DOCTORAL CONSORTIUM

CONCEPT MINING FOR DATA MANAGEMENT

DOI
https://doi.org/10.18690/um.fov.4.2026.74

ISBN
978-961-299-152-4

SWARUPA HARDIKAR[1, 2]

[1] HAN University of Applied Sciences, Arnhem, the Netherlands
swarupa.hardikar@han.nl
[2] Radboud University, Nijmegen, the Netherlands
swarupa.hardikar@han.nl

Keywords:
concept mining,
data management,
text mining,
terminology extraction,
term disambiguation

The PhD project addresses data ambiguity, a core issue hampering the management of organisational data, and aims to support data management through concept mining. We intend to (semi-)automatically analyse perceived meanings of terms and concepts within an organisation's data domain to aid terminological disambiguation. The concept mining system will deploy text mining and information retrieval techniques within unstructured data. We adopt a 3-phase approach of exploring the problem and requirements, iterative development and testing, and finally, an extrinsic evaluation in a real-life setting. Every component of the system will be intrinsically evaluated with metrics best suited to that component. The toolkit will also be evaluated qualitatively through studies with data managers for efficiency, quality and real-world applicability. The project will bridge the gap between domain expertise and automated methodologies, with expected contributions such as an evaluation framework for concept mining and advancement towards the research of terminology extraction.

University of Maribor Press

1 Introduction

1.1 The Importance of Data Meaning

Across organisational domains, people use language differently, and data owners may frequently develop their own interpretations of key terms (Alexopoulos, 2020). This may cause ambiguity in communication and a lack of data overview - for example, the term "model" might describe a machine learning system, a conceptual representation, or a 3D simulation depending on the domain the person using it operates in. It has been estimated by many studies that decision-making caused by the poor quality of information resulted in losses of thousands of dollars within organisations (Castillo et al., 2017). Per a McKinsey report, an average of 30% of time is spent by organisations on non-value added tasks due to poor data quality and availability (Petzold et al., 2020).

To reduce ambiguity and streamline communication, managing business vocabularies and clarifying the definitions of terms is crucial (Ross, 2020; Voskuil, 2015; Hoppenbrouwers, 2003). Consequently, to communicate about data meaning, having a description of the semantics is key.

1.2 The Challenge

Curating lexical assets such as business glossaries is typically the responsibility of a data steward or manager - however, the amount of organisational data can make the task resource-intensive (Henderson et al., 2017).

From multiple interviews with data management professionals, it was confirmed that data managers require insight into organisational term usage and meaning to make informed decisions about useful terms, concepts, and semantic descriptions. It was articulated more specifically that organisations are currently largely facing an issue with data semantics. Interviewees stated that reconciling terminological ambiguity was an active pain point which could potentially result in monetary losses.

As effective data management relies on the combination of qualitative domain expertise and automated methodologies, our approach respects the DAMA Data Management Body of Knowledge principles (Henderson et al., 2017) by focusing on the structured extraction and validation of domain-specific lexical assets.

Currently, such ambiguity is resolved by way of multiple consultations and meetings with domain experts. On the other hand, such domain experts are often a scarce resource (personal correspondence, Data Architect), and in addition, interviewees noted that such prolonged practices can contribute to employee fatigue (personal correspondence, Domain Architect). As such, we believe that (semi-)automated approaches within concept mining may expedite the process as well as better reflect the ground reality.

1.3 Concept Mining for Data Management

Efforts into the extraction and analysis of domain-specific terms within text data have been observed within the fields of natural language processing and terminology extraction (a subtask of information extraction). However, the latter still suffers from issues such as polysemy (usage of terms within different contexts) and domain adaptability (generalisability across a variety of domains) (Xu et al., 2025).

Given that effective data management relies on the marriage of qualitative domain expertise and automated methodologies, our approach aligns with the DAMA DM-BOK (Data Management Body of Knowledge) principles (Henderson et al., 2017) by focusing on the structured extraction and validation of domain-specific lexical assets, while also drawing upon components from FAIR (findable, accessible, interoperable and reproducible; Wilkinson et al., 2015) guidelines. The project will bridge the gap between qualitative domain expertise and automated methodologies.

