Radboud University Nijmegen

Faculty of Science

Using concept mining for business-oriented data management

An explorative study

Master Thesis Information Sciences

Supervisor:
dr. Stijn Hoppenbrouwers

Author:
Jaimy Göertz

Second reader:
dr. ir. Arjen de Vries

June 22nd 2023

Abstract

This master thesis is about concept mining and its process. Concept mining is a term that is not well-defined by existing literature. However, it shows potential to serve as a valuable addition to the data management field. Concept mining in this thesis will serve as a first major step towards the construction of a data model, which is the end-goal. An extensive literature was done to review the position of concept mining within the field and context. A process was also created with this academic literature using a list of requirements. This thesis gives an overview of the field that concept mining is in. Ten requirements were conceptualized in this thesis. Together they form the process of a future concept mining system. The first four requirements were extensively researched and constructed with existing techniques. These requirements are: identifying types, identifying instances, ranking types and the indication of homonyms and definitions. The other requirements, need more research in the future. Furthermore, the field of concept mining was laid out. This created a conceptual framework for the field, since concept mining is a new term. This includes important conditions like: importance, supervised vs unsupervised learning, evaluation, word types and the business case. Another focus was the implementation and added value of an additional concept mining system. The input and output of a future system were also laid out. This thesis can be used as a starting point for more concept mining research and for constructing data models out of knowledge bases and unstructured data.

Acknowledgements

There are several people I would like to thank for their help during the realization of this master thesis project:

- Stijn Hoppenbrouwers for his guidance in this thesis and for the opportunity of doing this project at HAN University of Applied Sciences.
- The Academy of IT and Media Design at HAN University of Applied Sciences. Especially Maya Sappelli and Bas Hofmans for their advice in the research process.
- Arjen de Vries for his advice about the field and his second readership.

Contents

1 Introduction

Data management is an ever-growing topic of research. Most organizations use internal information systems to map processes and terms that are relevant for that specific organization. This documentation is useful, but it can also be a burden to maintain and to find the necessary information. To lift this documentation to a higher level, some companies have the wish to construct a textual schema, like a knowledge graph, with this data. A term that can be useful for this is: concept mining. However, this term has not been defined properly by existing literature. That is the gap this study will try to fill, so organizations can lift their internal documentation to a higher level.

1.1 Context

The goal that we want to achieve is to construct a knowledge graph or another similar data structure from the unstructured information of an organization. The focus is not on structured data sources like databases. An example of an unstructured data source is a wiki with important concepts. The intention is to improve data management in all types of organizations. This means that organizations of all sizes have to be able to apply the concepts proposed in this research. These organizations can range from small local businesses to big international companies. This thesis will function as the first step to achieving the goal of constructing a knowledge graph or another similar data structure from the unstructured information of an organization. The first step would be to construct a knowledge base out of the given data. This will be the main focus of this specific thesis. However, the context itself also poses challenges related to company size, the amount of data available and the usefulness of a future concept mining project. We think that there is sufficient support in the field for concept mining and concept mining systems.

1.2 Research questions

The research questions have been described as one central question and five sub-questions. The sub-questions serve as a framework for this thesis.

Main question:
How can concept mining be used to define and give meaning to important terms in an unstructured textual source for business-oriented data management?

Sub questions:
1. How do we define "concept mining" based on existing literature?
2. What research is there to find about concept mining and other related concepts in the field?
3. What can concept mining add to the data management field?
4. Which requirements should a concept mining system have?
5. What should the process of concept mining look like?

