Demonstrating Narrative Bindings: Linking Discourses to Knowledge Repositories

Hermann Kroll
Institute for Information Systems
TU Braunschweig, Braunschweig, Germany
kroll@ifis.cs.tu-bs.de

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

Morris Kunz
Institute for Information Systems
TU Braunschweig, Braunschweig, Germany
morris.kunz@tu-braunschweig.de

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

Abstract

Scientific discourse is based on exchanging knowledge in the form of stringent and well-arranged argumentations. Capturing the full variety and flexibility of argumentations within a single knowledge repository is close to impossible. Therefore, we have designed a narrative model that, on the one hand, captures the key aspects of argumentations (entities, events, relations, etc.), and on the other hand, is a logical overlay on top of existing knowledge repositories. Hence, users may formulate a narrative and validate its plausibility with data of different knowledge repositories via narrative bindings. This paper describes and discusses the computation of narrative bindings against three different types of sources: the document collection PubMed, the knowledge graph Wikidata and the WHO data sets. We give insights into the computation of narrative bindings and discuss open research questions.

1 Introduction into Narrative Models

The idea of discourse in science means to exchange knowledge in the form of stringent, well-arranged, and interconnected arguments. Whether they are written or told, these argumentations come with a clear underlying structure to make them logically sound and convincing. Capturing such full-fledged argumentation structures in a comprehensive and structured fashion is still ongoing research, see e.g., [HG17, MM11].

On a purely representational level, the Resource Description Framework (RDF) recommends to encode and share knowledge in the form of triples called facts [MMM$^+$04]. Such facts are generally expected to be trustworthy and always valid [KKN$^+$20]. For instance, simple properties like persons' names or birth dates can be used without worrying about their validity. But indeed, sharing knowledge in arguments is more complicated than just somehow connecting valid facts [HCF02]. For example, the conclusions of clinical trials are only valid within the (often rather limited) scope of their surrounding argumentation, i.e., the trial's context spans the scope of the concluded piece of knowledge. When being restricted to encode knowledge as facts like in RDF, mapping

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Finlayson (eds.): Proceedings of the Text2Story'21 Workshop, Online, 1-April-2021, published at http://ceur-ws.org

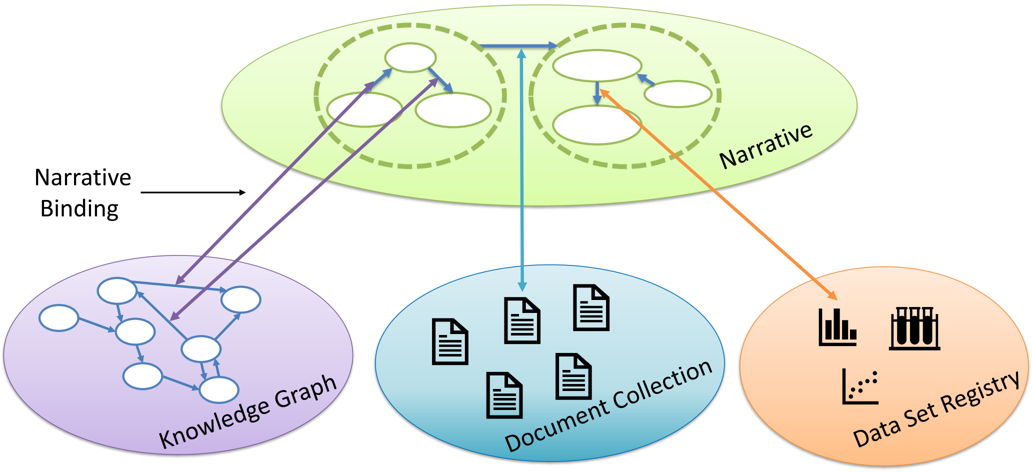

Figure 1: Narratives as Logical Overlays on top of Knowledge Repositories [KNB20]

detailed and complex argumentation structures is hardly possible, cf. [Suc20] for a good overview. What makes these structures so hard to catch?