Thus, we believe Concept Mining (CM) could be used to support data management. CM has been applied in other contexts such as formalisation (Yang et al., 2008), information retrieval (Liu et al., 2019) or data sharing for medical applications (Si et al., 2019). In this project, we focus specifically on methods to support data managers to gain overview of perceived interpretations by data users, and to support selecting and recording definitions.

2 Proposed Research

Given the above considerations, we formulate the following research question:

How can a concept mining system be designed, applied, and evaluated to support data managers in analysing concepts based on perceived interpretations of terms by data users?

The research question addresses (a) the translation of the needs of data managers into requirements and selection of text mining techniques ("*how to design*"), (b) the development, comparison and evaluation of the detection and/or generation of synonyms, homonyms and definitions for a given term using CM ("*how to apply*"), and (c) an evaluation of the concept mining system based on efficiency, effectiveness and usefulness ("*how to evaluate*").

The research question is supported by the following sub-questions:

1. How can we extract information from relevant data to determine the relevance of a concept given a specific goal of a data manager?
2. What information does a data manager need to adequately manage concepts with multiple interpretations (homonyms)?
   How can we extract homonymous concepts from data?
3. What information does a data manager need to find related concepts (synonyms)?
   How can we identify synonyms in data?
4. What information does a data manager need to determine which definition fits a concept in a specific context?
   How can we select, link or generate possible definitions from data?
5. How should information be represented to support different types of meaning descriptions?

Figure 1 draws upon the DIKW (data-information-knowledge-wisdom) model (Wien et al., 2006) to encapsulate our contribution in the development of data and information into meaning and knowledge, with subject matter expertise contributing to contextualisation and validation (note: we do not specifically adhere to the model's 'wisdom' layer; it is out of scope of the project).

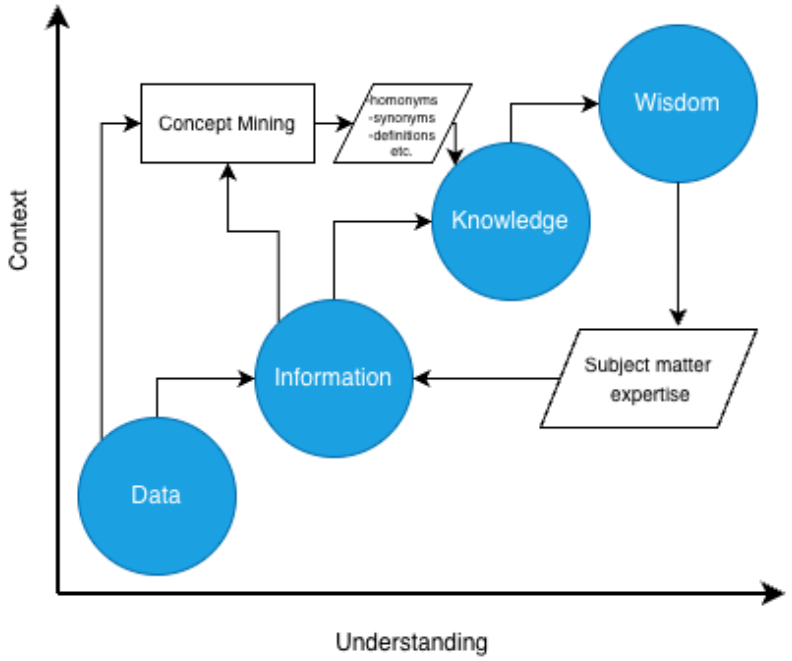

Figure 1: Concept Mining within the DIKW model
Source: Own

3 Research methodology and proposed experiments

Our research design is rooted in the Design Science Research (DSR) paradigm, aiming to improve scientific/technological bases by creating innovative artifacts that empower the working environments where they are implemented (Vom Brocke et al., 2020). In this case, the artifact would be a tool(kit) that deploys CM within organisations to support data managers.

The approach and methodology are also based on the DSR process model (Peffers et al., 2007). We first define the problem and objectives, i.e. the requirements (subtasks) of the solution (the concept mining toolkit). We then identify potential techniques that can help satisfy the requirements and test these techniques on the data. Finally, we evaluate the system quantitatively and qualitatively.

