Binding Data Narrations – Corroborating the Plausibility of Scientific Narratives by Open Research Data

Denis Nagel
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany
nagel@ifis.cs.tu-bs.de

Till Affeldt
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany
t.affeldt@tu-bs.de

Nils Voges
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany
n.voges@tu-bs.de

Ulrich Güntzer
Institute for Informatics,
University of Tübingen
Tübingen, Germany
ulrich.guentzer@informatik.uni-tuebingen.de

Wolf-Tilo Balke
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany
balke@ifis.cs.tu-bs.de

ABSTRACT

Narratives determine our worldviews, but need evidence to be believable. For *scientific narratives*, hard evidence usually is provided by specially curated research data and experiments within each publication. The aim is to strengthen the narratives' overall *plausibility*: the less plausible a narrative is, the more it is bound to be challenged. As recent advances in NLP enabled the scientific community to tackle important problems in fact-checking and more profound semantic interpretations of structured data, there is now also hope to unlock the rich narrative knowledge that such data sets can offer. Yet, current strategies to extract such narrative knowledge still heavily rely on exhaustive bottom-up analysis to cast insights from data into a human-understandable form. In this paper, we take a novel integration-based approach to design a system that reduces the task of finding narrative evidence to applying a sequence of simpler top-down matching tasks. Our BiND system builds upon an expressive definition of structured narratives. It uses them as templates for schema and instance matching against web tables, thereby computing flexible bindings between narratives and data. By combining structured narratives with a carefully chosen selection of statistical metrics to assess the inherent relationships between different attributes of the matched data, our system allows us to reliably identify the most plausible witnesses for a given narrative. We demonstrate the applicability of our system in the real world on the vast open data repository of the World Health Organization (WHO).

CCS CONCEPTS

• **Information systems** → **Retrieval tasks and goals**; *Digital libraries and archives.*

KEYWORDS

Narrative Bindings, Research Data, Digital Libraries

1 INTRODUCTION

Today's research produces a vast selection of interesting scientific narratives every day. These narratives are introduced into the scientific community through a plethora of scholarly objects, in particular research publications. Some of them become well-known and widely accepted over time. The plausibility of a narrative connecting smoking with an increased risk of lung cancer would most likely not be questioned by many people. However, this is usually not the case for the vast majority of scientific narratives, and concerns regarding their plausibility and credibility can quickly arise. As an example, consider the following narrative: *Links between smoking and TB have long been suspected, but new studies, coupled with reviews of older research, have provided definitive evidence of the connection. In addition to influencing the risk of contracting TB, developing the active form of the disease and ultimately dying from it, smoking is known to negatively influence the response to treatment and to increase the risk of relapse. Furthermore, recent studies have shown that second-hand smoke increases the risk of contracting the disease, especially in children*[1]. Except for domain experts, assessing the plausibility of such a narrative is a very challenging task.

To be believable, a narrative requires evidence and stringent argumentation to support its claims. In science, this evidence often boils down to data collected through studies or experiments. Recently, digital libraries have transformed from purely textual document collections to integrate associated data publications [15, 30]. An example is IRIS (Institutional Repository for Information Sharing) [41], the digital library of the World Health Organization (WHO), which offers access to a vast selection of public health documents. It is accompanied by the GHO (Global Health Observatory) [40], which collects and provides the data to support these documents.

This transformation process faces many problems, from documents lacking any associated data publications to missing links between document and data [15]. Even if the data is available and correctly linked to its associated publication: only considering the original data set might not be sufficient to assess a narrative's plausibility. The question of generalizability to similar settings might arise, or the derived conclusions might just not fit the current state of the art and may need further explanation. In these cases, consulting *validated external data* might provide the missing evidence.

[1] https://www.who.int/europe/news/item/22-03-2018-smoking-and-tuberculosis-a-dangerous-combination

However, to draw knowledge from all this data (original or external), its respective insights need to be extracted and represented in a human-understandable form. Unfortunately, this is very demanding since raw statistical research data is inherently difficult to interpret. This is especially true when the context of its creation is unknown to the user. Coupled with the high heterogeneity between data sets, revealing beneficial yet hidden insights is challenging even for state-of-the-art processes that are only beginning to profit from the first steps of automation. Until recently, classical bottom-up approaches such as data visualization [34, 35] were necessary to present interesting findings to the broader scientific community. These approaches are coupled with high manual expenditures, and still, many possible findings might be simply overlooked.

With the advent of language models and the general progress in the area of NLP, we are now able to directly leverage the raw data for various tasks. This leads to new and effective strategies for data interpretation, data set exploration and web table querying. Especially new systems for fact-checking [3, 12, 19] offer powerful tools to verify statements encountered in the scientific discourse. While these systems work well for simple factual claims, they usually lack the capabilities to assess complex narratives that embed multiple individual claims into a coherent context. They are often limited to individual data tables or focus solely on explicitly expressed relations. Additionally, the goal of fact-checkers is the verification of statements, which is a binary decision about their general truth, which is usually not possible for narratives.

Our goal is to provide users with a system to link a narrative to supporting data sets, allowing them to make reasoned decisions about its plausibility. Recently, we proposed using any narrative as a query template to identify relevant web tables and data sets[29]. We follow this conceptual idea by adopting the notion of structured narratives (see, e. g., [22]) but casting the proposed workflows into a practical system.

This paper presents BiND (Binding Narratives to Open Data), a novel top-down system for computing flexible bindings between scientific narratives and open research data. It casts the high-level task of computing narrative bindings [22], i.e., establishing supportive evidence between a scientific narrative and research data, as two simple and quite effectively solvable *semantic matching tasks*:

- In the first step, BiND deploys of-the-shelf *language models* in the form of state-of-the-art *sentence embeddings* to identify candidate data sets for a given narrative and the relevant attributes within. Because some narratives need additional qualifiers that may not apply to the whole data set, BiND subsequently enhances the schema matching by an *instance matching* to precisely select those parts satisfying the narrative's closer context. Consequently, BiND restricts potential data sets horizontally and vertically, leaving only those parts relevant to the narrative.
- In the second step, BiND uses the selected data to assess all of the narrative's claims. These can be expressed either explicitly or by *inherent statistical relationships* between those parts of the data sets identified as relevant in the first step. As an experimental setting in this paper, we focus on the biomedical domain.

Our central contribution is a practical system, casting the problem of computing semantic bindings as a two-step matching problem. We discuss all subtasks and their respective algorithms in detail, including an extensive discussion of open problems. Finally, we provide a proof-of-concept evaluation of real-world data repositories using the *Global Health Observatory* data sets provided by the World Health Organization (WHO).

2 NARRATIVES, RESEARCH DATA, AND BINDINGS

In this section, we introduce a model for narrative graphs and define the task of computing narrative bindings by building on the notations introduced in [22]. We use the last sentence of the introductory narrative as a running example: *Furthermore, recent studies have shown that second-hand smoke increases the risk of contracting the disease, especially in children.*

2.1 An Introduction to Narratives

Humans always expressed ideas, exchanged knowledge, or described observations by telling *stories*. These stories often go beyond the postulation of simple claims by combining individual claims and embedding them into a specific context to form complex narratives. The cornerstones of any narrative are individual *events*. These events can either describe specific occurrences, e. g., someone contracting tuberculosis, or they can be rather abstract. In our running example, such an abstract event would be *contracting tuberculosis* which incorporates all instances of people contracting the disease. We denote the set of all events by $\mathcal{E}$. One of the main aspects of narratives is the formulation of progression, either in a temporal or causal sense. This information is expressed via *narrative relations* that connect the individual events. In our example, the *exposure to smoking increasing the risk of contracting tuberculosis* is such a *narrative relation*. Hence, we define the set of narrative relations as

- $RN \subseteq \mathcal{E} \times \Sigma_N \times \mathcal{E}$

with Σ_N being the alphabet of narrative relation labels. Finally, a narrative can include factual information. It can be used to formulate contextual *constraints*, e. g., associate an event with a location or a group of people, or to formulate factual *claims*, e. g., 0.329% of the population in Angola contracted tuberculosis in 2020. This is expressed by using *factual relations* associating events with entities, entity types, or numerical information. Additionally, these relations can also describe the entities and types by stating properties, such as the birth date of a person, as well as relationships, e. g., between parent and child. Let $\mathcal{O}$ be the set of all entities/entity types and $\mathcal{L}$ be the set of all literals, i. e., numerical and lexicographical strings. We define the set of factual relations as

- $RF \subseteq (\mathcal{E} \cup \mathcal{O}) \times \Sigma_F \times (\mathcal{O} \cup \mathcal{L})$,

with Σ_F being the alphabet of factual relation labels. In our example, the contextual information of the narrative can be expressed using factual relations connecting each of the two events with the entity type *children*.

