Data Narrations - Using flexible Data Bindings to support the Reproducibility of Claims in Digital Library Objects

Denis Nagel[1], Till Affeldt[1] and Wolf-Tilo Balke[1]

[1]Institute for Information Systems, TU Braunschweig, Braunschweig, Germany

Abstract

Digital libraries support researchers by providing public access to a vast collection of state-of-the-art literature. The considerable variety of statements, claims, observations and insights that form the narrations of these documents can be used as a valuable groundwork for further research. However, when confronted with these narrations, concerns regarding their reproducibility might arise. Tackling these concerns usually requires a careful analysis of the underlying data sets and a search for similar repositories that support the questioned claims. In short, it is necessary to find repositories whose data narrations match those of the publication. Unfortunately, data analysis and mining are far too often reduced to basic statistical analyses that usually fail to be helpful. In this paper, we propose a novel idea to use structured narratives as a template to discover supporting data narrations, hence reducing the problem of assessing the reproducibility of a publication to a simple matching task between a document and data set. To realize this idea, we outline a novel two-step matching strategy by describing the individual steps along the lines of a pharmaceutical use case. We thereby identify the main open research tasks and discuss problems that need to be solved to develop a full-fledged matching algorithm.

Keywords

Narrative Intelligence, Open Data, Digital Libraries

1. Introduction

Digital libraries and their vast collections of documents represent an invaluable source of knowledge. They provide public access to state-of-the-art research across many domains of science. The scientific narrations provided through these documents form the groundwork for ongoing research that builds upon their claims, insights and observations. However, when working with such documents, concerns about the reproducibility of the encountered narrations might arise [1]. Often these narrations originate from research data that has been collected throughout extensive experiments, evaluations, scientific studies, or surveys. In recent years increasing efforts have been undertaken to make the ever-increasing amounts of research data publicly available by integrating them into the existing libraries and, in the best case linking them to their associated documents [2, 3].

Consider a researcher reading a document that has high relevance for her current work, but she is sceptical about a specific claim made by the authors. Is the claim plausible? Moreover, is the data that supports it broadly representative for her research domain, or is it applicable only to the document's specific use case? Answering these questions requires retracing the steps taken by the original authors by thoroughly analyzing the underlying data sets. Now, essentially two situations can occur. On the one hand, the required data set might not have been published or is challenging to find due to missing references by the authors, i.e., the critical link between document and data set might be unavailable [3]. On the other hand, if the data is readily available, it might still be unrepresentative for the domain of interest. Even if it is representative, a thorough data set analysis can result in a very time-consuming and exhausting process that many researchers might not be willing to take.

Recently scientific narratives have sparked much interest in the scientific discourse, and their application in digital libraries is the topic of an ongoing discussion [4]. Every document can contain several narrations that connect insights and statements to form a coherent story presented to the reader. For an example, consider the following pharmaceutical narrative: *Cardiovascular diseases are the leading reason for premature deaths worldwide and are caused by a multitude of risk factors, many of which are avoidable. To avoid premature deaths, it is thus essential to raise awareness of these risk factors.* One instance of this narrative can be encountered in [5], where multiple risk factors, like elevated blood pressure, are listed that are claimed to be causal for the occurrence of cardiovascular diseases. Hence, when reading [5], the question might arise whether these claims are plausible.

We believe that, by translating narrations such as our example into *structured narratives* [6], it is possible to assess the reproducibility of their statements, claims, insights and observations without a costly manual data analysis. Our proposal is based on the assumption that

DISCO'21 - Digital Infrastructures for Scholarly Content Objects at JCDL2021, September 30–October 01, 2021, Online

nagel@ifis.cs.tu-bs.de (D. Nagel); t.affeldt@tu-bs.de (T. Affeldt); balke@ifis.cs.tu-bs.de (W. Balke)

0000-0002-5832-9154 (D. Nagel); 0000-0001-6440-5654 (T. Affeldt); 0000-0002-5443-1215 (W. Balke)

Age (years)	Resting Blood Pressure (mm HG)	Serum Cholesterol (mg/dL)	Diagnosis of Heart Disease
63	145	233	0
67	160	286	2
67	120	229	1
37	130	250	0
(...)	(...)	(...)	(...)