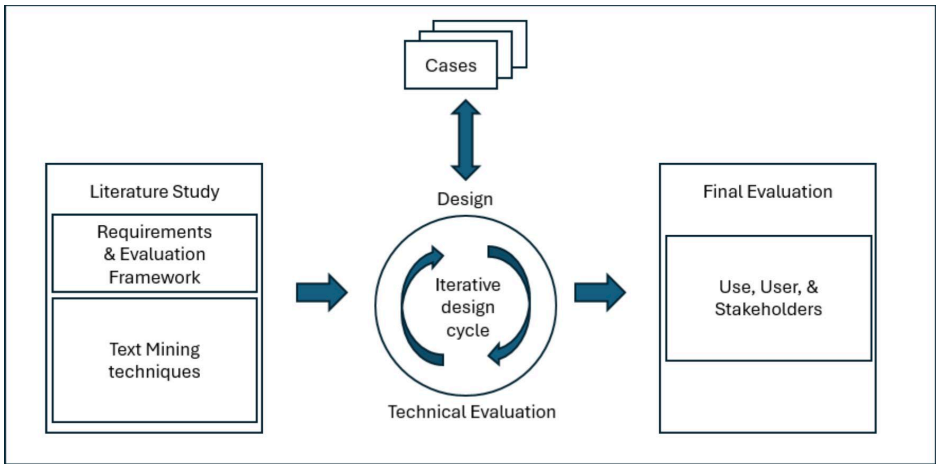

Figure 2: Phased design of the research approach based on the DSR model.
Source: Own

The three-phase approach spanning the duration of the PhD is shown below:

Phase 1: Exploring requirements and industry input

The phase is primarily a preparatory one. We have conducted an initial literature review to identify the issues surrounding good data management within organisations. We have also conducted research interviews with data professionals such as data architects, information managers and domain architects to validate the challenges of data management, the necessity of the informational value of data and the potential need for automating the extraction and analysis of concepts. Taking these into account, we have defined the research questions.

Phase 1 Steps:

i) A study of the relevant literature and analysis of SOTA text mining and IR techniques:
   a. Exploring pain points of data management and previous attempts at resolving them: Overview of literature
   b. Hypothesising and defining the problem

ii) Research interviews with data professionals (completed): We match the pain points discovered in the literature with findings from the industry, thereby defining the final problem statement

Phase 2: Iterative toolkit development via use cases and evaluation (Ongoing)

To implement the analyses conducted in the previous phase, we conduct a series of use case studies using concept mining. Each study consists of applying the aforementioned techniques, followed by evaluations. After every study, we aim to implement the findings into the next one, in combination to a revisitation of state-of-the-art (SOTA) techniques. In addition, we test the setup after each study for transferability and generalisability.

Phase 2 steps:

(i) Energy domain: abbreviation disambiguation for Dutch and English data (completed);
(ii) (ii-iii) Education and Banking domains. We also test the CM system for generalisability across domains.

Phase 3: Extrinsic evaluation of concept mining in a practical setting

In Phase 3, we plan to conduct a qualitative study involving experience, comments, feedback and advice from data management professionals from various organisations, pertaining to the utility and efficiency of the concept mining toolkit in various downstream data management tasks. The presentation of the concept mining system will be in the form of demos and live workshops where the user (typically a data professional) can implement it with their data. This will be followed by surveys and questionnaires that can be answered by the user, assessing the accuracy, efficiency and usability of the system.

Phase 3 steps:

(i) Qualitative assessment of concept mining
   a. Surveys and questionnaires to gather post-demo or post-live workshop responses from users

(ii) Test strategies based on the requirements of previous phases

4 Preliminary Results from Use Case #1: Acronym Disambiguation

Our first study explored the disambiguation of acronyms within domain-specific data with Alliander, a major energy company in the Netherlands. We investigate the applicability of state-of-the-art methods for multilingual acronym disambiguation (AD) on an open benchmark and an authentic domain-specific bilingual dataset to establish transfer from theory to practice.

Acronyms are often homonymous, with expansions that have similar or completely different meanings. It can be difficult to disambiguate them within organisational text, which may cause misinterpretation in downstream tasks, especially with in-the-wild domain- or company-specific acronyms. For example, the acronym 'AC' may stand for "Alternating Current" or "*Asbestcement*" (*Dutch: Asbestos cement*). The company lists numerous acronyms in use, with many highly domain-specific and company-specific usages. We found 176 acronyms with multiple long forms (expansions) each, which has raised concerns regarding their interpretation within the company (especially their usage within large language models).

