

to datasets. Pairing qualitative approaches with our current rule-based, semantic methods will allow us to develop models that are more sensitive to implicit data references in context.

Our study characterizes the role of data references in scholarly communication. Future work will support comparisons in data references across time, fields of study, and authors. Thus, it is also essential to document and standardize the information contained within data references by constructing a scholarly knowledge base. A large-scale knowledge base of data references will help stakeholders understand the role of datasets in scholarly communication and can support graph-based embeddings to accelerate scientific discovery through dataset recommendations.

6 CONCLUSION

Data producers and providers need to understand how their data are used and the kinds of impacts their data have across the scholarly landscape. We demonstrated how semantic mining methods support data reference detection, extraction, and large-scale analysis. We identified two main relationships between authors and the data that they reference in their work. First, we showed how authors establish their relationship to data through proximity relations. When authors use active agency to refer to data as objects of discourse, they are in direct contact with the data. Examples of these references indicate that narratives about hands-on work with data tend to follow a similar structure, making it possible to computationally distinguish data analysis from other types of uses. Second, data references perform two main types of functions: either they describe researchers' interactions with data or they provide readers with additional context about the data. Only a small fraction of the references described direct data interactions, suggesting that narratives mentioning data more often orient the reader than describe data manipulation or analysis.

In summary, our approach increases data stakeholders' ability to track data use while distinguishing indirect and direct relationships between data and authors who reference them. Our work introduces possibilities for assigning distinct forms of credit to (a) data that are manipulated in analysis (e.g., in synthesis studies from which new data products are derived) and (b) data that are mentioned (e.g., for their high-quality designs, questions, or sampling strategies) but not analyzed. The semantic narratives in data references demonstrate the research impact of data by revealing the analyses and findings that the data support.

ACKNOWLEDGMENTS

Many thanks to David Bleckley and Elizabeth Moss of the MICA team at the Inter-university Consortium for Political and Social Research (ICPSR) for their support of this research. This material is based upon work supported by the National Science Foundation under grants no. 1930645 and no. 2121789.

REFERENCES

[1] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 596–606.

[2] Micah Altman, Christine Borgman, Mercè Crosas, and Maryann Matone. 2015. Data citation synthesis group: Joint declaration of data citation principles. *Bulletin of the Association for Information Science and Technology* 41, 3 (2015), 43–45.

[3] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 344–354.

[4] American Political Science Association. 1976. Supplementary Empirical Teaching Units in Political Science. *PS* 9, 3 (1976), 395–405. http://www.jstor.org/stable/418018

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

[6] Katarina Boland, Dominique Ritze, Kai Eckert, and Brigitte Mathiak. 2012. Identifying references to datasets in publications. In *Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, September 23-27, 2012. Proceedings 2*. Springer, Paphos, Cyprus, 150–161.

[7] Christine L Borgman. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press, Cambridge, MA.

[8] Jan Brase, Michael Lautenschlager, and Irina Sens. 2015. The tenth anniversary of assigning DOI names to scientific data and a five year history of DataCite. *DLib Mag.* 21, 1/2 (Jan. 2015).

[9] Peter Buneman, Dennis Dosso, Matteo Lissandrini, and Gianmaria Silvello. 2022. Data citation and the citation graph. *Quantitative Science Studies* 2, 4 (Feb. 2022), 1399–1422.

[10] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3586–3596. https://doi.org/10.18653/v1/N19-1361

[11] Jean-Claude Cosset, Charles Martineau, and Anis Samet. 2014. Do political institutions affect the choice of the US cross-listing venue? *Journal of Multinational Financial Management* 27 (2014), 22–48.

[12] Blaise Cronin. 1981. The need for a theory of citing. *J. Doc.* 37, 1 (Jan. 1981), 16–24.

[13] Blaise Cronin and Cassidy R Sugimoto. 2014. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. MIT Press, Cambridge, MA.

[14] Mihai Dascalu. 2014. Computational Discourse Analysis. In *Analyzing Discourse and Text Complexity for Learning and Collaborating: A Cognitive Approach Based on Natural Language Processing*, Mihai Dascălu (Ed.). Springer International Publishing, Cham, 53–77.

[15] Alfred DeMaris. 2014. Combating unmeasured confounding in cross-sectional studies: evaluating instrumental-variable and Heckman selection models. *Psychological methods* 19, 3 (2014), 380.