Stripping argumentations down to their basic components, their backbone is formed by **narrative relations** between **arguments**, i.e., narrative relations describe the temporal and causal structure between individual arguments [HCF02, Tou58]. Disassembling such a comprehensive structure into reusable pieces remains a mostly unsolved problem, since some arguments may only be valid if other arguments were made before. Instead of focusing on the extraction and disassembly of argumentation structures, we argue for a logical overlay on top of existing knowledge repositories. The basic idea is similar to peer-to-peer networks, where high-level structures group the underlying IP-network in logical layers. That is why we have designed a formal **narrative model**, published in [KNB20].

On the one hand, our narrative model allows the formulation of complex structures by featuring factual knowledge, events, entities, actors, and narrative relations. On the other hand, the model is designed to be a logical overlay on top of arbitrary knowledge repositories, see Fig. 1. We have designed **narrative bindings** to bind parts of the narrative model against different sources. These narrative bindings give evidence for a narrative's relation by validating its plausibility with a knowledge repository's data. So instead of integrating different sources into a single one, we argue to bind the knowledge in a logical overlay for different applications. We believe that each knowledge repository has its unique purpose and cannot be integrated without losing information.

Hence, two questions must be answered: How can we model narrative structures? We have already discussed this question in detail [KNB20]. The second question is yet open: Suppose a narrative model and a set of knowledge repositories were given: How can we find suitable narrative bindings automatically or at least semi-automatically? This paper gives an overview about the automatic computation of narrative bindings, answering the previous question. We demonstrate and discuss how these bindings are computed against the document collection PubMed, the knowledge graph Wikidata, and the WHO data sets (Global Health Observatory of the WHO). We believe that narratives as logical overlays on top of existing knowledge repositories offer a novel way to represent scientific discourse. Possible applications would range from hypothesis testing (formulate a hypothesis as a narrative and test if narrative bindings exist) up to inferring new knowledge (infer new knowledge if a narrative could be bound) [SWB+14].

2 Computing Narrative Bindings

Let us explain our narrative model by giving a biomedical example narrative: *Aspirin is a drug that treats headaches and can be administered as a tablet. In a patient's treatment, aspirin is applied as a tablet to treat headaches. This treatment leads to an observation of an adverse effect, namely the medical condition alicylate toxicity.* This little narrative is a scientific discourse about an adverse effect of an *aspirin* treatment. Let us decompile this example narrative into smaller pieces: *aspirin* is a drug, *headaches* and *alicylate toxicity* are diseases and *tablet* is a dosage form. We represent them as **entities**, i.e., biomedical concepts of interest. The process of *a patient's treatment* has some temporal component, i.e., it starts at some point in time and has a

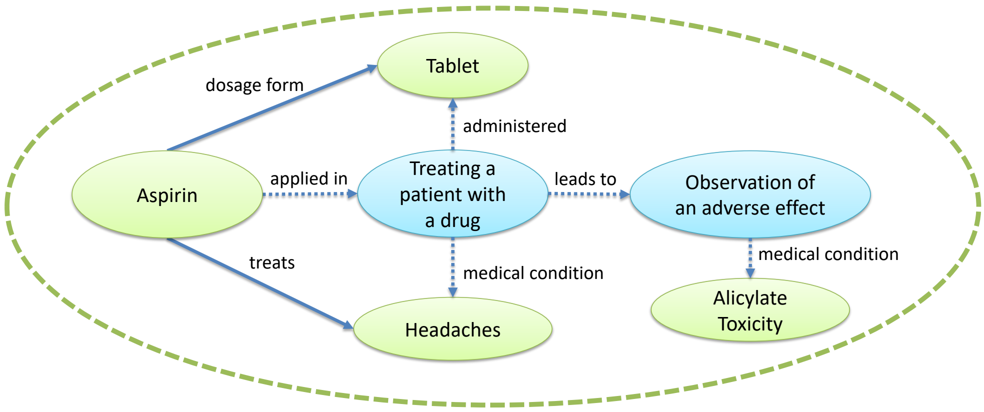

Figure 2: Example Narrative: Aspirin is a drug that treats headaches and may be administered as a tablet. In a patient's treatment, aspirin is administered as a tablet to treat headaches. The treatment leads to the observation of an adverse effect, namely alicylate toxicity.

