CrowdIQ: An Ontology for Crowdsourced Information Quality Assessments

Davide Ceolin[1], Dafne van Kuppevelt[2] and Ji Qi[2]

[1]*Centrum Wiskunde & Informatica, Science Park 123, 1098XG, Amsterdam, The Netherlands*

[2]*Netherlands eScience Center, Science Park 402, 1098 XH Amsterdam, The Netherlands*

Abstract

Fact-checking is a common journalistic practice adopted to verify the truthfulness of claims and information items. Because of the demanding nature of fact-checking, a significant amount of research has been devoted to the use of crowdsourcing to scale up this practice. The idea to use laypeople to fact-check information allows accessing a vast amount of human computation resources, but introduces an issue of reliability: when these tasks are performed by laypeople instead of experts, their quality might be questioned. In this paper, we introduce an ontology for modeling crowdsourced datasets of information quality assessments. We emphasize that we allow modeling information about the items evaluated as well as important metadata such as the authors of such assessments. The goal of this model is to favor interoperability among different datasets of the same kind, as well as to support internal analyses of the dataset themselves in terms of bias and reliability of the collected assessments.

Keywords

Information Quality, Crowdsourcing, Data Modeling

1. Introduction

Fact-checking is a common journalistic practice employed in order to test the veracity of claims to be reported in newspapers. In order to fact-check information, journalists look up information sources, appraise their reliability, extract information out of them, and then reason on the resulting information set. This sequence of operations is rather time-consuming, while the amount of claims that require fact-checking online is constantly growing. On the hand, this means that the workload of specialists is high and, on the other hand, their possibility to intervene real-time is limited. Crowdsourcing has revealed a useful tool to allow scaling up fact-checking. Crowd workers can, in fact, be instructed so to produce expert-like assessments, and by collecting multiple assessments about the same item (wisdom of the crowds), quality can be assured. In order to maximize the usefulness of the fact-checking (and, more in general, information quality assessment) datasets obtained from crowdsourcing, it is important to annotate them with specific metadata. This paper describes a lightweight ontology to describe and annotate crowdsourced information quality assessments. The ontology allows characterizing information about the items being assessed, the author of the assessment, and information

quality details. We leverage existing ontologies like the Data Quality Vocabulary (DQV) and Schema.org (Schema) and we select and specialize their elements to serve this specific niche of information. The paper is structured as follows. Section 2 presents related work. Section 3 introduces the ontology, while Section 4 discusses example applications. Section 5 concludes.

2. Related Work

Relevant to this line of work is the Data Quality Vocabulary (DQV) [1], that allows defining quality dimensions and metrics for annotating quality of data. Our model extends and specializes DQV in order to model specifically crowdsourced information quality assessments. In turn, this links to the extensive line of research on data quality modeling and measuring [2]. While we can observe similarities with the measurement and modeling of data quality, the information quality assessments we are interested in aim at assessing the information content of items, rather than their data serialization. For an overview on information quality and its philosophy, we refer the reader to the book edited by Floridi and Illari [3]. The Data Cube Vocabulary [4] is also a predecessor of our model, since it allows modeling multidimensional metadata. We see our model as a specialization of this vocabulary as well.

The work presented in this paper is also relevant to the field of FAIR data principles, as it aims at favoring findability (especially on principle F2) and interoperability (I1) of data. We refer the reader to the work of Poveda et al. [5] who provide an extensive analysis on this topic.

3. The CrowdIQ Ontology

The ontology we present here aims at achieving the highest interoperability while allowing modeling the necessary information resulting from the crowdsourcing tasks of interest. In the ontology, we identify three main "superclasses" that represent the core of the information we intend to model. We describe them as follows, while Figure 1 provides an overview.

Target Item   The target item represents the item to be evaluated. This ontology means to model quality assessments on information items online, therefore this class specializes the DigitalDocument class of Schema.org.

Worker   The worker class is meant to characterize the author of a quality assessment, and is thus modeled from the Person class of Schema.org. The instances of this class could be more or less populated depending on the level of anonymity granted to the crowd workers. Information to populate instances of the worker class can be provided both by the crowdsourcing platform or by the worker who responds to demographic questions in the crowdsourcing task.