1.3 Methods

The method of this study will be the design paradigm: "design science", as described by Hevner et al. (2004). It means that artefacts will be produced in an iterative design process of assessing and refining. This framework consists of four phases. The key phases are the "develop" and "justify" phases. Here, you iteratively develop and evaluate your artefacts. These two contribute to the "environment" and "knowledge base" which are the application phases. The artefacts will be developed in redesign cycles. The requirements will be validated by multiple researchers from the field (transferability). This will happen through three interviews. The people that were interviewed were: an enterprise architect, a business consultant information, data & analytics and a data domain manager. The data domain manager is from the company that had the first request for the project this thesis was based on. Quotes and information from these interviews will be used throughout this thesis. The literature review will be documented as systematically as possible with: keywords and an explanation of the process. The literature review will be written by an iterative process of gathering academic knowledge and writing. A further explanation of the methods can be found in chapter two.

1.4 Scope

The emphasis of this research lies on concept mining and the applications of this part of the process. It is important to keep its place in the data model construction process in mind, but the full process will not be implemented. Only the creation of a knowledge base through concept mining. To keep this thesis practically applicable, mathematics will be kept to a minimal. This follows the explorative nature of this study. The research will be based on scientific literature. The scope is fully described by the case and context, as described earlier. To evaluate the results, interviews with experts from the field will be conducted.

1.5 Outline

Chapter two will be a description of the used methods in this research. This report will continue with defining concept mining in chapter three, based on existing literature. This includes defining important concepts with a similar meaning to concept mining to establish its position within the field. Chapter three will also look at the ten requirements for a concept mining system. This is the first step towards creating a process flow. It also looks at the added value of concept mining in general. This chapter could be summarized by the what questions. Like: what is concept mining, and what is the added value? Chapter four will be a very explorative study into the concept mining field that will also look at the input and output of the future system. It also establishes some important topics of research. Then the concept mining process is ready to be described in chapter five. This chapter applies the requirements to a proposed pipeline. It delves deep into everything that is necessary to successfully apply the first four requirements to create a process. The process in combination with the requirements are the base for improving data management in organizations with concept mining. Chapter four and five answer the how question: how to do concept mining? The chapters will be mostly supported by academic literature and examples from our specific context.

2 Methods

This chapter acts as a justification for the research process. Firstly, the usage of the general research method will be explained. Then two important parts of the research process will be explained. This thesis was established at HAN University of Applied Sciences in Nijmegen. The methods are the starting point for this research, since they are necessary to conduct literature research.

2.1 Design science

As explained in the brief methods section in the introduction, the design paradigm: "design science" was used in the design process. In figure 1 the design science paradigm is illustrated. After the figure, the usage of all components will be explained and for every component of the figure a concrete list of the solution will be given.