Figure 1: Excerpt from a pharmaceutical data set [7], containing clinical data of 303 patients, collected by the Cleveland Clinic Foundation (publicly available through the UCI machine learning repository [8])

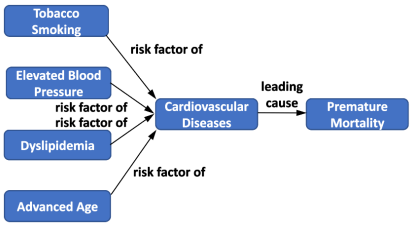

Figure 2: Excerpt of a structured narrative extracted from a PubMed publication [5]. The nodes of the narrative graph represent important entities, events and literals, while the labeled edges represent the relations between them

the intrinsic relations between individual values inside data sets also form implicitly expressed narrations that we call *data narrations*. Our core idea is that *whenever a data set is schematically suitable to reproduce some scientific narrative, it should provide a data narration that can be successfully aligned to that narrative.*

With structured narratives at hand, the problem of assessing reproducibility can thus be reduced to a simple matching approach. Given a structured narrative extracted from a document of interest, in a first step, data sets suitable for the matching can be discovered by considering the available meta-data, i.e., data set descriptions, table headers, or column titles. We then align this meta-data to the events and entities described in the extracted narrative, resulting in a set of possible candidate data sets. The second step looks into the actual data to verify whether the relations between the entities and events of the narrative also occur expressed by the data.

The contributions of this paper can be summarized as follows: We propose structured narratives as a tool to assess the reproducibility of a document's statements, claims and insights. For this, we outline a novel approach to discover data sets fitting to the document's narrations based on a simple two-step matching strategy.

2. Preliminaries

Data Sets Data sets usually store empirical data gained through experiments, measurements, studies, or surveys. A problem often encountered when working with data sets is very high heterogeneity in structure and schema formatting. For ease of understanding, we thus consider data sets in the scope of this paper to store (mainly numerical) information in a tabular format. We denote the set of all possible data sets that comply with this format by DS. Each data set $D_i = \{Var, T\} \in DS$ consists of a set of variables Var, with each $var_j \in Var$ representing a single column and a set of tuples T, each representing an individual record ($t_i \in T$) of the data set which comprises of a value $d_{ij} \in T$ for each variable $var_j \in Var$. Figure 1 shows an example for a pharmaceutical data set.

Structured Narratives For this paper, we define structured narratives according to the definition found in [6]. Therein narratives are defined as directed edge- and node-labeled graphs $N = (V, R)$. The nodes represent events with a temporal component, entities (i.e., real-world objects and concepts) and literals, such as numerical values and lexicographical strings. In contrast the edges represent the relations between them, which could be the participation of some entity in an event, a causal or temporal relationship between events or simple facts. As such, the set of nodes is defined as $V \subseteq E \cup L \cup \Gamma$ with E being the set of entities, L being the set of literals, and Γ being the set of events. Then the set of edges is defined as $R \subseteq (E \cup \Gamma) \times \Sigma \times (E \cup \Gamma \cup L)$, with Σ being the alphabet of available edge labels. Figure 2 shows an excerpt of a structured narrative extracted from [5].

Narrative Bindings and Data Narrations Narratives can consist of any arbitrary statements and claims without any indication of their plausibility. Hence, in [6] narrative bindings are introduced to connect parts of the narrative to a knowledge repository of any type in the sense of substantiation. Semantically a narrative binding between a narrative and a repository indicates that the repository supports the statements made throughout the narrative. As such, we can define *narrative bindings* as follows. Let $N = (V, R)$ be a narrative and KR the set of all knowledge repositories. A narrative binding is a tuple $nb = (r, kr) \in R \times KR$. We say that e is bound to kr by nb.

Narratives can be encountered not only in natural languages, such as documents, novels, or human speech, but also behind the intrinsic relations that the individual values of a data set implicitly express. We call this special type of narrative a data narration. Using the notion of narrative bindings, we can define a data narration as follows. Let $DS \subset KR$ be the set of all data sets. A narrative $N = (V, R)$ is called a data narration of a data set $D \in DS$, iff there exists a $nb = (r_i, D)$, for every $r \in R$.