duration. The treatment involves the entities *aspirin* (applied drug), *tablet* (administered as) and *headaches* (medical condition to treat). The *patient's treatment* leads to an *observation of an adverse effect* that involves the medical condition *alicylate toxicity*. Both, *patient's treatment* and the *observation of an adverse effect*, are **events**, i.e., they describe some expected states and may have a temporal component. The example narrative might thus be interpreted by having some order, e.g., the treatment should happen before the adverse effect is observed. The whole narrative is depicted in Fig. 2. We visualize entities as green nodes, events as blue nodes and relations as directed-labeled edges between them. For this paper, we understand a narrative as a directed edge-labeled graph with nodes and edges. We distinguish nodes into three different types: 1. entities represent important and relevant concepts, 2. literals represent values like numeric expressions or strings, and 3. events represent some states or state changes. Next, an edge represents the relationship between two different nodes, i.e., an edge might express some entity's participation in an event or a relationship between entities and literals. We divide relations into two different sets: **factual relations** between entities and literals, and **narrative relations** between events and entities involved in events. Entities might here be seen as actors as well, i.e., a person might be included in a factual relation but also in a narrative relation by participating in some event. However, the terms entity and event might be very similar in a given scenario, e.g., an election might be understood as a) the election process and modeled as an event, or b) just as an important concept and modeled as an entity. We believe that such a decision depends on the specific use case. More details about our narrative model and bindings may be found in our previous work [KNB20].

We have designed our narrative model to be as generalizable as possible, i.e., we do not restrict entities, events, and relations. In our research, we formulate scientific discourses based on our model. But indeed, the model might be applied in other scenarios as well. In this paper, we are interested in validating a narrative's plausibility. Therefore, we have designed the so-called **narrative binding**. A narrative binding binds a single relation of a narrative against some knowledge repository. Suppose there is a binding between a relation and a knowledge repository. In that case, we assume the relation to be plausible because we have found evidence for it. For example, *aspirin treats headaches* might be bound against some scientific publication or the DrugBank database [WFG+17]. Here, the corresponding binding validates the plausibility of the *aspirin-headaches relation*. However, it should be noted that the binding just gives evidence for the plausibility. It might be possible that a knowledge repository, especially when considering scientific publications, states wrong information against which we compute a narrative binding. For this paper, we assume knowledge repositories to contain valid knowledge. Subsequently, the main question remains open: *How can we compute suitable narrative bindings against different kinds of knowledge repositories? And especially, what are the challenges with this?* The following will describe how narrative bindings could be computed against three different sources, namely the document collection PubMed [Nat21], the knowledge graph Wikidata [VK14] and the Global Health Observatory of the WHO [Wor21]. Therefore, we describe the current state of our prototype and discuss open challenges. We develop our prototype to assist biomedical researchers in their daily work, i.e., the prototype should validate whether a biomedical discourse is plausible.