To facilitate an automated computation of narrative bindings, it is necessary to cast narratives into a structured format. For this, we define *narrative graphs* modeling narratives as directed edge- and node-labeled graphs.

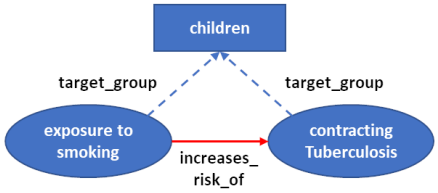

Figure 1: The narrative graph of our running example, consisting of two events connected by a narrative relation. Both events are contextualized by factual relations (dashed).

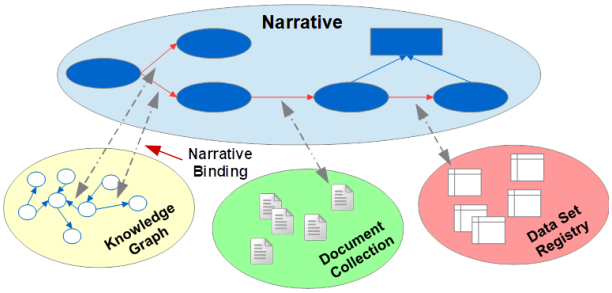

Figure 2: A narrative used as a logical overlay on top of knowledge repositories as proposed in [22]. Each binding connects the narrative with scientific evidence.

Definition 2.1. We define a narrative graph n as $n = (V, E)$ with

- $V \subseteq \mathcal{E} \cup \mathcal{O} \cup \mathcal{L}$ being nodes
- $E \subseteq V \times (\Sigma_N \cup \Sigma_F) \times V$ being edges such that for each $e \in E$ it holds that $e \in RN$ or $e \in RF$

To reference the components of a specific narrative n, we use a subscript n, e. g., $\mathcal{E}_n$ for its events. Figure 1 shows a narrative graph for our introductory example.

2.2 Problem Definition

Recently, we postulated the idea of using a narrative as a logical overlay on top of different knowledge repositories, e. g., document collections, data sets, knowledge graphs, etc. [22]. The goal is to find connections between different parts of the narrative overlay and connect them with those repositories that provide information to support their claims. These connections are then called *narrative bindings*. See figure 2 for a visualization. In this paper, we build upon this idea but focus solely on structured data sets as the underlying repositories. Additionally, we also aim to find connections between a narrative and those data sets that directly contradict its claims. While the structure and presentation of tabular data are highly heterogeneous, for the scope of this paper, we only consider those data sets representing their data in a relational format, i. e., the values of each individual record are semantically connected. We define a data set as follows:

Definition 2.2. A data set d with c columns and r rows is defined by $d \subseteq Dom_1 \times ... \times Dom_c$ with

- the set $A_d := \{a_1, ..., a_c\}$ of d's attributes with $Dom_1, ..., Dom_c$ as respective value domains

Country	Year	Percentage of children exposed to second-hand smoke in the home	Incidence of Tuberculosis in children (aged 0-14)
Argentina	2012	33.8	0.024
Bangladesh	2009	57.0	0.221
China	2010	66.7	0.076
Egypt	2009	64.1	0.019
(...)	(...)	(...)	(...)

Figure 3: A data set that can be used as a witness for the introductory narrative in figure 1 to support or contradict its claims.

- a set of records $R_d = \{r_j := (v_{ja_1}, v_{ja_2}, ..., v_{ja_c})\}$ $(1 \leq j \leq r)$, where v_{ja_i} denotes the j-th record's value for the i-th attribute

The contents of a data set can additionally be described by a caption which we denote as cap_d.

We can now define the task of *computing narrative bindings for data narrations* as follows: Given a narrative graph n and a repository $D = \{d_1, ..., d_m\}$ of data sets, the task is to find those data sets $d_i \in D$ that provide evidence for or contradict n (or parts of n). We denote such a data set d_i as a *witness* for n. As an example consider figure 3, which shows a data set acting as a witness for the narrative graph in figure 1.

Before we can formally define narrative bindings, we make two assumptions: (1) Events of a narrative usually translate to attributes of a data set. Data sets are mainly used to store empirical and statistical information gathered throughout experiments or studies. These experiments often observe specific events in differing contexts. When storing the results in a data set these events correspond to a specific attribute, while each record of the data set represents a different occurrence of that event, i. e., a context in which this event is observed. In our example, the third and fourth columns represent the events of children being exposed to second-hand smoke and children contracting tuberculosis, respectively. (2) A data set is *context-compatible* to a narrative n if it respects all constraints expressed through factual relations in n. In our example, the narrative constrains both events to the context of children. The data set in figure 3 is context-compatible as the attributes matching these events also adhere to this context.

Hence, by expanding the definition in [22] accordingly, we define narrative bindings as follows:

Definition 2.3. Given a narrative graph n and a data set repository $D := \{d_1, ..., d_m\}$, a narrative binding $nb_n(D)$ of D to n exists, iff

- a function $eb : \mathcal{E}_n \rightarrow \{A_{d_1}, ..., A_{d_m}\}$ is able to map each event in n to columns of some data sets in D
- these data sets are context-compatible to n

We call D *bound* to n if $nb_n(D)$ exists. Finally, we need to determine, whether the bound repository *supports* or *contradicts* the given narrative. We consider a narrative n to be supported by a bound repository D, if D can provide evidence for each of the narrative relations and claims expressed by factual relations in n.

Definition 2.4. Given a narrative binding $nb_n(D)$. We call n to be *supported* by D iff

- for every factual relation $r_{cl} = (e_j, l_{cl}, p) \in RF_n$ expressing a claim, the requirement expressed by the label l_{cl} can be met by $eb(e_j)$
- for every narrative relation $r_{nr} = (e_k, l_{nr}, e_l) \in RN_n$, the dependency expressed by the label l_{nr} can be statistically observed between $eb(e_k)$ and $eb(e_l)$

Otherwise, we call n to be *contradicted* by D.

For an intuition consider the narrative relation *(exposure to smoking, increases_risk_of, contracting Tuberculosis)* from our running example. Suppose the data set in figure 3 is bound to the narrative by mapping the third and fourth columns to its two events. Then the data set supports the narrative if a positive influence from the third to the fourth column can be statistically observed.

3 RELATED WORK

To the best of our knowledge, our BiND system represents the first approach to automatically compute narrative bindings against structured data sets. However, the computation of bindings against textual data provided by document collections has recently been explored in [23, 24]. Currently, only a few models consider data narrations, with one recent example being introduced in [8]. The motivation is similar to our approach, but the use case differs quite drastically. However, many related tasks have recently received much attention.

Argumentation mining aims at identifying the argumentative structure of scientific publications [1]. The goal is to identify arguments or claims within a text document, the relations between them, and each argument's internal structure [10, 11, 31]. Hence, argumentation mining might be a valuable starting point for the extraction of narratives from natural language.

While argumentation mining is generally applied to unstructured texts, many approaches aim to extract knowledge from structured data. However, like [16, 17], they usually focus on explicitly expressed factual knowledge. Contextual information or inherent implicit relations are often not considered.