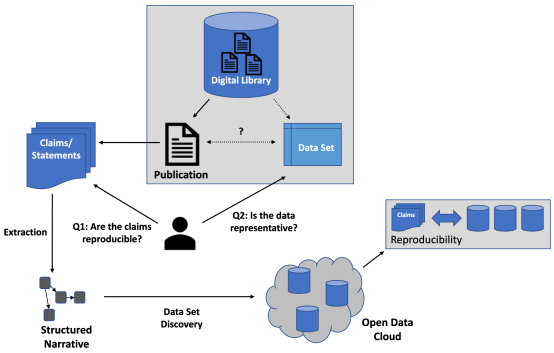

Figure 3: Our proposed outline to assess the reproducibility of claims using structured narratives

3. Related Work

Reproducibility of scientific results has always been one of the core aspects of good scientific practice. Unfortunately, there are many cases in which it is not or only insufficiently taken into account. As such, terms such as reproducibility crisis can be encountered frequently in the literature. Hence, many authors propose new strategies to improve this situation. Recent examples are [9], where a new system to collect provenance information for data science pipelines is presented, and [10] where the authors developed an integrative platform in the context of the semantic web to capture the provenance information for individual experiments.

In order to tackle the problem of reproducibility, an increasing amount of research data has recently been collected and integrated into digital libraries [3, 11, 12]. At the same time, it is often difficult to find the links connecting a publication to the underlying data. In [3] the authors thus introduced a specialized digital library that offers integrated access to both the documents and their associated research data. Contrary to such strategies, our approach aims at finding arbitrary research data suitable to reproduce the claims from some document, even if it was not the source for these statements.

4. Method Description

In this section, we present our envisioned idea of using structured narratives to assess the reproducibility of scientific claims and statements in more detail (for a visualization, see figure 3). We identify the steps required to develop a full-fledged matching algorithm along the lines of a pharmaceutical use case and discuss difficulties and open tasks that need to be solved.

Problem Description Let us reconsider the PubMed publication [5], in which the narrative described in section 1 can be encountered. For simplicity, we focus on the following excerpt of the narrative, which is visualized in 2: The paper claims that cardiovascular diseases (CVDs) are the leading cause of premature mortality worldwide. It lists many risk factors, like tobacco smoking, elevated blood pressure, dyslipidemia and advanced age, among others, that are stated to be associated with CVDs. Our goal is now to assess whether these claims can be reproduced using available open research data sets. We assume that only data sets whose intrinsic data narrations support these claims can be used to reproduce them. The basic idea for our approach is that by translating both, the intrinsic messages of the data set and the scientific narration into structured narratives, we can reduce the problem of assessing the reproducibility of the claims to a simple matching problem.

Extraction of Structured Narratives In order to allow for an automated matching between publications and research data, we need to extract the narrations encountered throughout the document and translate them into a structured narrative. Hence, it is necessary to analyse how each building block of a narrative graph is expressed in natural language. For this, we have to combine multiple disciplines from natural language processing (NLP), such as named entity recognition and event detection for the nodes and relation extraction for the edges between them. By applying a manual extraction on our example, we can identify six biomedical concepts and entities, namely *Cardiovascular Diseases*, *Premature Mortality*, as well as the four different risk factors (*Tobacco Smoking*, *Elevated Blood Pressure*, *Dyslipidemia* and *Advanced Age*). Furthermore, the narration claims that there is a causal dependency between CVDs and premature mortality, as well as a relation between CVDs and each of the risk factors. Hence, for our example a structured narrative as defined in section 2 and denoted by $N_e = (V_e, R_e)$ could look as shown in figure 2. As a manual extraction can often result in a cumbersome process, developing a sophisticated strategy for the automated extraction of narratives is crucial for the practical applicability of our approach and thus an important task for future work. Different approaches, such as [13], are already actively discussed, thus providing valuable groundwork.

Identification of Data Narrations Discovering data sets that are suitable to match N_e requires awareness of their intrinsic data narrations. The current state-of-the-art approach to make data narrations visible is the application of techniques for data visualization [14, 15]. However, such techniques can only be applied on top of data analysis, i.e., if the intrinsic relations of a data set are already known, which is a requirement that is rarely met in the realm of open data. Especially in large-scale data

sets, finding exactly those narrations associated with the topic of interest is coupled with extensive manual labour and is often not feasible. We can apply a more feasible top-down strategy by relying on the structured narrative as a template for the general data narration.