[16] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *J. Assoc. Inf. Sci. Technol.* 65, 9 (Sept. 2014), 1820–1833.

[17] Monica Duke and Alex Ball. 2012. How to cite datasets and link to publications: A report of the digital curation centre. In *23rd International CODATA Conference*. CODATA, Taipei, Taiwan.

[18] Lizhou Fan and Todd Presner. 2022. Algorithmic Close Reading: Using Semantic Triplets to Index and Analyze Agency in Holocaust Testimonies. *Digital Humanities Quarterly* 16, 3 (2022).

[19] Kathleen Marie Fear. 2013. *Measuring and Anticipating the Impact of Data Reuse*. Ph. D. Dissertation. University of Michigan.

[20] Lisa M Federer. 2019. *Who, what, when, where, and why? Quantifying and understanding biomedical data reuse*. Ph. D. Dissertation. University of Maryland.

[21] Martin Fenner, Mercè Crosas, Jeffrey Grethe, David Kennedy, Henning Hermjakob, Philippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone, and Timothy Clark. 2016. A Data Citation Roadmap for Scholarly Data Repositories. *Scientific Data* 6, 1 (Dec. 2016).

[22] Gates Foundation. 2021. Bill & Melinda Gates Foundation Open Access Policy. https://openaccess.gatesfoundation.org/

[23] National Science Foundation. 2011. Chapter II–Proposal Preparation Instructions. https://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp

[24] National Science Foundation. n.d.. Open Data at NSF. https://www.nsf.gov/data/

[25] Qin Gao. 2008. Social benefits in urban China: Determinants and impact on income inequality in 1988 and 2002. In *Understanding Inequality and Poverty in China*. Springer, Palgrave Macmillan, London, 173–217.

[26] Kathleen Gregory, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. 2019. Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. *J Assoc Inf Sci Technol* 70, 5 (May 2019), 419–432.

[27] Kathleen Gregory, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2020. Lost or found? Discovering data needed for research. *Harvard Data Science Review* 2, 2 (April 2020).

[28] Martin Halla, Friedrich G Schneider, and Alexander F Wagner. 2013. Satisfaction with democracy and collective action problems: the case of the environment. *Public Choice* 155, 1 (2013), 109–137.

[29] Karen L. Hanson, Tim DiLauro, and Mark Donoghue. 2015. The RMap Project: Capturing and Preserving Associations amongst Multi-Part Distributed Publications. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (Knoxville, Tennessee, USA) *(JCDL '15)*. Association for Computing Machinery, New York, NY, USA, 281–282. https://doi.org/10.1145/2756406.2756952

[30] Jenny Heddes, Pim Meerdink, Miguel Pieters, and Maarten Marx. 2021. The Automatic Detection of Dataset Names in Scientific Articles. *Data* 6, 8 (Aug. 2021), 84.

[31] Myriam Hernández-Alvarez and José M Gomez. 2016. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering* 22, 3 (May 2016), 327–349.

[32] Tony Hey, Stewart Tansley, and Kristin Tolle. 2009. *The Fourth Paradigm: data-intensive scientific discovery*. Microsoft Research, Redmond, Washington.

[33] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[34] Daniel W Hook, Simon J Porter, and Christian Herzog. 2018. Dimensions: building context for search and evaluation. *Frontiers in Research Metrics and Analytics* 3 (2018), 23. https://doi.org/10.3389/frma.2018.00023

[35] The Chan Zuckerberg Initiative. 2019. Essential Open Source Software for Science (EOSS)-Chan Zuckerberg Initiative, 05 2019.

[36] Chenyue Jiao and Peter T Darch. 2020. The role of the data paper in scholarly communication. *Proc. Assoc. Inf. Sci. Technol.* 57, 1 (Oct. 2020).

[37] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics* 6 (2018), 391–406.

[38] Dagmar Kern and Brigitte Mathiak. 2015. Are there any differences in data set retrieval compared to well-known literature retrieval?. In *Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, September 14-18, 2015, Proceedings 19*. Springer, Poznań, Poland, 197–208.

[39] Thomas Krämer, Andrea Papenmeier, Zeljko Carevic, Dagmar Kern, and Brigitte Mathiak. 2021. Data-seeking behaviour in the social sciences. *International Journal on Digital Libraries* 22 (2021), 175–195.

[40] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage, Thousand Oaks.