Causal relationships which play an important role in many narratives are especially difficult to detect from raw data alone. This has led to a multitude of strategies for causality detection, including Bayesian networks [13, 28, 36] and data mining techniques [25, 26]. While BiND does not support binding causal relationships, advances in causal discovery might enable this support in the future.

While the previous works all focused on information extraction, much effort has also been put into data analysis to facilitate access and interpretation of open research data. These approaches face similar problems to BiND, including the heterogeneity of data sources and lack of a formal data set schema [37]. They often rely heavily on entity linking [2, 5, 27] or directly transform the data into RDF-triples [37]. Contrary to these approaches BiND does not aim to interpret individual data sets bottom-up, but rather to identify data to support a given narrative in a top-down approach.

Another similar research topic regarding data narrations is data visualization. Most commonly, the goal is to convey insights from data sets to a target group [34, 35, 39, 42]. The main difference compared to computing narrative bindings, is that these insights are obtained through bottom-up data analysis focusing on the data set, while narrative bindings are computed top-down, focusing on the narrative.

When confronted with vast repositories, effective data discovery is essential in identifying relevant data sets. This includes the difficulty of querying unknown relational data sets. Text-to-SQL Advances such as [14, 21] tackle this problem by allowing automated translation of natural language queries into suitable SQL statements. As complex information needs often require more than one table to provide the desired knowledge, a crucial task is also to identify join-compatible data sets [4, 6, 9]. Contrary to pure data discovery strategies, BiND combines data discovery with approaches for data interpretation to successfully compute bindings between a narrative and data.

Fact-checking approaches offer a way to quickly verify a claim against a trusted source of truth, usually a collection of textually represented facts [12] or tabular sources [3]. They provide a useful tool against misinformation and can verify factual claims about an entity's attributes (e. g., "John E. was re-elected in 1972") as well as claims about aggregated attributes [19] (e. g., "Three democratic candidates were elected in 1972"). However, they usually only work on a single provided source of information, and explicitly stated facts. Information represented through inherent statistical dependencies is usually ignored. BiND on the other hand, supports the assessment of both explicitly and implicitly expressed knowledge that could span multiple tables coupled with a data exploration step to automatically identify relevant data sets.

4 BIND: COMPUTING FLEXIBLE BINDINGS BETWEEN NARRATIVES AND STRUCTURED DATA

In the following, we outline BiND, a system for automated computation of narrative bindings. We describe challenges and subtasks in detail. Figure 4 shows the system overview.

4.1 Input

BiND requires two types of input. First, BiND needs a narrative graph n provided by the user, which we refer to as the *narrative query*. Its format is similar to the Resource Description Framework (RDF), describing each relation as a subject-predicate-object triple. Additionally, the relation type, i. e., narrative relation, factual claim, or constraint, needs to be annotated to each triple. Every narrative query that suffices definition 2.1 is accepted by the system. For now, we rely on a manual construction of narrative queries, i. e., translating natural language texts into narrative graphs, as an automated extraction of narratives from text is still a subject of ongoing research [18]. This manual process consists of three simple steps: (1) The main events of interest must be identified. (2) Relevant factual relations can be added to each event and (3) narrative relations can be added between each pair of events.

Furthermore, a data set repository $D = \{d_1, d_2, ..., d_m\}$ has to be specified. This repository can contain any relational data set $d_i \in D$ that conforms to definition 2.2. BiND will only consider those $d_i \in D$ to find witnesses for n.

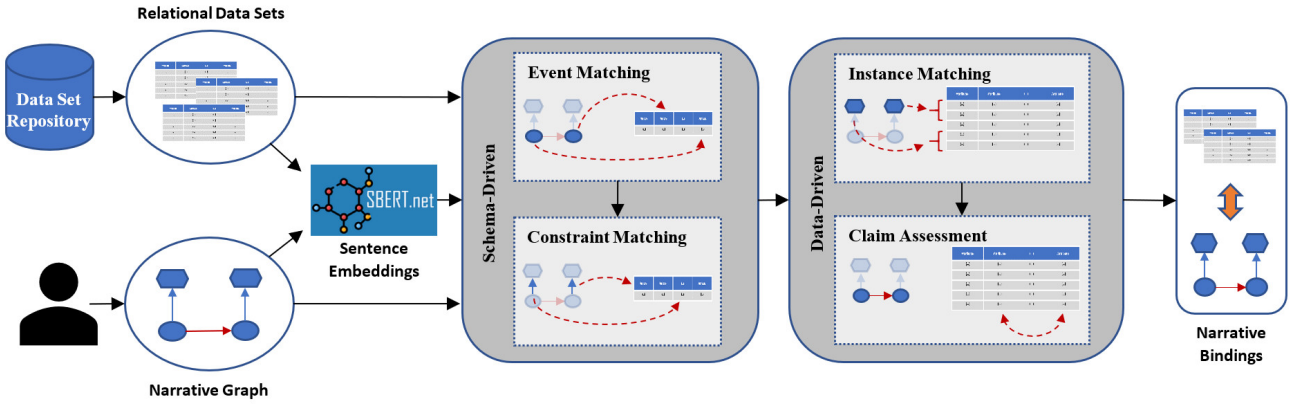

Figure 4: Overview of the BiND system

4.2 Event Matching

The first challenge of our matching pipeline consists of the *event matching*, i. e., identifying those attributes across all data sets in D that correspond to an event $e \in \mathcal{E}_n$. We call such an attribute a *quantifying attribute* for e. In our running example, the attribute *Incidence of Tuberculosis in children (aged 0-14)* in figure 3 can be considered as quantifying attribute for the event *contracting Tuberculosis* in figure 1. The goal of the event matching is twofold. It aims to (1) find an entry point into the data by identifying those data sets $d_i \in D$ that could be candidate witnesses for n and (2) exclude as many irrelevant data sets as possible as early as possible to reduce computational overhead in downstream tasks. The focus lies on events, as they form the focal point in any narrative structure. It is important to note that each $e \in \mathcal{E}_n$ can be mapped to as many attributes as possible (but only one per data set). The event matching function of BiND is defined as follows.

- $em_n : \mathcal{E}_n \rightarrow \mathcal{P}(\{A_{d_i} | d_i \in D\})$

Let $e \in \mathcal{E}_n$ and $d_i \in D$. After the event matching, d_i is called a *candidate witness* for n if for any $a \in A_{d_i}$ it holds that $a \in em(e)$.

For the event matching, the challenge is to identify the semantics for a data set's attributes. Essentially, we can draw from three parts of a data set to get insight into these semantics. (1) Its schema, i. e., attribute names, (2) its extension, i. e., its record's values, and (3) surrounding text, i. e., meta-data.

If available, attribute names can be a prime candidate to derive the semantics of their respective attribute. Unfortunately, in open data, the quality of attribute names is highly heterogeneous and many data sets use only abstract and very inexpressive attribute labels, such as *"Value"*, *"Indicator"* or *"Dimension"*. This problem has already been extensively discussed in the literature [33, 38].

Attribute extensions only help in an entity-centric context, i. e., in data sets where most of the values represent entities. When most values consist of numerical data, only their distribution can give an insight into an attribute's semantics.

Finally, the textual surroundings of a data set can offer information that can not be derived from the data itself. However, BiND needs to identify the semantics of each individual attribute and not the data set as a whole. Hence, the challenge is to identify those parts referring to a specific attribute.

Due to these considerations, BiND takes a hybrid approach. It mainly relies on attribute names but also considers textual surroundings. We decided not to rely on attribute extensions to identify candidate witnesses purely schema-driven. Analyzing the actual data would not be feasible due to the large number of data sets in D as well as their possibly vast size.