First Step: Matching Events and Entities In the first matching step, we focus on the narrative's nodes, i.e., the entities and events that partake in the narration. Let us consider the causal relation $r = (\textit{Elevated Blood Pressure}, \text{risk factor of}, \textit{Cardiovascular Diseases}) \in R_e$, as well as the data set $D_e = (Var_e, T_e) \in DS$ shown in figure 1 [8]. In order to assess whether there exists a successful narrative binding between r and D_e, we have to identify those data values suitable to represent the respective nodes of r. We do so by matching the narrative's nodes to the values of the data both horizontally and vertically.

The horizontal matching for a node $v \in V_e$ aims at identifying the subset of variables, i.e. $Var_e^v \subseteq Var_e$ that can be semantically associated with v. By looking into D_e, we can see that each tuple represents an individual patient and captures different properties, like the blood pressure level in the second column, or observed events, like a diagnosis regarding heart health in the last column. Usually, descriptive meta-information gives insights about the variables of the data set, thus providing valuable hints about the entities and events referred to in the data. By referring to such meta-information, we can infer that only the second and fourth columns contain data values relevant for a binding of r. While the process of horizontally restricting the data is relatively straightforward in our use case, this is unfortunately not guaranteed. The high heterogeneity in the structure and formatting of meta-information in open data sets makes this a non-trivial task. For example, additional information about the context of a clinical study could be attached externally as a description of the data set, e.g., the study containing only patients with diagnosed diabetes. Similarly, such information may be directly embedded into the data, e.g., the measuring units such as *mm HG* could be included as individual variables of the data set. In order to find precise data narrations, it is therefore essential to consider all available meta-information. Finding ways to effectively identify and assign the correct meta-information to the correct data thus remains a challenging research task.

At this point, it has to be noted that for a successful matching between data set and narrative, both components must draw from a shared vocabulary. Primary candidates for such a vocabulary can be found in extensive and, preferably, well-curated ontologies. For many of these ontologies, specialized NLP-tools (such as SciSpacy [16] for biomedical terms) exist that allow for complete annotation pipelines. Using these tools on both the narrative's node labels and the meta-information of the data set yields annotations that we can apply in the matching.

The vertical matching for a node $v \in V_e$ aims at identifying the subset of tuples, i.e. $T_e^v \subseteq T_e$ whose substitutions of the variables in Var_e^v fulfill constraints imposed by additional qualifiers for v. When considering the node *Elevated Blood Pressure* for example, it becomes apparent that we are only interested in those tuples of the data set that show a sufficiently high value for the respective variable. Although in some cases it might be sufficient to define these qualifiers relative to the actual values, e.g., higher than the average, it will in most cases be inevitable to rely on external domain knowledge.

As the matching result, we receive for each node $v \in V_e$ the set of semantically associated data values consisting of the substitutions for the variables in Var_e^v of each tuple in T_e^v.

Second Step: Matching Relations Once we matched all nodes to their associated data values, we can now focus on the relations of the narrative. Essentially for N_e to be a feasible data narration for D_e, the relations occurring between the individual nodes of the narrative have to also occur between those values matched to them in the previous step. Here it is important to note that a narrative might express a large variety of different relation types. Identifying these relations requires the application of specialized metrics and strategies for each type. While this might sound difficult to realize, we believe that most narrative relations can be assigned into three categories: correlation, causation, and temporal relations. By being able to make reasoned decisions about whether such a relationship occurs between data set values, we can thus expect to be able to handle most narratives. Finding suitable metrics and strategies for these three types is thus a focus of our ongoing research. Deciding whether two columns correlate with each other is a trivial task that can be solved by applying metrics such as the Pearson correlation coefficient. Causation, on the other hand, is complicated to assess. Here, relying on metrics as deployed in clinical studies, such as the relative risk, that build upon the idea of control groups to analyse the effect some factor has on an observed event might be a promising first step. Additionally, we currently analyse the applicability of more sophisticated rule-based approaches as an easy-to-use strategy for causality assessment.