2.1 Bindings against PubMed

PubMed is the world's most extensive biomedical library with more than 30 million assets (2020). Computing narrative bindings against such an extensive collection requires efficient methods. On the one hand, information retrieval methods may therefore be considered to find relevant information in a document's text [CL96]. For example, the Lucene project [Luc21] offers a variety of suitable methods like indexing, similarity measures, and retrieval methods. On the other hand, converting unstructured text into structured information could also be done to extract relevant information from text a-priori. For such a conversion, methods like entity linking, information extraction, and more come into mind, see [WDRS20] for a good overview. For our prototype, we apply a preprocessing step to convert document texts into an intermediate graph representation. These graph representations will be queried later to compute narrative bindings. In the preprocessing phase, we apply an entity linking and information extraction step to retrieve how entities are connected within the single texts. Entity linking is done by utilizing and developing biomedical annotation tools and suitable vocabularies. Information extraction is done by utilizing Natural Language Processing toolkits (like Stanford CoreNLP [MSB+14], Stanford Stanza [QZZ+20] and OpenIE6 [KAA+20]). We build upon open information extraction to bypass the need for training data. Our system cleans the heterogeneous output with entity-based filters and word embeddings. We keep only information about documents containing relevant entity annotation and interactions between them. We start by selecting the most relevant part of PubMed, i.e., all PubMed documents that contain information about drugs. Our entity linking has detected drugs in around 5.6 million different documents. In the future, we will incrementally increase the number of processed documents. Hence, the current prototype searches through these 5.6 million documents for now and is currently evaluated by ten biomedical experts. The evaluation should determine how good narrative bindings against scientific publications can answer their information needs. They formulate their information need as short narratives, and the prototype retrieves suitable narrative bindings. In detail, we analyze how good the annotation, extraction, cleaning, and retrieval performance is and how helpful the bindings are for their use cases. For example, *aspirin treats headaches in patients* could be bound against publications that support *aspirin treats headaches* and *aspirin treats patients*. In extension to that, we plan to implement support for literal and event annotation. Literals like numeric expressions or dates could be found via regular expressions, e.g., applied dosages or time information. Finding arbitrary events requires either designing pre-known vocabularies (think about pharmaceutical methods) or utilizing the latest retrieval methods. Here, we may consider methods like exploring entity-centric events [SG18] or event summarization [SG16]. We currently analyze whether the latest textual entailment methods could help estimate whether a relation involving events and entities is mentioned within a paragraph [LOG+19]. Given a relation between an event and an entity, we estimate which paragraphs contain the information and send the paragraph plus the relation to a language model for textual entailment to make the decision. This approach will be further evaluated to enhance the retrieval quality for narrative bindings against document collections.

2.2 Bindings against Wikidata

Knowledge graphs are well-suited for the storage of factual knowledge about entities. That is why we develop a component to compute narrative bindings against Wikidata (WD), a large, high-quality open knowledge graph [VK14]. Wikidata is known to cover knowledge in many domains and hence, is a good candidate to validate arbitrary narratives. Besides, Wikidata includes mappings between several vocabularies, e.g., the WD entity *aspirin(Q18216)* could easily be translated to the DrugBank identifier *DB00945*. Then, more relevant information could be retrieved from DrugBank. Given a narrative, we obtain all labels from entities, literals, and the relationships between them. These labels are not restricted to knowledge graph resource identifiers. Accordingly, we perform an entity linking against all WD entities and relations through their (English) labels and synonyms in an inverted index. We mark relations as *factual* if a match is found. Once the initial entity linking step is done, finding narrative bindings for narrative components with factual relations is straightforward. For every combination of subject, relation, and object, we execute a SPARQL query to check if a fact is present on WD and obtain narrative bindings. Homonymous entity and relation names require some disambiguation when linking. We deal with homonyms by evaluating more than one match. On the one hand, it is likely that SPARQL queries will not find matches for erroneously linked entities and relations. On the other hand, the user can easily resolve false positives when bindings are shown and explained by the WD entities' description. When one or more narrative bindings were found, we may use so-called qualifiers to provide evidence about a fact's validity. Wikidata introduces these qualifiers to attach provenance information to facts [HHK15, Wik21]. For example, *aspirin is a medication* could be bound against the WD fact *aspirin(Q18216), instance of (P31), medication(Q12140)*, attached with the qualifier linking the fact to its source. Here, the qualifier would reveal that the information was gathered from DrugBank. On the one hand, computing narrative bindings against a knowledge graph might be straightforward. On the other hand, a practical knowledge graph might come with various heterogeneity issues, i.e., duplicated entity entries, synonymous relationships, and missing information [KFEB20]. However, modern methods like *KnowlyBERT* propose to utilize modern neural language models to bypass some of these issues [KFEB20].