Algorithm 1 describes the event matching. For each $e \in \mathcal{E}_n$, BiND iterates over all data sets $d_i \in D$ and conducts a pairwise comparison between each attribute $a \in A_{d_i}$ and e. To assess their semantic similarity, BiND relies on the state-of-the-art sentence transformer [32]. It is used to compute the *cosine similarity* between all pairs of attribute names and event labels, denoted by $coSim(e, a) = sc$ with $sc \in [-1, 1]$. For each data set d_i and event e, BiND identifies the attribute a_{max} resulting in the highest similarity. If this similarity is higher than a threshold θ_{EM}, the pair (e, a_{max}) is considered as an event match and d_i as a candidate witness for n.

Else, BiND considers the cosine similarity between the data set's caption cap_{d_i} and the label of e and checks whether it exceeds θ_{EM}. In this case, d_i is considered a *preliminary match* for e as the event could not be directly matched to a specific attribute. The remaining data sets (neither candidate witness nor preliminary match) are discarded and not used in any downstream tasks. We denote the set of all event matches by EM and the candidate witnesses for n by $Cand_n \subseteq D$.

It is important to note that (1) each event of the narrative can be matched to at most one attribute of each data set, and (2) different events of the same narrative can be matched against attributes of different data sets. We made the first decision as we assume no two attributes of a single data set to conform to the same event. While more complex events might only be expressed by a combination of attributes, the computation of such bindings is out of scope for this paper. The second decision allows BiND to compute bindings combining knowledge from multiple data sets. Especially for complex narratives, it is unlikely that a single data set can act as a witness for the whole narrative. By matching each event independently, we can compute joins of individual data sets that are, as a whole, sufficient to do so. In this case, the data sets have to be join-compatible. We will discuss the problem of join compatibility later in the claim assessment step.

Algorithm 1 Event Matching

Input:
Data Set Repository $D = \{d_1, d_2, ..., d_m\}$
Narrative Graph n with set of events $\mathcal{E}_n = \{e_1, e_2, ..., e_k\}$
$Cand_n = \emptyset$
$EM = \emptyset$
for all $e \in \mathcal{E}_n$ **do**
  $PH_e = \emptyset$
  for all $d_i \in D$ **do**
    $(a_{max}, sc) \leftarrow EM(e, d_i)$
    if $sc > \theta_{EM}$ **then**
      $EM = EM \cup \{(e, a_{max})\}$
      $Cand_n = Cand_n \cup d_i$
    else
      if $CoSim(cap_{d_i}, e) > \theta_{EM}$ **then**
        $PH_e = PH_e \cup d_i$
        $Cand_n = Cand_n \cup d_i$
Return EM and all PH_e

Algorithm 2 Constraint Matching

Input:
Candidate witnesses $Cand_n = \{d_1, d_2, ..., d_m\}$
Event Matches EM
for all (e, d_p) with $e \in \mathcal{E}_n$ and $d_p \in PH_e$ **do**
  $(a_j, sc_p) \leftarrow IntConst(Const_e, A_{d_p})$
  if $sc_p > \theta_{EM}$ **then**
    $EM = EM \cup \{(e, a_j)\}$
for all $em = (e, a) \in EM$ with $a \in A_d$ **do**
  $CM_e = \emptyset$
  for all $c = (e, l, o) \in Const_e$ **do**
    $(a_{max}, sc_a) \leftarrow CM(l, A_d)$
    $sc_d \leftarrow CapSim(o, cap_d)$
    if $sc_a > sc_d$ **then**
      $CM_e = CM_e \cup (c, a_{max})$
  $rsc \leftarrow RefineScore(EM)$
  if $rsc < \theta_{CM}$ **then**
    $EM = EM \setminus \{em\}$
    $Cand_e = Cand_e \setminus \{d\}$
UpdateWitnesses()
Return EM

4.3 Constraint Matching

After the event matching, BiND aligned each event to its quantifying attributes across all data sets and identified candidate witnesses for the narrative graph. As described in chapter 2, constraints expressed by factual relations can be used in a narrative graph to specify the context of an event, e. g., a location or time or describing its participants (age, sex, nationality, etc.). We denote the set of all constraints c directly connected to an event $e \in \mathcal{E}_n$, i. e., $c = (e, l, o)$, by $Const_e$. The goal of the *constraint matching* step of BiND's pipeline is to identify those candidate witnesses that contain the necessary information to assess these constraints. This assessment is then done in the subsequent instance matching step.

The challenge is to identify how (if at all) a constraint is expressed in a data set. We have to distinguish two cases. The constraint is either (1) expressed in the data set as an attribute or (2) it is expressed by the metadata.

If a constraint is expressed as an attribute, this attribute has to be included in the same data set as a quantifying attribute of the analyzed event. Hence, for each event match $em = (e, a)$ with $a \in A_d$ the task for BiND is to identify those attributes of d that conform to constraints attached to e. The constraint matching function can accordingly be defined as follows.

- $cm_n : RF_n \rightarrow \mathcal{P}(\{A_{d_i} | \exists (e, a) \in EM \text{ with } a \in A_{d_i}\})$

If the constraint is expressed by metadata, the data set itself is defined in a way that satisfies the required constraint, e. g., for our example a data set named *Incidence of Tuberculosis in Children* satisfies the target group constraint. Hence, BiND has to check if this is the case. For the scope of this paper, we only consider the data set caption as available metadata.

Finally, BiND has to identify quantifying attributes in each preliminary match identified in the previous step. For this, it relies on the observation that some data sets as a whole can quantify a single event. In these cases, the data set's attributes usually describe contexts in which this event is analyzed, e. g., a specific year or location. This is the same type of information expressed in a narrative graph by the value of a constraint, i. e., its target node. If a narrative graph specifies a constraint of an event e, whose value matches such an attribute, then it can be used as quantifying attribute for e.

Algorithm 2 describes the constraint matching. First, BiND iterates over all pairs of events $e \in \mathcal{E}_n$ and their preliminary matches and identifies a quantifying attribute a_j using cosine similarity scores. If this is successful, (e, a_j) is added to EM. Otherwise, the preliminary match is removed from further consideration. Next, BiND takes each event match and tries to match the event's constraints to the data set. According to the two cases described above, each constraint can either be matched to an attribute or the metadata. Hence, for each combination of event match $em = (e, a)$ (with $a \in A_d$) and constraint $c = (e, l, o)$ BiND compares the cosine similarity sc_a between the label l and the best matching attribute a_{max}, with the cosine similarity sc_d between the value o and the data set caption cap_d. If $sc_d > sc_a$ then c is considered to be matched by the metadata, otherwise, the pair (c, a_{max}) is considered a *constraint match* of em, with the set of all constraint matches of em being denoted by CM_e. For each event match its similarity score is then adjusted by multiplying it with the average of the individual constraint's matching scores (sc_d or sc_a) and dividing it by two. If the adjusted score does not exceed a given threshold θ_{CM}, the event match is removed. Finally, all candidate witnesses without an attribute occurring in an event match $em \in EM$ are discarded.

By only checking if a data set contains all information to assess whether it matches the narrative's context, and not doing the assessment itself, this step can be kept purely schema-driven. This allows BiND to remove as many candidate witnesses as possible, before looking into the actual data.

4.4 Instance Matching

After the constraint matching, BiND has identified all attributes across D that are relevant for the given narrative graph, i. e., the data sets have been *vertically* partitioned.