[41] William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience. *Journal of Narrative & Life History* 7, 1 (1997), 3–38.

[42] Sara Lafia, Lizhou Fan, and Libby Hemphill. 2022. A Natural Language Processing Pipeline for Detecting Informal Data References in Academic Literature. *Proceedings of the Association for Information Science and Technology* 59, 1 (2022), 169–178. https://doi.org/10.1002/pra2.614

[43] Sara Lafia, Lizhou Fan, Andrea Thomer, and Libby Hemphill. 2022. Subdivisions and Crossroads: Identifying Hidden Community Structures in a Data Archive's Citation Network. *Quantitative Science Studies* 3, 3 (May 2022).

[44] Sara Lafia, Andrea Thomer, Elizabeth Moss, David Bleckley, and Libby Hemphill. 2023. How and Why do Researchers Reference Data? A Study of Rhetorical Features and Functions of Data References in Academic Articles.

[45] Loet Leydesdorff. 1998. Theories of citation? *Scientometrics* 43, 1 (Sept. 1998), 5–25.

[46] Kai Li and Chenyue Jiao. 2022. The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts. *J. Assoc. Inf. Sci. Technol.* 73, 6 (June 2022), 834–846.

[47] Matthew S Mayernik, David L Hart, Keith E Maull, and Nicholas M Weber. 2017. Assessing and tracing the outcomes and impact of research infrastructures. *J. Assoc. Inf. Sci. Technol.* 68, 6 (June 2017), 1341–1359.

[48] Hailey Mooney. 2011. Citing data sources in the social sciences: do authors do it? *Learn. Publ.* 24, 2 (April 2011), 99–108.

[49] Hailey Mooney and Mark P. Newton. 2012. The anatomy of a data citation: Discovery, reuse, and credit. *J. Libr.. Inf. Sci.* 1, 1 (2012).

[50] Elizabeth Moss and Jared Lyle. 2018. Opaque data citation: Actual citation practice and its implication for tracking data use.

[51] Alondra Nelson. 2022. Ensuring free, immediate, and equitable access to federally funded research. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

[52] Allard Oelen, Markus Stocker, and Sören Auer. 2022. TinyGenius: Intertwining Natural Language Processing with Microtask Crowdsourcing for Scholarly Knowledge Graph Creation. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 5, 5 pages. https://doi.org/10.1145/3529372.3533285

[53] National Institutes of Health. 2023. 2023 NIH Data Management and Sharing Policy. https://oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy

[54] United States Department of Health and Human Services. National Center for Health Statistics. 1987. National health interview survey. https://doi.org/10.3886/ICPSR09195

[55] Hyoungjoo Park, Sukjin You, and Dietmar Wolfram. 2018. Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *J. Assoc. Inf. Sci. Technol.* 69, 11 (Nov. 2018), 1346–1354.

[56] Norman Paskin. 2005. Digital Object Identifiers for scientific data. *Data Sci. J.* 4 (2005), 12–20.

[57] Irene V Pasquetto, Christine L Borgman, and Morgan F Wofford. 2019. Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review* 1, 2 (Nov. 2019).

[58] Jean-Christophe Plantin. 2019. Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science. *Sci. Technol. Human Values* 44, 1 (Jan. 2019), 52–73.

[59] Jesse J Plascak, Yamile Molina, Samantha Wu-Georges, Ayah Idris, and Beti Thompson. 2016. Latino residential segregation and self-rated health among Latinos: Washington state behavioral risk factor surveillance system, 2012–2014. *Social Science & Medicine* 159 (2016), 38–47.

[60] David Pride and Petr Knoth. 2020. An Authoritative Approach to Citation Classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 337–340.

[61] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

[62] J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning* 1 (1986), 81–106.

[63] Eleanor J Read. 2007. Data services in academic libraries: Assessing needs and promoting services. *Reference & User Services Quarterly* 46, 3 (2007), 61–75.

[64] John Robst, Solomon Polachek, and Yuan-Ching Chang. 2007. Geographic proximity, trade, and international conflict/cooperation. *Conflict Management and Peace Science* 24, 1 (2007), 1–24.

[65] Belen Saldias and Deb Roy. 2020. Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*. Association for Computational Linguistics, Online, 78–86. https://doi.org/10.18653/v1/2020.nuse-1.10

[66] Tracy Schifeling, Jerome P Reiter, and Maria Deyoreo. 2019. Data Fusion for Correcting Measurement Errors. *Journal of Survey Statistics and Methodology* 7, 2 (2019), 175–200.

