



SCHEMA INFERENCE WITH EXPLAINABLE AI FOR DATA ENGINEERING IN GOVERNMENT INSTITUTIONS

SYSTEMATIC REVIEW PROTOCOL

 Christian Koch and  Dirk Riehle

1. INTRODUCTION

Data engineering is an integral part of the data science process [1] [2]. In CRISP-DM the equivalent process stage is called “data preparation” [3, pp. 23-26]. It comprises tasks such as data ingestion, data transformation, and data quality assurance. In order to fulfill these tasks, schema inference is an important capability [4]. Its goal is to detect the structure of a dataset and to derive metadata on hierarchies, data types, etc. [5] Artificial intelligence (AI) has the potential to automate schema inference and thus increase the efficiency of the data science process. However, as government institutions are subject to special regulations, explainability of AI models can be a mandatory requirement. Transparency of AI models is particularly relevant for high-risk systems in domains such as employment services, law enforcement and critical infrastructure.* Goal of this research protocol is to plan a systematic review of literature on schema inference with explainable AI (XAI) for data engineering in government institutions. The document forms the first element of a series of publications in the quest for AI models that a.) increase the efficiency of schema inference and b.) are explainable to both end users and regulators in government organizations.

2. REVIEW QUESTION

Purpose of the planned review is to systematically collect and analyze relevant literature on the following question: *What explainable AI models can government institutions use for schema inference in data engineering?* Key focus of the review lies on schema inference in data engineering and explainability. It is assumed that AI models meeting

Key words and phrases. Schema Inference, Explainable Artificial Intelligence, Explainable Machine Learning, Data Engineering, Data Science, Government.

*This document follows the concept of high-risk systems as defined in the Artificial Intelligence Act of the European Union [6]

these requirements are in principle qualified for the use in government institutions—even in high-risk environments.

3. INCLUSION CRITERIA

In accordance with the search strategy described below, the review will include all studies that address the use of explainable AI for schema inference in the field of data engineering. For the term "artificial intelligence", we rely on Tegmark's definition. Here, AI is the ability of a machine to "accomplish complex goals" [7, p. 39]. Thus, we implicitly identify schema inference as such a complex task. Since "machine learning" (ML) is an important technique in AI development, we include this expression in the list of initial search terms. For the notion "schema", we refer to Mlynková [8, p. 16], who interprets an XML schema as a formal, context-free grammar, following Berstel and Boasson [9]. Since our primary interest is in data engineering, we define schema inference accordingly as the process of describing the structure of a dataset using a formal language. Our analysis does not target particular schema types, like XML or database schemas, but considers all variants.

As pointed out by Vilone and Longo [10], there is no universally accepted notion of "explainability" in the context of XAI. In our review, we build on a human-centred definition, formulated by the Defense Advanced Research Projects Agency (DARPA). Accordingly, XAI techniques "enable end users to better understand, trust, and effectively manage artificially intelligent systems" [11, p. 1] [12]. This definition does not necessarily demand interpretability, where users are expected to understand the inner workings of an AI model. Instead, the definition allows for black-box algorithms that generate results in a way that is trustworthy to end users.

Although the focus of this review lies on government entities, publications from other sectors, such as banking or manufacturing, will be considered as well. We include all English literature; publications in other languages will be excluded from the study. Our main interest is in peer-reviewed articles. However, if the search strategy does not yield a sufficient number of documents, we will also consider other publication types.

4. SEARCH STRATEGY

Adapting the example of Pearson et al. [13], the search strategy comprises the following three stages:

- (1) Limited search to identify relevant keywords contained in title and abstract of the results.
- (2) Terms identified in this way, and the synonyms will be used in an extensive search of literature.

- (3) Reference lists and bibliographies of the articles collected from those identified in stage two will be searched.

Our search will be performed using the subsequent search engines/digital libraries:

- Google Scholar
- ACM Digital Library
- IEEE Xplore

For the initial search, we use the following terms:

- ("artificial intelligence" OR "machine learning") AND ("explainable" OR "explainability") AND "schema inference" AND "data"

To not overly limit the search results, we omit the term "government" from our initial list. Moreover, we exclude the word "engineering" and only include the term "data". In addition to the classical search described above, we will also address our question to the following large language models (LLMs):

- ChatGPT
- LlamaChat

Responses from LLMs are discussed in a dedicated section of the resulting paper, separate from regular search results.