The task of the *instance matching* is now to *horizontally* partition the candidate witnesses. This means for each event $e \in \mathcal{E}_n$ to identify those records that contain the values specified by the constraints in $Const_e$. In our running example, the target group is restricted to *children*, allowing us to remove all other records. The instance matching function is defined as:

- $im_n : \mathcal{E}_n \rightarrow \mathcal{P}(\{R_{d_i} | d_i \in Cand_n\})$

To identify suitable records satisfying a constraint, it is only necessary to look for their matched attributes. Then, the respective attribute's entry of each record has to be compared to the constraint's value. Compared to the previous two steps, an application

Algorithm 3 Instance Matching

Input:
Candidate witnesses $Cand_n = \{d_1, d_2, ..., d_m\}$
Event Matches EM
for all $em = (e, a) \in EM$ with $a \in A_d$ **do**
  $RR_{d,em} \leftarrow FindRelRec(d, CM_{em})$
  if $RR_{d,em} = \emptyset$ **then**
    $EM = EM \setminus \{em\}$
UpdateWitnesses()
Return EM

Algorithm 4 Claim Assessment

Input:
Narrative Graph n with set of events $\mathcal{E}_n = \{e_1, e_2, ..., e_k\}$
Candidate witnesses $Cand_n = \{d_1, d_2, ..., d_m\}$
Event Matches EM
$NB = \emptyset$
for all $(d_1, d_2, ..., d_k) \in Cand_n^{e_1} \times Cand_n^{e_2} \times ... \times Cand_n^{e_k}$ **do**
  for all $em = (e_i, a) \in EM$ with $a \in A_{d_i}$ and $i \in \{1, ..., k\}$ **do**
    for all $c = (e, l_c, x) \in Claim_e$ **do**
      $val_c \leftarrow FactCheck(c, d)$
    for all $nr = (e, l_{nr}, e_y) \in RN_n$ **do**
      $ps_{nr} \leftarrow StatCheck(nr, d)$
  $NB = NB \cup \{(d_1, d_2, ..., d_k, sup)\}$ with $sup \in [0, 1]$
Return NB

of SBERT would not be feasible. This is due to (1) the possibly vast number of records that would have to be compared, (2) many values (especially in larger statistical data sets) being numerical, thus rendering a sentence embedding useless, and (3) the necessity for constraints to be precisely matched. At the same time, values are usually much easier to interpret than attribute names, as these are usually categorical consisting of a fixed set of values, e. g., *female* or *male* for the attribute *sex* or contain only numerical values, e. g., for an attribute *age*.Hence, BiND applies a simple string comparison.

Algorithm 3 describes the instance matching. BiND iterates over each event match $em = (e, a) \in EM$ with $a \in A_d$. For each record $r_j \in R_d$ BiND checks every constraint $c = (x, l, y) \in Const_e$ by comparing the value y of its matched attribute a_c with the record entry r_{ja_c}. Each record matching the values of all constraints in $Const_e$ is considered a *relevant record* for em with $RR_{d,em}$ denoting the set of all relevant records for em in d. Every event match without relevant record is removed. Then, every candidate witness without any attribute occurring in an event match is removed.

4.5 Claim Assessment

Until now, BiND has partitioned the candidate witnesses, leaving for each event match $em \in EM$ only the relevant parts of the data. The task of the *claim assessment* is to assess, whether the identified candidate witnesses support or contradict the claims of the narrative graph. The claim assessment needs the ability to solve three main subtasks: (1) Validate factual claims by analyzing the relevant records of events, (2) compute statistical dependencies between the relevant records for events connected by a narrative relation, and (3) if their quantifying attributes occur in different tables, compute the required joins.

To assess a factual claim, BiND has to know how it is expressed in the data. Thus, it relies on an alphabet $\Sigma_C \subseteq \Sigma_F$ of relation labels with clearly defined semantics often occurring in factual claims. Each label is assigned to its own validation algorithm, ranging from validating specific values to computing aggregations (e. g., count, max, min). According to definition 2.1, factual claims never connect two events, thus each claim can be validated using a single data set.

For the second task, BiND is again provided with an alphabet $\Sigma_{CN} \subseteq \Sigma_N$ of narrative relation labels with clearly defined semantics that are coupled to specific statistical metrics, i. e., algorithms for their assessment. However, there are two main differences: (1) Narrative relations always connect two events, thus requiring a join of tables, and (2) they are highly context-dependent and can not generally be considered valid or invalid.

The challenge for BiND when computing joins is that it has no access to schema information for the data sets in D and as such no knowledge about possible join candidates. Hence, we apply a simple join heuristic. Assume $nr = (e_i, e_j) \in RN_n$, as well as d_i and d_j contain a quantifying attribute for e_i and e_j, respectively. A join between d_i and d_j is then computed in two steps. First, all attribute pairs $(a_i \in A_{d_i}, a_j \in A_{d_j})$ with identical attribute names are identified as potential *join candidates*. This is to avoid comparing contextually incompatible data, e. g., the prevalence of tobacco use in Germany during the year 2005 with the prevalence of high blood pressure during the year 2010 in France. From these candidates, all pairs containing a quantifying attribute of e_i or e_j or an attribute matched to one of their constraints are removed. This is done to allow analyzing narrative relations imposing the same type of constraints on an event, e. g., restricting it to a specific year, while the constraint's value differs between both join partners. An example is a narrative relation stating the prevalence of high blood pressure increased in 2020 compared to 2010.

Algorithm 4 describes the claim assessment. Let $Cand_n^e$ denote the set of candidate witnesses containing a quantifying attribute of e. For each narrative binding, i. e., combination of candidate witnesses, $(d_1, ..., d_k)$ with $d_i \in Cand_n^{e_i}$ BiND checks whether each claim and narrative relation of n is supported. For this, all relevant event matches $em = (e_i, a)$ are analyzed. BiND then validates each claim $c = (e, l_c, x)$ using the specific algorithm assigned to the relation label l_c. Analogously, BiND checks each narrative relation $nr = (e, l_{nr}, e_y)$ using the algorithm assigned to label l_{nr}. Finally, all narrative bindings are presented to the user. Thereby, each binding is assigned a binary value (0 if any claim or narrative relation is considered to contradict the data and 1 otherwise).

It is important to note that using predefined relation alphabets restricts the possible narratives that BiND can interpret. This is a necessary requirement as interpreting the semantics of arbitrary relation labels and translating them into applicable algorithms is a difficult task for which no solution currently exists. However, a small set of relation types is usually sufficient to allow for a binding against most narratives as most claims can be assigned to a few broad categories. Many state-of-the-art fact-checkers such as [3] are similarly restricted on a limited set of fact types. Furthermore, BiND's claim assessment is designed to be extendable, with the possibility to integrate new relation types using simple algorithms.

4.6 Discussion

For the scope of this paper we focused on the core functionalities of the BiND system. However, there are many interesting open problems that we aim to tackle in the future.

One challenge are constraints integrated into an event. As an example, consider the event *suffering high blood pressure* which specifically asks for *high* blood pressure values. Thus, we have to distinguish between people with normal or low blood pressure and those with high blood pressure. Now two situations could occur: (1) A data set includes an attribute incorporating the constraint, e. g., *Prevalence of high blood pressure in men*, or (2) there is an attribute quantifying the event, but not explicitly satisfying the asked for constraint, e. g., an attribute *Measured blood pressure value*. While BiND supports the first case, the second case requires external domain knowledge, which BiND has no access to. For example, the qualifier *high* can translate to values higher than 140mm Hg. However, the semantics are highly context-dependent. *High* blood pressure values can, for example, not be handled the same way as *high* physical activity. Solving this problem is even harder if these qualifiers have no clear definition.

A subtask in BiND's claim assessment step is to join individual data sets that include quantifying attributes for different events in the same narrative. While the heuristic applied by BiND allows to sufficiently compute bindings for simple narratives, complex narratives might necessitate more sophisticated join algorithms. Consider the following narrative: *The probability for cardiovascular problems is higher for children of people that themselves suffered cardiovascular problems*. Such narratives are currently not supported by BiND as they require an understanding of complex semantic relationships (such as a parent-child relationship) in data sets where usually no schema information is available.