Thus, we conducted a two-phase exploration for highly domain-specific and bilingual AD, creating a dataset of 176 examples for domain-specific bilingual (Dutch+English) data to help us reflect in-the-wild acronym usage in research.

Experimental setup:

We first curated a bilingual AD dataset with original Dutch and English examples from the company data. For a second dataset, we used an open benchmark for scientific acronym disambiguation in English.

We then tested the performance of two language models on the two datasets: AcroBERT (a pre-trained English AD language model) (Chen et al., 2023) and Ministral-3 8B (an open generative language model) (Liu et al., 2026), focusing on AD on highly domain-specific acronyms.

As an alternative resource-friendly tactic in the possible absence of an internal knowledge base, we also evaluated the use of Web search result pages in locating the correct acronym expansion given context keywords. This was done using DDGS,[1] a meta-search Python library for retrieving results across multiple search engines.

Results and Conclusion:

Table 1: Breakdown of performance on the Company Dataset by domain specificity and language

Category	Subset	Model	Accuracy
Domain-specificity	Company-specific	Ministral	0.5753
		AcroBERT	0.5342
	Domain-specific	Ministral	0.7869
		AcroBERT	0.5738
	Global	Ministral	0.9048
		AcroBERT	0.5952
Language	English	Ministral	0.8077
		AcroBERT	0.7692
	Dutch	Ministral	0.6867
		AcroBERT	0.5267

Source: Own

Table 2: Acronym expansion matches found in Web search result pages (max. 50 aggregated results)

Category	Subset	Model	Recall
Overall	All acronyms	39/103	37.86
Domain-specificity	Global	22/42	52.38
	Domain-specific	17/61	27.87
Language	English	15/23	65.22
	Dutch	24/80	30.00

Source: Own

For the first phase, we found that Ministral-3 8B can often identify the appropriate expansion in a non-English, domain-specific setting without finetuning, but it cannot generalise well enough to unseen data. For AcroBERT, despite its lower

[1] https://pypi.org/project/ddgs/

scores, its comparable performance in correctly identifying company-specific acronym expansions could indicate that finetuning it with a Dutch AD dataset would already improve performance, without the need for training on company-specific data, making it a resource-efficient alternative that handles unseen data better.

For the usage of search results from the Web, we found that our attempt at the extraction of domain-specific acronym expansions did not yield satisfactory recall. However, we believe that there may be promise in using it as a low-resource tool in the absence of an internal knowledge base to get diverse perspectives on the real-world usage of a given acronym.

We thus concluded that there remains a gap between theoretical and practical implementations of AD tools for multilingual domain-specific text.

5 Future Development

The above study indicated that in the case of non-English and domain-specific text data, there is still room for improvement in getting theoretically tested models ready for application in practice. On the other hand, it was an experiment on a smaller scale, examining acronyms as homonymous terms encountered on a daily basis. For the next study, we aim to implement these findings to scale beyond acronym disambiguation and study term-level ambiguity in unstructured company text. In doing so, we expect to analyse homonymous (and synonymous) usages of terms of interest and provide an overview of these usages to aid terminological disambiguation.

As highlighted in Section 3, we plan on conducting analyses across domains, iteratively developing the concept mining system to ensure generalisability across domains. The next implementation will address the education domain, focusing on commonly used terms and concepts across departments (domains) within the same organisation. Consequently, we will compare the official definitions of the terms with their actual usage in documentation, thereby analysing diverse interpretations. The interpretations will then help expand the overview of the terms for a data manager and potentially aid disambiguation efforts.

Acknowledgment

The PhD project is supported by the DEMAND project under SVB/ SPR.ALG.02.015 and the HAN Doctoral Grant.

References

Alexopoulos, P. (2020). *Semantic modeling for data: avoiding pitfalls and breaking dilemmas* (First edition). Beijing Boston Farnham Sebastopol Tokyo: O'Reilly.