[67] Gianmaria Silvello. 2018. Theory and practice of data citation. *Journal of the Association for Information Science* 69, 1 (2018), 6–20.

[68] Henry G Small. 1978. Cited Documents as Concept Symbols. *Soc. Stud. Sci.* 8, 3 (Aug. 1978), 327–340.

[69] Reid Swanson, Elahe Rahimtoroghi, Thomas Corcoran, and Marilyn Walker. 2014. Identifying Narrative Clause Types in Personal Stories. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, Philadelphia, PA, U.S.A., 171–180.

[70] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sydney, Australia, 103–110.

[71] Terry Therneau, Beth Atkinson, Brian Ripley, and Maintainer Brian Ripley. 2015. Package 'rpart'. https://cran.pau.edu.tr/web/packages/rpart/rpart.pdf

[72] Andrea K Thomer, Dharma Akmon, Jeremy York, Allison R B Tyler, Faye Polasak, Sara Lafia, Libby Hemphill, and Elizabeth Yakel. 2022. The craft and coordination of data curation: complicating "workflow" views of data science. *Proceedings of the ACM on Human Computer Interaction (PACM HCI)* 6, 414 (2022), 1-29 pages.

[73] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd international conference on World wide web*. Association for Computing Machinery, New York, NY, USA, 515–526.

[74] Mark D Wilkinson, Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S Grethe, Jaap Heringa, Peter A C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J Lusher, Maryann E Martone, Albert Mons, Abel L Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine

Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016).

[75] Morgan F Wofford, Bernadette M Boscoe, Christine L Borgman, Irene V Pasquetto, and Milena S Golshan. 2020. Jupyter Notebooks as Discovery Mechanisms for Open Science: Citation Practices in the Astronomy Community. *Comput. Sci. Eng.* 22, 1 (Jan. 2020), 5–15.

[76] Dustin Wright and Isabelle Augenstein. 2021. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 1796–1807. https://doi.org/10.18653/v1/2021.findings-acl.157

[77] Laura A Wynholds, Jillian C Wallis, Christine L Borgman, Ashley Sands, and Sharon Traweek. 2012. Data, data use, and scientific inquiry: two case studies of data practices. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (Washington, DC, USA) *(JCDL '12)*. Association for Computing Machinery, New York, NY, USA, 19–22.

A EXAMPLE SENTENCES AND TRIPLE NETWORKS

A.1 Example Sentences for Agency

The following example sentences correspond to the five triples in Section 3.2 in the order of occurrence:

(1) We obtain these data from the Mannheim Eurobarometer Trend File, 1970 File, -2002 This is an integrated set of data covering harmonized variables for the years 1970 through 2002 that allow a cross-time (and cross-country) comparison. (This sentence is from publication [28].)
(2) ADR data are collected from the major depository bank websites: Bank of New York, Citibank, Deutsche Bank, and JPMorgan. (This sentence is from publication [11].)
(3) SETUPS data can be ordered for use with SPSS, OSIRIS, or as a card image. (This sentence is from publication [4].)
(4) The BRFSS is the largest, continuously conducted, telephone health survey in the world (Centers for Disease Control and Prevention (CDC) 2012). (This sentence is from publication [59].)
(5) Following Reuveny (2003:255) who argues "it would be beneficial for the field of international relations to go back and routinely use events data" our primary source of data on conflict and cooperation between dyads is the World Event Interaction Survey (WEIS). (This sentence is from publication [64].)

A.2 Example Sentences for Limitations

The following example sentences correspond to the two triples in Section 5.3 in the order of occurrence:

(1) To illustrate the application of IVR and HSM, I used data from the General Social Survey (GSS). (This sentence is from publication [15].)
(2) First, we use survey-weighted inferences to estimate population totals of (Y | X) from the 2010 NSCG. (This sentence is from publication [66].)

B ANNOTATION RESULTS

Figure 6 shows the annotation results that both annotators agree on. We use these results to create the decision trees in Section 3.3.