5 EVALUATION

In the following, we present the results of two different experiments to show the applicability of BiND (the code, including all data sets and narratives can be found at our github[2]). In the first experiment, we evaluate the system in a real-world environment by using the Global Health Observatory of the WHO and narratives extracted from the organization's data stories. As no tools exist for finding narrative bindings on data sets, finding a suitable baseline is challenging. While fact-checkers can effectively validate factual claims, they lack the exploratory aspect needed to compute narrative bindings. Still, to show that BiND's claim assessment can compare to state-of-the-art fact-checking tools we select four tools and compare them to BiND in a second experiment. All experiments are executed on a server running Ubuntu 18.04.1 with an IntelCore i9 7940X, 3.1 GHz, having 28 cores and 128GB RAM.

5.1 Evaluation on the Global Health Repository

Data Repository. The first evaluation is run on the *Global Health Observatory (GHO)* [40], a large-scale open data repository that is maintained by the WHO. The data is partitioned into multiple sub-repositories, such as the *Global Health Estimates*, or the *HEAT Plus Data Repository*. Due to this decentralized data access resulting in high manual expenditures to collect data sets, we restrict this evaluation to those data sets available through the GHO API and those included in the HEAT Plus Data Repository, resulting in a set of 2,807 individual tables. It has to be noted that the *HEAT Plus Data Repository* combines multiple indicators in each table. To match the remaining WHO data sets we split these tables, resulting in one separate table per indicator. We also annotate the raw data with metadata used in the table's visualization on the WHO's website.

Table 1: Average runtime for the individual steps of BiND's matching pipeline - keep in mind that the similarity scores only need to be computed once, and can be reused

	Full Evaluation (in s)		
Step	GHO	TabFact	$C19_s$
Computing Similarity Scores	106.28	846.42	5.86
Loading Data Sets and Narratives	34.59	7.23	0.29
Event Matching	2.13	1.76	0.003
Constraint Matching	0.003	1.73	0.008
Instance Matching	18.7	2.48	0.026
Claim Assessment	150.97	0.01	0.0007
Total (wo/Similarity Computation)	206.39	13.21	0.33

Narratives and Ground Truth. A major difficulty in evaluating narrative bindings results from the lack of suitable benchmarks. To create a feasible ground truth, we rely on the *Data Stories*, articles published on the website of the WHO[3]. They present insights gathered from data sets available through the GHO. As these data sets are explicitly referenced, the Data Stories provide proxies for correct narrative bindings. From these articles, we manually extract 42 narratives, identify the referenced data sets, and within them those attributes quantifying each event. We denote this ground truth, consisting of 93 correct narrative-witness pairs, as WHO_{Strict}.

Now the problem arises, that these data stories do not reference every data set that could be considered for a plausible narrative binding. Hence, we manually analyzed each of the 2,807 tables and created a second, relaxed ground truth by extending WHO_{Strict}. We denote this ground truth as WHO_{Rel}. For each narrative-witness pair (n, d) in WHO_{Strict}, we also considered each data set d' as a plausible witness for n that (1) contains the same attribute as d in a different context but still adheres to the context of n or (2) describes a more specific topic than d. As such WHO_{Rel} is a more accurate representation of the narrative bindings we expect BiND to identify. Compared to WHO_{Strict}, the number of plausible narrative-witness pairs increases significantly from 93 to 393.

Evaluation Outline. To evaluate BiND's performance, we focus on its two main tasks, i. e., (1) identifying suitable candidate witnesses containing data relevant to the narrative and (2) assessing the plausibility of the narrative's claims. For the first task, we look at the candidate witnesses remaining after each of the first three steps of BiND's pipeline and compare them with WHO_{Strict} and WHO_{Rel}. We analyze precision, recall, and the reduction factor of discarded data sets. For every downstream step, we apply each preceding step using those thresholds where the highest average performance over both, WHO_{Strict} and WHO_{Rel} was achieved. For the second task, we evaluate BiND's accuracy in correctly assessing the claims as being supported by the data. In table 1 we measure the runtime each step needs to process all 42 narratives over the whole repository (without parallelization).

[2] https://github.com/DenisNagel93/BiND

[3] https://www.who.int/data/stories

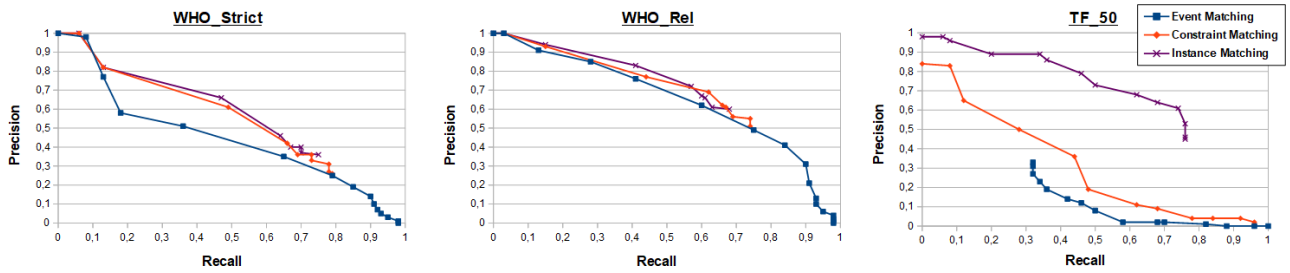

Figure 5: Precision-Recall Curves for the Candidate Witnesses after each step of the pipeline

Table 2: Highest Recall at which a Reduction Factor (RF) of at least 99% and 95% are achieved

	max. Recall at an RF of 99% / 95%	
Step	GHO (WHO_{Rel})	TabFact (TF_{50})
Event Matching	84.0 / 93.0	- / 50.0
Constraint Matching	75.0 / 75.0	44.0 / 68.0
Instance Matching	68.0 / 68.0	76.0 / 76.0

Event Matching. For the event matching, we compute the average precision, recall, and reduction factor values over all 42 narratives for different thresholds θ_{EM}, ranging from 0.0 to 1.0 (in steps of 0.05). We do this for WHO_{Strict} and WHO_{Rel} separately. The precision-recall-curves are shown in the first diagram of figure 5. The maximum recall BiND can retain while discarding at least 99% (95%) of the data sets, i. e., a reduction factor of 0.99 (0.95), can be seen in table 2.

BiND can reliably prune the repository by simply matching attribute names and metadata to the events of a narrative. Our results show that BiND's event matching performs best across the threshold range from 0.6 (Prec.: 41%, Rec.: 84%, WHO_{Rel}) to 0.75 (Prec.: 76%, Rec.: 41%, WHO_{Rel}). Reaching a high recall is essential as all further tasks only remove candidate witnesses thus potentially lowering the recall.

The low precision values when comparing WHO_{Strict} to WHO_{Rel} can be explained by the fact that the GHO contains multiple data sets for many indicators analyzed by the WHO. Hence, it offers many suitable witness candidates for most events. However, the data stories only reference a few data sets for each indicator, thus WHO_{Strict} considers many plausible candidates as false positives, resulting in a lower precision.

Regarding the reduction factor, BiND discards around 99% of all data sets (averaged over all narratives), while keeping a high recall of up to 84%. Hence, in the event matching BiND can already reliably distinguish relevant from irrelevant data indicating the high importance of events for their narrative.

Constraint Matching. We first apply the event matching with $\theta_{EM} = 0.65$. We then apply the constraint matching with different thresholds θ_{CM}, ranging from 0.0 to 1.0 (in steps of 0.05). The results are shown in figure 5. After the constraint matching step, BiND removed those candidate witnesses that do not contain the information necessary to assess the narrative's constraints. In doing so, the overall precision clearly improves for most recall values. The constraint matching performs best in the θ_{CM}-range between 0.6 (Prec.: 55%, Rec.: 74%, WHO_{Rel}) and 0.8 (Prec.: 69%, Rec.: 62%, WHO_{Rel}). When considering the results achieved for WHO_{Rel} after the event matching with $\theta_{EM} = 0.65$ (Prec.: 49%, Rec.: 75%), the constraint matching achieves an increase in precision of 6 percentage points while only sacrificing 1 percentage point in recall. A similar observation holds for WHO_{Strict}. Note, that the recall can not exceed the results of the previous step as candidate witnesses are only removed. This also explains the lower recall values in table 2.