Castillo, L. F., Raymundo, C., & Mateos, F. D. (2017). Information Architecture Model for Data Governance Initiatives in Peruvian Universities. *Proceedings of the 9th International Conference on Management of Digital EcoSystems*, 104–107. Presented at the Bangkok, Thailand. doi:10.1145/3167020.3167036

Chen, L., Varoquaux, G., & Suchanek, F. M. (2023). GLADIS: A general and large acronym disambiguation benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023).

Henderson, D., Earley, S., & Data Administration Management Association (Eds). (2017). *DAMA-DMBOK: data management body of knowledge* (Second edition). Basking Ridge, New Jersey: Technics Publications.

Hoppenbrouwers, S. J., Proper, H. A., & van der Weide, T. P. (2005). A fundamental view on the process of conceptual modeling. *International Conference on Conceptual Modeling*, 128–143. Springer.

Liu, A. H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., Sablayrolles, A., Héliou, A., You, A., Ehrenberg, A., Lo, A., Eliseev, A., Calvi, A., Sooriyarachchi, A., Bout, B., ... Ramzi, Z. (2026). Ministral 3. arXiv. doi:10.48550/arXiv.2601.08584Liu, B., Guo, W., Niu, D., Wang, C., Xu, S., Lin, J., … Xu, Y. (2019). A User-Centered Concept Mining System for Query and Document Understanding at Tencent. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1831–1841. doi:10.1145/3292500.3330727

Liu, B., Guo, W., Niu, D., Wang, C., Xu, S., Lin, J., … Xu, Y. (2019, July). A User-Centered Concept Mining System for Query and Document Understanding at Tencent. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1831–1841. doi:10.1145/3292500.3330727

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. doi:10.2753/MIS0742-1222240302

Petzold, Z., Roggendorf, M., Rowshankish, K., & Sporleder, C. (2020). Designing data governance that

delivers value. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-digital/ourinsights/designing-data-governance-that-delivers-value#

Ross, R. G. (2020). *Business knowledge blueprints: enabling your data to speak the language of the business* (Second edition). Place of publication not identified: Business Rule Solutions, LLC.

Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, *26*(11), 1297–1304. doi:10.1093/jamia/ocz096

Voskuil, J. (07 2015). Enterprise vocabulary management A lexicographic view. *Wacana*, *16*, 411. doi:10.17510/wacana.v16i2.384

Veyseh, A. P. B., Dernoncourt, F., Tran, Q. H., & Nguyen, T. H. (2020). What does this acronym mean? Introducing a new dataset for acronym identification and disambiguation. In D. Scott, N. Bel, & C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics (pp. 3285–3301). International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.292

Vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to design science research. In *Design science research. Cases* (pp. 1–13). Springer.

Wien, J. J. F., Jansen, J. M. L., Zeeuw, K., & Roosenschoon, O. R. (01 2006). Frameworks and Interactive Tools for Scientific Knowledge Systematization. *MODSIM07 - Land, Water and Environmental Management: Integrated Systems for Sustainability, Proceedings*.

Xu, K., Feng, Y., Li, Q., Dong, Z., & Wei, J. (2025). Survey on terminology extraction from texts. *Journal of Big Data*, *12*(1), 29. doi:10.1186/s40537-025-01077-x

Yang, K.-M., Kim, E.-H., Hwang, S.-H., & Choi, S.-H. (2008). *Fuzzy concept mining based on formal concept analysis. 2*, 279–290.

About the author

I am a 2nd year PhD candidate at Radboud University (RU) and HAN University of Applied Sciences (where I also work as a junior researcher). I have a background in text mining, and my areas of interest include natural language processing, language modelling, information retrieval and data semantics. I am currently supervised by Prof.dr. Arjen de Vries (promoter, RU), Lec.dr. Stijn Hoppenbrouwers (co-promoter, HAN), and Dr. Maya Sappelli (daily supervisor, HAN).

Summary

The research addresses the problem of data ambiguity within organisations, which is known to be a core issue hampering data management. We aim to support data management through concept mining: (semi)-automatically analysing concepts within an organisation's data domain, given a list of terms of interest. In doing so, we intend to analyse perceived meanings of terms and concepts, support the detection of synonyms, homonyms and find or generate suitable definitions. The concept mining system is a "toolkit" for data managers, incorporating text mining techniques. It will be evaluated for efficiency, quality and real-world applicability.