IQ Assessment   This class models the information quality assessment that the worker provides about the target item. This class requires the specification of a quality dimension (e.g., precision, accuracy, truthfulness; see [3] for an overview of information quality dimensions) and of a metric. E.g., the precision of an online document might be expressed using a 5-level Likert scale. We also record provenance information (time, platform of creation).
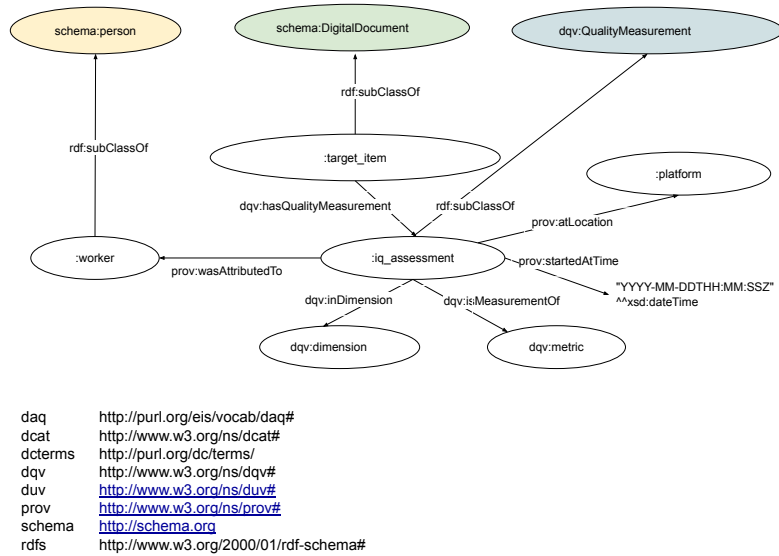
daq   http://purl.org/eis/vocab/daq#
dcat   http://www.w3.org/ns/dcat#
dcterms   http://purl.org/dc/terms/
dqv   http://www.w3.org/ns/dqv#
duv   http://www.w3.org/ns/duv#
prov   http://www.w3.org/ns/prov#
schema   http://schema.org
rdfs   http://www.w3.org/2000/01/rdf-schema#

Figure 1: Overview of the CrowdIQ Ontology

4. Applications

4.1. Data Conversion Requirements

The proposed ontology is meant to produce and annotate Linked Data representation of crowd-sourced information quality assessments. Given that these data are often produced in CSV format, we refer the reader to the CSV on the Web standard [6] to convert CSV files in RDF or JSON-LD. We also provide an example metadata file online[1] for converting CSV files using our ontology.

4.2. Analysis workflow in Knime

In the real-world research project "The Eye of the Beholder" [2], we aim to building or leveraging a platform for scholars to train machine learning pipelines to automatically assess quality of online information. To train these pipelines, we make use of a dataset collected from a crowd sourcing platform called CrowdFrame [7]. The training data contains information from three parts, namely worker information, document information, and assessment information. Details about the corresponding crowdsourcing experiment is available at [8].

Through investigation and comparison, we chose the KNIME [9], a free and open-source data analysis platform. KNIME integrates various components for machine learning, data mining, and reporting. Users can visually create workflows with these components and execute part or all of them, and subsequently use interactive views to examine results, models, etc.

---

[1]Available at https://github.com/EyeofBeholder-NLeSC/assessments-ontology/blob/main/metadata.json

[2]https://www.esciencecenter.nl/projects/the-eye-of-the-beholder-transparent-pipelines-for-assessing-online-information-quality/

The desired tool is composed of several KNIME workflows working in tandem with each other. These workflows automate the process of data exploration, model training, result interpretation, and pipeline comparison. To achieve this goal, we designed the CrowdIQ ontology to design the data interfaces for these workflows, and we also need to validate the training data files corresponding to this ontology before feeding the to those workflows.

4.3. User stories

The requirements for integrating the ontology in KNIME are specified as follows.