Instance Matching. For the instance matching, we choose $\theta_{EM} = 0.65$ and compute precision and recall for varying θ_{CM}. After the instance matching step, BiND removed all witness candidates not matching the narrative's context. Note that the GHO's data sets only contain a low variety of contextual information (usually restricted to *year*, *country*, *sex*) with mostly the same values (each table contains data for most of the WHO's member nations and spans similar year ranges). Hence, the impact of the instance matching step is relatively minor. It achieves a slight increase in precision, while trading a few percent of recall. At the best performing threshold of $\theta_{CM} = 0.6$, for WHO_{Rel} the precision increases by 5 points reaching 60%, while the recall drops by 6 points reaching 68%. A similar observation holds for WHO_{Strict}. In this step, reaching a high precision has a higher priority than a good recall, as all remaining candidate witnesses are considered to be successfully bound to the narrative. Hence, any false positive negatively impacts the claim assessment step.

Claim Assessment. To evaluate the claim assessment, we first run the previous steps using $\theta_{EM} = 0.65$ and $\theta_{CM} = 0.6$. In the claim assessment BiND decides for the bound data sets whether they support or contradict the narrative. As ground truth, we assume every claim stated in one of WHO's data stories to be supported by the referenced data sets. We see that BiND can effectively analyze the plausibility of narrative claims reaching an accuracy of **71.15%**.

Overall Results. The evaluation shows that BiND can effectively be applied in a real-world environment to compute narrative bindings between scientific narratives and structured open data. When looking at table 1 this task can be solved in an acceptable time taking in the worst case only around 5 minutes to evaluate 42 narratives over 2,807 data sets. On average, this means BiND takes around 5 seconds per narrative.

Table 3: Verification accuracy of BiND's claim assessment compared to various state-of-the-art fact checkers (*reported in [20],reported in [7] for the full TabFact benchmark)**

Tool	TF_{50}	$C19_s$	$C19_{sd}$
TabFact	57.77	76.0*	/
TAPAS	81.0**	64.0*	/
AggChecker	/	50.0	50.0
Scrutinizer	/	80.0	85.0
$BiND_{wE}$	72.34	60.0	80.0
BiND	50.72	55.0	76.65

5.2 Comparison with existing Fact Checkers

Data Repository. For the second evaluation, we consider two data set repositories. The first is the *TabFact* benchmark[3] (TF), which was designed with the purpose of evaluating language inference on structured data and is used by many fact-checkers for evaluation. It consists of 16,570 web tables that were extracted from Wikipedia. They offer a stark contrast to the GHO data, which is highly curated and consists mainly of numerical data, while the Wikipedia tables are usually entity-centric and their quality highly heterogeneous. We therefore also measure the precision, recall, and reduction factor after each of BiND's steps and present them alongside the previous results (see figure 5). The second repository ($C19_s$) is a small Covid-19 data set used to evaluate the fact-checker *Scrutinizer*[20]. It is a curated data set of 8 individual tables consisting mainly of numerical data.

Narratives and Ground Truth. For the TabFact benchmark, the authors collected 117,854 statements referring to the data sets, which they manually annotated as *entailed* or *refuted*. Each table is assigned a varying number refuted and entailed statements. From these, we select 47 statements referring to the first 25 different tables. We always select the first entailed and the last refuted statement that is supported by the relation alphabets (Σ_C and Σ_{CN}) of BiND. We denote this subset of statements by TF_{50}. We then manually translated them into narrative graphs. As ground truth, we only consider the data set a statement is assigned to as correct narrative binding. For the $C19_s$ data set[4], the authors collected 20 claims, that are also annotated as entailed or refuted. For our evaluation, we translated all 20 claims into narrative graphs.

Evaluation Outline. We evaluate BiND's fact-checking capabilities by comparing it to four state-of-the-art fact-checking tools, i. e., Aggchecker[19], Scrutinizer [20], TabFact[3], and Tapas[14]. However, these baselines have a few limitations. Except for Scrutinizer, they all require the table a claim is validated on as input. To allow for a fair comparison, we provide the same information to BiND, thus not applying its exploratory aspect. We denote this version of BiND as $BiND_{wE}$. Scrutinizer is also the only baseline able to analyze claims requiring multiple tables to be validated. Finally, AggChecker can only interpret numerical data, and as such can not be applied to most statements of TF_{50}.

[4]https://zenodo.org/record/5128604/#.YskYIoTP2Ul

Evaluation Results. The evaluation shows that BiND can reach a comparable performance to state-of-the-art fact-checkers on dedicated fact-checking benchmarks, while none of the baseline tools were applicable on the WHO data set. On the TF_{50} benchmark, $BiND_{wE}$ clearly outperforms Tabfact (72.34% vs. 57.77%), while AggChecker is not applicable on the provided statements. When no information about the required table is given, BiND's accuracy drops to 50.72%, which indicates the higher complexity of BiND's task compared to simple fact-checking. However, TAPAS still outperforms $BiND_{wE}$. This is not surprising as TAPAS is listed as the best-performing tool on the original TabFact benchmark and is specifically designed for validating factual claims, which is not the case for BiND.

On the $C19_s$ data set BiND only achieves a relatively low accuracy of 55% (60% for $BiND_{wE}$) (however still outperforming AggChecker and only being slightly outperformed by TAPAS). These values can be explained by the fact that, while all attributes in the table of $C19_s$ are labeled with dates, many of the claims do not specify a date they refer to. On a closer look, AggChecker and Scrutinizer also validate the respective claims on the wrong parts of the data. Scrutinizer only achieves an accuracy of 80% as it accordingly labels these claims as refuted, coincidentally matching their annotation. Therefore, we also evaluate all three approaches on a modified version of the claims where the missing dates are explicitly added ($C19_{sd}$). We can see that the accuracy of BiND drastically improves, reaching up to 76.65% (vs. 85% for Scrutinizer).

6 CONCLUSION

In this paper, we introduce BiND, which, to the best of our knowledge, represents the first system to automatically compute narrative bindings between scientific narratives and structured data sets. These narrative bindings are valuable in providing evidence for narratives, whose plausibility is questioned. Our evaluation of the Global Health Observatory shows that BiND can be successfully deployed to great effect in a real-world environment. Compared to state-of-the-art fact-checkers, BiND achieves comparable accuracy in validating individual claims. On the other hand, most fact-checkers lack the exploratory capabilities of BiND in automatically identifying data sets containing relevant data to answer these claims. Additionally, BiND sets itself apart by supporting on-the-fly computation of joins between individual data sets, that only in conjunction support a given narrative.

ACKNOWLEDGMENTS

Supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): PubPharm – the Specialized Information Service for Pharmacy (Gepris 267140244).

REFERENCES

[1] Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A Neural Transition-based Model for Argumentation Mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Association for Computational Linguistics, Online, 6354–6364. https://doi.org/10.18653/v1/2021.acl-long.497

[2] Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. 2015. TabEL: Entity Linking in Web Tables. In *The Semantic Web - ISWC 2015.* Springer International Publishing, Cham, 425–441.

[3] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *ArXiv* abs/1909.02164 (2019).