2.3 Bindings against Scientific Data

Large-scale data sets store experimental and statistical information. Contrary to structured knowledge graphs, simple data sets usually vary in many different ways, from the formatting of the data to the availability of descriptive meta-information. The heterogeneity of data sets makes computing bindings against them a challenging task [BCN20, BBN19]. Our first approach is therefore restricted to data sets that, on the one hand, store their information in a relational fashion, and on the other hand, also provide meta-data describing their individual columns. In order to compute a narrative binding between a narrative and a data set, we propose a two-step matching algorithm: 1. Computation of a meta-data-alignment to recognize which data set entries contain descriptive information about the narrative's nodes (entities and events). 2. Estimating whether the actual data justifies the narrative's relations between the aligned nodes. Hence, nodes are linked to columns and relations are expressed by the dependencies between the columns' values. For example, the WHO database [Wor21] has information about cholesterol levels and the prevalence of diseases. Hence, we may compute a binding between a narrative's relation like *cardiovascular diseases are caused by high cholesterol levels* and the WHO database. Hence, a binding against a data set can give evidence for a narrative's causal relation. Estimating causation is a difficult task, and for now, we build upon simple correlation metrics, e.g., the Pearson correlation coefficient. We utilize causality metrics such as relative risk [KBAP78], which have been established throughout clinical studies to minimize errors and remove confounders. Detecting causality is a challenging research area on its own [SBMU00]. Another application of such bindings can be seen in a semi-automated computation of narrative explanations for data sets, similar to the task of data visualization [SH10]. In practice, the heterogeneity of data sets can be a major issue for the computation of narrative bindings. Brickley et al. argue that data sets are published in different domains, have different structures and do not always provide useful metadata [BBN19]. Hence, the automatic understanding and analysis of data sets remains a hard challenge due to this heterogeneity. In the future, we will investigate methods tackling such problems in practice.

3 Discussion

In summary, we have developed our prototype to compute narrative bindings automatically. First, the kind of relation (factual or narrative) does not determine a knowledge repository's type. Evidence for factual information might be found in knowledge graphs. However, it could also be bound against a scientific publication or a data set. Second, computing narrative bindings requires browsing through extensive collections of knowledge. For example, if users want to test a hypothesis in an online-fashion, such a system must be responsive. Combining smart index techniques with the latest natural language methods seems promising here. The next step is to understand how useful a logical overlay is in real-world applications like hypothesis testing or inferring new knowledge [SWB$^+$14]. Testing a hypothesis could be done by formulating it as a narrative and searching for suitable narrative bindings as evidence. If the narrative could be bound, it might be likely that the hypothesis is valid. Inferring new knowledge can be done by introducing *variables* as narrative nodes. Here, we might utilize our example narrative by replacing the *alicylate toxicity* by a variable *?diseases*. We can infer more side effects of an *aspirin treatment* when we find a suitable substitution for that variable by finding bindings for a corresponding narrative. Our current retrieval prototype for PubMed can search through 5.6 million documents in nearly real-time and is under evaluation by ten biomedical domain experts. Already performed interviews and a first questionnaire indicate the prototype's usefulness for testing new hypotheses by retrieving precise document hits as bindings. We plan to enhance the prototype by integrating more knowledge repositories and featuring retrieval with events. The first techniques to compute bindings against knowledge graphs and data sets are currently under development and will be evaluated further.

Although our model comes with advantages, many questions are still not solved yet. How good is the quality of narrative bindings? Natural language processing methods are known to be error-prone due to the high complexity of natural language [WDRS20]. Practical knowledge graphs still have heterogeneity issues that may lead to false bindings [KFEB20]. Even worse, the fully-automatic understanding of a data set's structure is close to impossible. And even if all of these problems would be solved, the next question is: How can we ensure that the narrative bindings feature a similar context? Computing bindings for a single narrative against several sources facilitates the need for *context-compatibility* [KKN+20], e.g., experimental data of mouse treatments should not be combined with humans' clinical trial data. In this case, it might be sufficient to check whether the treatments' target groups are identical. However, depending on concrete use cases, such decisions might get more complicated, e.g., *pre-existing conditions*, *doses* and more.

All these questions are hard-to-answer if we think about bottom-up methods, i.e., extracting all causations from a data set or the argumentation structure of some document [WDRS20]. Our model proposes a top-down method, i.e., a user formulates the narrative she is looking for. If our system knows the narrative structure already, retrieval methods should yield higher quality. Think about modern NLP methods here: extracting all relations between arbitrary concepts in texts seems to be nearly impossible. But on the other hand, estimating if a relation is mentioned within a small paragraph already achieves high quality [LOG+19]. For now, we have built a retrieval prototype to compute narrative bindings against PubMed. We are currently working on the integration of Wikidata and the WHO data sets. In the future, we will increase the quality of retrieval methods and demonstrate the benefit of our narrative model for applications like hypothesis testing or inferring new knowledge.