Upload data from the crowdsourcing platform Once the crowd sourcing task has been completed by enough workers, our user wants to load the resulting assessments into KNIME for further analysis. The platform provides our user with a URL of the metadata of the file (the url of the data itself can be inferred from the metadata). The KNIME component that we develop takes this URL as an input. The component downloads the data and metadata, and interprets the data according to the CrowdIQ ontology. It executes some sanity checks: it returns an error when, for example, no documents or no users are present. It outputs three tables: for workers, documents and assessments respectively. It also outputs a visualization of the realization of some of the attributes in the data: e.g., which dimensions and worker attributes are present.

Specify metadata When the user wants to provide data from a platform that does not give machine readable metadata compatible with our ontology, the user needs to specify which fields relate to which attributes in the data. Instead of providing an URL to the metadata as above, the user should be able to specify in a graphical user interface, for each column of their CSV data, which attribute in the ontology it represents. Depending on the format of the data that the user provides, the CSV data can either consist of one table with all information combined, or three different tables for workers, assessments, and documents respectively.

4.4. Proposed solutions

To account for the user stories as described in the previous section, we propose three types of validation to integrate in KNIME workflows. First, if the training data is stored in CSV format, a CSV-on-the-Web validation is essential. In this step, metadata annotations are added to the input CSV files (manually or by referring to an existing metadata to interpret the data files (encoding, data types of fields, etc.). At the same time, these files will be converted into a linked data format. Next, the obtained linked data should also be tested for consistency with the ontology. This is done through an ontology-based validation that maps the data fields to the expected classes/properties defined in the ontology and further cleans the data. Finally, the resulting data should also be validated against the constraints defined in the ontology. This can include various checks such as missing/duplicate properties, missing or improper type arcs, and inconsistent value ranges. In addition, we propose to extend the CrowdFrame platform to output a metadata file describing the output CSV according to the CSV on the web standard. In the interface for creating a new crowd sourcing task, the user should be able to specify which fields correspond to which attributes in the ontology, for those fields for which it cannot be

inferred automatically. A prototype of a Knime pipeline that reads CSV and validates against the ontology, can be found on GitHub[3]. It wraps the csvw library[4] in a Python scripting node.

5. Conclusion

In this paper, we introduce CrowdIQ, an ontology for modeling crowdsourced information quality assessments. The ontology aims at achieving minimal commitement, while specializing existing models to represent common information about information quality crowdsourced tasks. The ontology is described along with a possible utilization scenario. The goals of this model are twofold: allow interoperability among existing crowdsourced datasets, as well as to allow inspecting each dataset separately in order to assess its reliability and bias. Future work will aim at extending the model further and at providing pipeline automation components that leverage the model's expressivity.

Acknowledgements

References

[1] E. Hyvonen, R. Albertoni, A. Isaac, Introducing the data quality vocabulary (dqv), Semant. Web 12 (2021) 81–97. URL: https://doi.org/10.3233/SW-200382. doi:10.3233/SW-200382.

[2] V. C. Storey, R. Y. Wang, Modeling quality requirements in conceptual database design., in: I. N. Chengalur-Smith, L. Pipino (Eds.), IQ, MIT, 1998, pp. 64–87.

[3] L. Floridi, P. Illari (Eds.), The Philosophy of Information Quality, Springer, 2014.

[4] D. R. Richard Cyganiak, The RDF Data Cube Vocabulary, 2014.

[5] M. Poveda-Villalón, P. Espinoza-Arias, D. Garijo, O. Corcho, Coming to terms with fair ontologies, in: Knowledge Engineering and Knowledge Management: 22nd International Conference, EKAW 2020, Springer-Verlag, Berlin, Heidelberg, 2020, p. 255–270.

[6] J. Tennison, Csv on the web: A primer, https://www.w3.org/TR/tabular-data-primer/, 2016.

[7] M. Soprano, K. Roitero, F. Bombassei De Bona, S. Mizzaro, Crowd_frame: A simple and complete framework to deploy complex crowdsourcing tasks off-the-shelf, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22, ACM, 2022, p. 1605–1608.

[8] M. Soprano, K. Roitero, D. La Barbera, D. Ceolin, D. Spina, S. Mizzaro, G. Demartini, The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale, Information Processing Management 58 (2021) 102710.

[9] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz Information Miner, in: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007.

---

[3]https://github.com/EyeofBeholder-NLeSC/assessments-ontology
[4]https://github.com/cldf/csvw