5. DATA COLLECTION

Following the example of Pearson et al. [13], full copies of the articles identified in the search meeting the inclusion criteria based on their title and abstract are collected for data analysis. Publications identified through reference lists and bibliographies will be considered based on their title. Two reviewers will independently search for articles. Discrepancies in reviewer selection will be resolved in a meeting before finalizing the data collection.

6. DATA ANALYSIS

In the analysis stage, we will systematically arrange and analyze collected literature. In a first step, literature is divided into categories. These categories are defined by reviewers in an iterative process. In the second step, literature will be evaluated both quantitatively and qualitatively. The quantitative analysis focuses primarily on document metadata such as number of items collected, year of publication, category, etc. Qualitative analyses will focus on the text of the selected literature. Here we utilize the general inductive approach proposed by Thomas [14]. In addition, our qualitative analysis will apply methods of natural language processing. Examples are n-grams and their pointwise mutual information (PMI), as suggested by Curch and Hanks [15].

7. IMPLEMENTATION AND OUTLOOK

In order to ensure the quality of this protocol before implementation, its content was validated using the checklist of Moher et al. [16]. Stages planned in this protocol may be performed in several iterations. Each iteration will pass all the phases described above. Newly identified or changed requirements may lead to an update of this protocol and an adaptation of the review procedure.

APPENDIX A. GLOSSARY

Term	Definition
AI	See Artificial Intelligence
Artificial Intelligence	Ability of a machine to accomplish complex goals.
Explainable Artificial Intelligence	Machine learning techniques enabling end users to better understand, trust, and effectively manage artificially intelligent systems.
ML	Machine Learning
Schema Inference	Process of describing the structure of a dataset using a formal language.
XAI	See Explainable Artificial Intelligence

REFERENCES

- [1] Oscar Romero and Robert Wrembel. Data engineering for data science: two sides of the same coin. In *Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22*, pages 157–166. Springer, 2020.
- [2] Meike Klettke and Uta Störl. Four Generations in Data Engineering for Data Science: The Past, Presence and Future of a Field of Science. *Datenbank-Spektrum*, 22(1):59–66, 2022.
- [3] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. CRISP-DM 1.0: Step-by-step data mining guide. SPSS, 2000.
- [4] Vraj Shah and Arun Kumar. The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. DEEM’19, New York, NY, USA, 2019. Association for Computing Machinery.
- [5] Pavel Koupil, Sebastián Hricko, and Irena Holubová. A universal approach for multi-model schema inference. *Journal of Big Data*, 9(1):1–46, 2022.
- [6] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *EUR-Lex*, 2021.
- [7] Max Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Penguin, 1 edition, 2017.
- [8] Irena Mlynková. An analysis of approaches to xml schema inference. In *2008 IEEE International Conference on Signal Image Technology and Internet Based Systems*, pages 16–23, 2008.
- [9] Jean Berstel and Luc Boasson. Xml grammars. In Mogens Nielsen and Branislav Rován, editors, *Mathematical Foundations of Computer Science 2000*, pages 182–191, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.
- [10] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.
- [11] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. Darpa’s explainable ai (xai) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021.
- [12] Defense Advanced Research Projects Agency. Explainable artificial intelligence. <https://www.darpa.mil/program/explainable-artificial-intelligence>. Accessed 2023-09-18.
- [13] Alan Pearson, John Field, and Zoe Jordan. *Evidence-Based Clinical Practice in Nursing and Health Care*, pages 173–176. John Wiley & Sons, Ltd, 2006.
- [14] David R. Thomas. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2):237–246, 2006.
- [15] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.
- [16] David Moher, Larissa Shamseer, Mike Clarke, Davina Gherzi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A Stewart. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic reviews*, 4(1):1–9, 2015.

CHRISTIAN KOCH, TECHNISCHE HOCHSCHULE NÜRNBERG GEORG SIMON OHM, NUREMBERG /
FAU ERLANGEN–NUREMBERG, GERMANY

Email address: christian.koch@th-nuernberg.de

DIRK RIEHLE, FAU ERLANGEN–NUREMBERG, GERMANY

Email address: dirk.riehle@fau.de