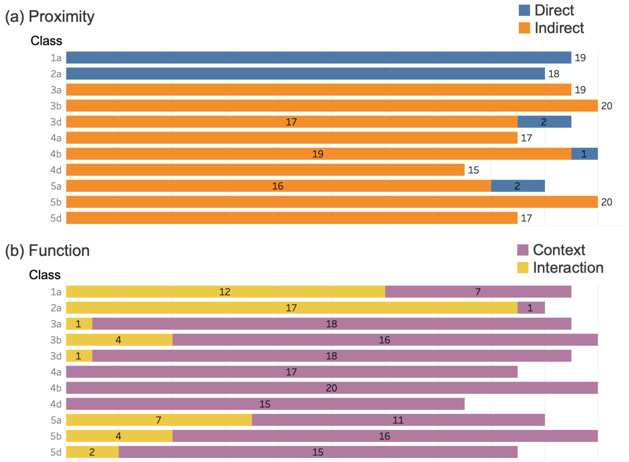

Figure 6: The annotated samples

C AGENCY CLASSIFICATION ALGORITHM

Algorithm 1 shows the details of agency classification. This algorithm is converted from the Action, Orientation, and Evaluation (AEO) characterization algorithm [18] and is based on spaCy [33], an NLP package in Python.

Algorithm 1 Agency classification algorithm based on semantic structures

```
Require: spaCy Tokens for a Relation $R$ and spaCy Tokens for an Object $O$, Evaluation Verbs $list_{evaluation}$, Orientation Verbs $list_{orientation}$, Possible Action Verbs $list_{posact}$
Ensure: Triples Agency class $C_{AEO}$;

# Step 1: Initialization
Initiate and assign 0 to each of the following variables $r_{has\_evaluation}$, $r_{has\_orientation}$, $r_{has\_posact}$, $r_{has\_be}$, $r_{has\_have}$, $r_{has\_to}$, $r_{has\_neg}$, $r_{has\_VBG}$, $r_{num\_verb}$, $o_{is\_adj}$, and $o_{has\_no}$
# Step 2: Value Assignments for $R$
for $r \in R$ do
        Add the number of part of speech tags of $VERB$ or $AUX$ to $r_{num\_verb}$
    if lemma of $r \in list_{evaluation}$ then Assign 1 to $r_{has\_evaluation}$
    else if lemma of $r \in list_{orientation}$ then Assign 1 to $r_{has\_orientation}$
    else if lemma of $r \in list_{posact}$ then Assign 1 to $r_{has\_posact}$
    else if lemma of $r\ is\ word\ be$ then Assign 1 to $r_{has\_be}$
    else if lemma of $r\ is\ word\ have$ then Assign 1 to $r_{has\_have}$
    else if lemma of $r\ is\ word\ to$ then Assign 1 to $r_{has\_to}$
    else if semantic dependency tree tag of $r\ is\ label\ neg$ then Assign 1 to $r_{has\_neg}$
    else if semantic dependency tree tag of $r\ is\ label\ VBG$ then Assign 1 to $r_{has\_VBG}$
    end if
end for
# Step 3: Value Assignments for $O$
for $o \in O$ do
    if lemma of $o\ is\ word\ no$ then Assign 1 to $o_{has\_no}$
    end if
end for
for $o \in O$ do
    if part of speech tagger of $o\ is\ label\ ADJ$ then Assign 1 to $o_{is\_adj}$
    end if
    if part of speech tagger of $o \in labels\ NOUN, PROPN, PRON$ then Assign 0 to $o_{is\_adj}$ and end For loop
    end if
end for
# Step 4: Agency Class Decision
if $r_{has\_evaluation}$ and $o_{is\_adj}$ then $C_{AEO}$ = Evaluation
else if $r_{has\_posact}$ then $C_{AEO}$ = Agency_Possible
else if $r_{has\_orientation}$ then $C_{AEO}$ = Orientation
else if $r_{has\_neg}$ or $o_{has\_no}$ then $C_{AEO}$ = Orientation
else if $r_{has\_have}$ then
    if $r_{has\_to}$ then $C_{AEO}$ = Agency_Passive
    else $C_{AEO}$ = Orientation
    end if
else if $r_{has\_be}$ then
    if $o_{is\_adj}$ then $C_{AEO}$ = Evaluation
    else if $r_{has\_VBG}$ then $C_{AEO}$ = Agency_Active
    else if $r_{num\_verb} > 1$ then $C_{AEO}$ = Agency_Passive
    else if $r_{num\_verb} = 1$ then $C_{AEO}$ = Orientation
    end if
else $C_{AEO}$ = Agency_Active
end if
```