References

[BBN19] Dan Brickley, Matthew Burgess, and Natasha Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, WWW '19, page 1365–1375, New York, NY, USA, 2019. Association for Computing Machinery.

[BCN20] Omar Benjelloun, Shiyu Chen, and Natasha Noy. Google dataset search by the numbers. In Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal, editors, *The Semantic Web – ISWC 2020*, pages 667–682, Cham, 2020. Springer International Publishing.

[CL96] Jim Cowie and Wendy Lehnert. Information extraction. *Commun. ACM*, 39(1):80–91, 1996.

[HCF02] Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? In *Science*, volume 298, pages 1569–1579. American Association for the Advancement of Science, 2002.

[HG17] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. In *Computational Linguistics*, volume 43, pages 125–179. MIT Press, 2017.

[HHK15] Daniel Hernández, Aidan Hogan, and Markus Krötzsch. Reifying rdf: What works well with wikidata? *SSWS@ ISWC*, 1457:32–47, 2015.

[KAA+20] Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3748–3761, Online, 2020. Association for Computational Linguistics.

[KBAP78] D. Katz, J. Baptista, S. P. Azen, and M. C. Pike. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, 34(3):469–474, 1978.

[KFEB20] Jan-Christoph Kalo, Leandra Fichtel, Philipp Ehler, and Wolf-Tilo Balke. Knowlybert - hybrid query answering over language models and knowledge graphs. In *The Semantic Web – ISWC 2020*, pages 294–310, Cham, 2020. Springer International Publishing.

[KKN+20] Hermann Kroll, Jan-Christoph Kalo, Denis Nagel, Stephan Mennicke, and Wolf-Tilo Balke. Context-compatible information fusion for scientific knowledge graphs. In *Digital Libraries for Open Knowledge*, pages 33–47, Cham, 2020. Springer International Publishing.

[KNB20] Hermann Kroll, Denis Nagel, and Wolf-Tilo Balke. Modeling narrative structures in logical overlays on top of knowledge repositories. In *Conceptual Modeling*, pages 250–260, Cham, 2020. Springer International Publishing.

[LOG+19] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[Luc21] Apache Lucene. Apache lucene. `https://lucene.apache.org`, 2021.

[MM11] Raquel Mochales and Marie-Francine Moens. Argumentation mining. In *Artificial Intelligence and Law*, volume 19, pages 1–22, 2011.

[MMM+04] Frank Manola, Eric Miller, Brian McBride, et al. Rdf primer. *W3C recommendation*, 10(1-107):6, 2004.

[MSB+14] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[Nat21] National Library of Medicine. Pubmed. `https://pubmed.ncbi.nlm.nih.gov`, 2021.

[QZZ+20] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages, 2020.

[SBMU00] Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2-3):163–192, 2000.

[SG16] Andreas Spitz and Michael Gertz. Terms over load: Leveraging named entities for cross-document extraction and summarization of events. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 503–512, New York, NY, USA, 2016. Association for Computing Machinery.

[SG18] Andreas Spitz and Michael Gertz. Exploring entity-centric networks in entangled news streams. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 555–563, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[SH10] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.

[Suc20] Fabian M Suchanek. The need to move beyond triples. In *Text2Story@ ECIR*, pages 95–104, 2020.

[SWB+14] Scott Spangler, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, and et al. Automated hypothesis generation based on mining scientific literature. In *Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD '14, page 1877–1886. ACM, 2014.

[Tou58] Stephen Edelston Toulmin. *The uses of argument*. Cambridge Univ. Press (Cambridge), 1958.

[VK14] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[WDRS20] Gerhard Weikum, Luna Dong, Simon Razniewski, and Fabian Suchanek. Machine knowledge: Creation and curation of comprehensive knowledge bases, 2020.

[WFG+17] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, and Ana et al. Marcu. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2017.

[Wik21] Wikidata. Help: Qualifiers. `https://www.wikidata.org/wiki/Help:Qualifiers`, 2021.

[Wor21] World Health Organization. Global health observatory. `https://www.who.int/data/gho`, 2021.