[4] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. 2012. Finding Related Tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (Scottsdale, Arizona, USA) *(SIGMOD '12)*. Association for Computing Machinery, New York, NY, USA, 817–828. https://doi.org/10.1145/2213836.2213962

[5] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2020. TURL: Table Understanding through Representation Learning. *Proc. VLDB Endow.* 14, 3 (nov 2020), 307–319. https://doi.org/10.14778/3430915.3430921

[6] Y. Dong, K. Takeoka, C. Xiao, and M. Oyamada. 2021. Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 456–467. https://doi.org/10.1109/ICDE51399.2021.00046

[7] Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 281–296. https://doi.org/10.18653/v1/2020.findings-emnlp.27

[8] Faten El Outa, Matteo Francia, Patrick Marcel, Veronika Peralta, and Panos Vassiliadis. 2020. Towards a Conceptual Model for Data Narratives. In *International Conference on Conceptual Modeling (ER)*. Springer, 261–270. https://doi.org/10.1007/978-3-030-62522-1_19

[9] Mahdi Esmailoghli, Jorge-Arnulfo Quiané-Ruiz, and Ziawasch Abedjan. 2022. MATE: Multi-Attribute Table Extraction. *Proc. VLDB Endow.* 15, 8 (jun 2022), 1684–1696. https://doi.org/10.14778/3529337.3529353

[10] Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A Weakly-supervised Approach to Argumentative Zoning of Scientific Documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., 273–283. https://aclanthology.org/D11-1025

[11] Ivan Habernal and Iryna Gurevych. 2015. Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2127–2137. https://doi.org/10.18653/v1/D15-1255

[12] Naeemul Hassan, Anil Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10 (08 2017), 1945–1948. https://doi.org/10.14778/3137765.3137815

[13] David Heckerman and John S Breese. 1996. Causal independence for probability assessment and inference using Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 26, 6 (1996), 826–831. https://doi.org/10.1109/3468.541341

[14] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4320–4333. https://doi.org/10.18653/v1/2020.acl-main.398

[15] Daniel Hienert, Dagmar Kern, Katarina Boland, Benjamin Zapilko, and Peter Mutschke. 2019. A Digital Library for Research Data and Related Information in the Social Sciences. In *19th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 148–157. https://doi.org/10.1109/JCDL.2019.00030

[16] Vinh Thinh Ho, Koninika Pal, Simon Razniewski, Klaus Berberich, and Gerhard Weikum. 2021. Extracting Contextualized Quantity Facts from Web Tables. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 4033–4042. https://doi.org/10.1145/3442381.3450072

[17] Vinh Thinh Ho, Koninika Pal, and Gerhard Weikum. 2021. *QuTE: Answering Quantity Queries from Web Tables*. Association for Computing Machinery, New York, NY, USA, 2740–2744. https://doi.org/10.1145/3448016.3452763

[18] Muhammad Nihal Hussain, Hayder Al Rubaye, Kiran Kumar Bandeli, and Nitin Agarwal. 2021. Stories from Blogs: Computational Extraction and Visualization of Narratives. In *Proceedings of Text2Story - Fourth Workshop on Narrative Extraction From Texts*. CEUR-WS.org, 33–40.

[19] Saehan Jo, Immanuel Trummer, Weicheng Yu, Xuezhi Wang, Cong Yu, Daniel Liu, and Niyati Mehta. 2019. Verifying Text Summaries of Relational Data Sets. In *Proceedings of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands) *(SIGMOD '19)*. Association for Computing Machinery, New York, NY, USA, 299–316. https://doi.org/10.1145/3299869.3300074

[20] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: fact checking statistical claims. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2965–2968. https://doi.org/10.14778/3415478.3415520

[21] George Katsogiannis-Meimarakis and Georgia Koutrika. 2021. A Deep Dive into Deep Learning Approaches for Text-to-SQL Systems. In *Proceedings of the 2021 International Conference on Management of Data*. Association for Computing Machinery, New York, NY, USA, 2846–2851. https://doi.org/10.1145/3448016.3457543

[22] Hermann Kroll, Denis Nagel, and Wolf-Tilo Balke. 2020. Modeling Narrative Structures in Logical Overlays on top of Knowledge Repositories. In *International Conference on Conceptual Modeling (ER)*. Springer, 250–260. https://doi.org/10.1007/978-3-030-62522-1_18

[23] Hermann Kroll, Jan Pirklbauer, Jan-Christoph Kalo, Morris Kunz, Johannes Ruthmann, and Wolf-Tilo Balke. 2021. Narrative Query Graphs for Entity-Interaction-Aware Document Retrieval. In *The 23rd International Conference on Asia-Pacific Digital Libraries (ICADL)*. Springer, Springer, Online, 80–95. https://doi.org/10.1007/978-3-030-91669-5_7

[24] Hermann Kroll, Florian Plötzky, Jan Pirklbauer, and Wolf-Tilo Balke. 2022. What a Publication Tells You – Benefits of Narrative Information Access in Digital Libraries. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Cologne, Germany, 1–8. https://doi.org/10.1145/3529372.3530928

[25] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, and Bingyu Sun. 2013. Mining causal association rules. In *13th International Conference on Data Mining (Workshops)*. IEEE, 114–123. https://doi.org/10.1109/ICDMW.2013.88

[26] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma. 2015. From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 2 (2015), 1–27. https://doi.org/10.1145/2746410

[27] Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. 2010. Annotating and Searching Web Tables Using Entities, Types and Relationships. *Proc. VLDB Endow.* 3, 1–2 (sep 2010), 1338–1347. https://doi.org/10.14778/1920841.1921005

[28] Sucheta Nadkarni and Prakash P Shenoy. 2001. A Bayesian network approach to making inferences in causal maps. *European Journal of Operational Research* 128, 3 (2001), 479–498. https://doi.org/10.1016/S0377-2217(99)00368-9

[29] Denis Nagel, Till Affeldt, and Wolf-Tilo Balke. 2021. Data Narrations - Using flexible Data Bindings to support the Reproducibility of Claims in Digital Library Objects. In *1st International Workshop on Digital Infrastructures for Scholarly Content Objects (DISCO@JCDL2021)*. CEUR Proceedings, Vol-2976, CEUR Proceedings, Vol-2976, Urbana-Champaign, IL, USA, 19–23.

[30] Julianna Pakstis, Hannah Calkins, Christiana Dobrzynski, Spencer Lamm, and Laura McNamara. 2019. Advancing Reproducibility Through Shared Data: Bridging Archival and Library Practice. In *19th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 49–52. https://doi.org/10.1109/JCDL.2019.00017

[31] Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. 98–107. https://doi.org/10.1145/1568234.1568246

[32] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR* abs/1908.10084 (2019). arXiv:1908.10084 http://arxiv.org/abs/1908.10084

[33] Dominique Ritze, Oliver Lehmberg, and Christian Bizer. 2015. Matching HTML Tables to DBpedia. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics* (Larnaca, Cyprus) *(WIMS '15)*. Association for Computing Machinery, New York, NY, USA, Article 10, 6 pages. https://doi.org/10.1145/2797115.2797118

[34] María Teresa Rodríguez, Sérgio Nunes, and Tiago Devezas. 2015. Telling stories with data visualization. In *Proceedings of the 2015 Workshop on Narrative & Hypertext*. 7–11. https://doi.org/10.1145/2804565.2804567

[35] Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

[36] Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. 2000. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery* 4, 2-3 (2000), 163–192. https://doi.org/10.1023/A:1009891813863

[37] Zareen Syed, Tim Finin, Varish Mulwad, Anupam Joshi, et al. 2010. Exploiting a web of semantic data for interpreting tables. In *Proceedings of the Second Web Science Conference*.

[38] Petros Venetis, Alon Halevy, Jayant Madhavan, Marius Paşca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. 2011. Recovering Semantics of Tables on the Web. *Proc. VLDB Endow.* 4, 9 (jun 2011), 528–538. https://doi.org/10.14778/2002938.2002939

[39] Jun Wang and Klaus Mueller. 2017. Visual causality analysis made practical. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 151–161. https://doi.org/10.1109/VAST.2017.8585647

[40] World Health Organization. 2023. Global Health Observatory. https://www.who.int/data/gho

[41] World Health Organization. 2023. Institutional Repository for Information Sharing. https://apps.who.int/iris

[42] Cindy Xiong, Joel Shapiro, Jessica Hullman, and Steven Franconeri. 2019. Illusion of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 853–862. https://doi.org/10.1109/TVCG.2019.2934399