Interpreting the Matching Results With the outlined two-step matching strategy, we can now assess the reproducibility of a narration by analysing which claims and insights available open data sets can support, thus giving valuable insights into the narration's plausibility. It is thereby possible to discover multiple data sets whose data narrations align to a single narrative. In that case, we can assume that this narrative as a whole is plausible across many domains and maybe even generally applica-

ble. On the other hand, it might be possible that certain parts of the narrative rarely match any data, which could indicate that the respective substories of the narrative are very context-dependent. Thus, even if only parts of a narrative result in successful bindings, it is still possible to draw valuable conclusions about individual claims. For some cases, it might even be feasible to allow for some form of data set augmentation, i.e., combining individual bindings against multiple different data sets to support the narration as a whole.

5. Conclusion and Future Work

Assessing whether claims and statements encountered in scientific publications are representative and thus reproducible in various use cases usually requires a thorough and careful analysis of the underlying data. For this, it is necessary to identify the data narrations formed by the intrinsic relations inside the data sets. In this paper, we outline an approach that relies on translating the claims of scientific publications into structured narratives that form valuable templates for the discovery of additional data sets that can support these claims. For this, we propose a novel two-step matching strategy. As a first step, we rely heavily on the meta-information provided with each data set in order to identify those data values that align to the entities and events encountered in the narrative. This first step thus allows us to identify the relevant parts of the data set that we need to analyse in the second step. We then compute a narrative binding between the data set and the narrative. If the intrinsic relations between the individual data values match the relations expressed in the narrative, we consider the data set to be successfully bound to the narrative. In this case, we argue that the data can reproduce the claims encountered in the publication. In the near future, we would like to build upon this paper by developing and describing the complete matching algorithm in detail and evaluate its practicality in a large-scale evaluation of real-world data. For this, we will focus on solving the open research questions discussed throughout this paper.

References

[1] M. Pawlik, T. Hütter, D. Kocher, W. Mann, N. Augsten, A link is not enough – reproducibility of data, Datenbank-Spektrum 19 (2019).

[2] J. Pakstis, H. Calkins, C. Dobrzynski, S. Lamm, L. McNamara, Advancing reproducibility through shared data: Bridging archival and library practice, in: 19th ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2019.

[3] D. Hienert, D. Kern, K. Boland, B. Zapilko, P. Mutschke, A digital library for research data and related information in the social sciences, in: 19th ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2019.

[4] C. Meghini, V. Bartalesi, D. Metilli, Representing narratives in digital libraries: The narrative ontology, Semantic Web 12 (2021).

[5] S. Zaninovic, I. Nola, Management of measurable variable cardiovascular disease' risk factors, Current Cardiology Reviews 14 (2018).

[6] H. Kroll, D. Nagel, W.-T. Balke, Modeling narrative structures in logical overlays on top of knowledge repositories, in: International Conference on Conceptual Modeling (ER), Springer, 2020.

[7] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, V. Froelicher, International application of a new probability algorithm for the diagnosis of coronary artery disease, The American Journal of Cardiology 64 (1989).

[8] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.

[9] L. Rupprecht, J. C. Davis, C. Arnold, Y. Gur, D. Bhagwat, Improving reproducibility of data science pipelines through transparent provenance capture, Proc. VLDB Endow. 13 (2020).

[10] S. Samuel, Integrative data management for reproducibility of microscopy experiments, in: The Semantic Web - 14th International Conference, ESWC, volume 10250, 2017.

[11] F. Limani, A. Latif, K. Tochtermann, Linked publications and research data: Use cases for digital libraries, in: 22nd International Conference on Theory and Practice of Digital Libraries, TPDL, volume 11057, Springer, 2018.

[12] T. Friedrich, A. O. Kempf, Making research data findable in digital libraries: A layered model for user-oriented indexing of survey data, in: IEEE/ACM Joint Conference on Digital Libraries, JCDL, IEEE Computer Society, 2014.

[13] M. N. Hussain, H. A. Rubaye, K. K. Bandeli, N. Agarwal, Stories from blogs: Computational extraction and visualization of narratives, in: Proceedings of Text2Story - Fourth Workshop on Narrative Extraction From Texts, CEUR-WS.org, 2021.

[14] M. T. Rodríguez, S. Nunes, T. Devezas, Telling stories with data visualization, in: Proceedings of the 2015 Workshop on Narrative & Hypertext, 2015.

[15] E. Segel, J. Heer, Narrative visualization: Telling stories with data, IEEE Transactions on Visualization and Computer Graphics 16 (2010).

[16] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: Proc. of the 18th BioNLP Workshop and Shared Task, ACL, 2019.