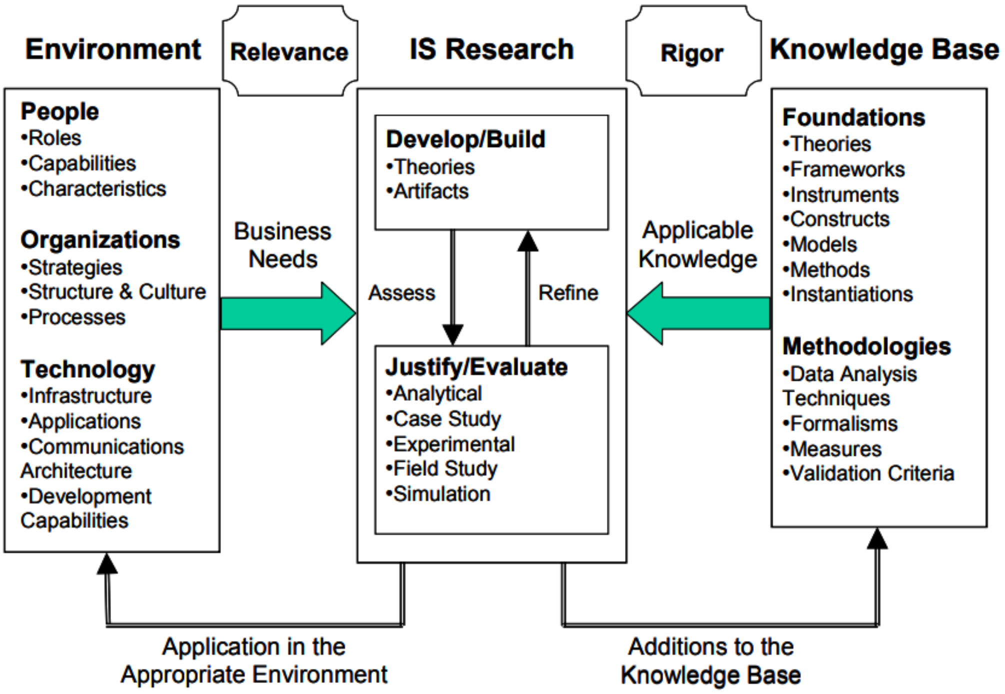


Figure 1: Design science from (Hevner et al., 2004)

2.1.1 IS Research

The develop/build phase is one of the central phases, together with justify/evaluate. Artefacts are central in the develop/build phase. An artefact is: "an object made by humans with the intention that it is used to address a practical problem" (Weigand, Johannesson, & Andersson, 2021). The artefacts in this thesis are the requirements list, process definition, interviews and a small experiment. These artefacts focus on the explorative nature of this study. That is why no actual system will be constructed. The justify/evaluate phases in our case consists of interviews and heaps of academic sources. The main part of the justification consists of academic sources. They are the key evaluation in this thesis. The interviews as described in the next paragraph are an extra justification measure to make sure the plans are actually executable. All solutions can be found in the lists underneath.

Artefacts:

- Process definition
- Interviews
- Requirements list
- Small experiment related to requirement 3

Theories:

- Concept mining definition
- Input, output and conditions
- Process definition
- Added value and implementation (business side)

Justification:

- Interviews
- Combination of existing techniques, so the individual techniques have proved themselves
- Requirements verified by client

2.1.2 Environment (Relevance)

On the environment side, the business needs are the central point. There are three types of people to take into account. The data manager, users of internal documentation and data engineers. The data manager will be the operator of a future system, and the rest are stakeholders and users of the output. The people will be further examined in chapter 4.1. Their point-of-views will always be taken into account within the context. Organizations are also a vital part of our context. All types of organizations will have to be supported, as long as they have unstructured data.

This has close links to the technology that organizations might have. The goal is to keep the future system as accessible as possible. This means that the amount of technology knowledge has to be kept to a minimum. The business needs will also be discussed extensively in chapter 3.3. The environment is crucial for determining technologies that will be used to do concept mining. The environment will often be mentioned when making choices about concept mining. All relevance points can be found in the lists underneath.

People:

- Data manager
- Users of internal (unstructured) documentation
- Data engineers

Organizations (requirements):

- Organizations with unstructured data
- Organizations with the capacity for extensive data management

Technology:

- Infrastructure: add-on to existing data management systems
- Development capabilities: organizations with the capacity for extensive data management

2.1.3 Knowledge base (Rigor)

On the right side of design science, we have the knowledge base. The foundations of this thesis will be mostly based on academic literature, because of its explorative nature. These foundations will be kept as practical as possible, so a concept mining system can be easily constructed. The theories, models and instantiations will come from an academic literature review to create something new. In terms of methodologies, part of the validation will be done by conducting interviews and getting validation from professionals at HAN University of Applied Sciences and a professional from an external company. Because the study is mostly explorative, the validity of this thesis will only be as good as the validity of the academic literature. This is why the validation criteria will have to be top-notch and strict. That is why a specific list of validation criteria for the academic literature review has been constructed. Those validation criteria will be evaluated in the discussion section of this thesis. All academic sources that are used to prove a point have to match these criteria.

There are a few sources that are exceptions. These are sources that are used for definitions (Cambridge Dictionary) and one source purely for illustration (Wikipedia). The academic literature criteria are based on an article by Pradeep & Wijesekera (2019), who made a selection of criteria used by four credible American universities. All lists that correspond to the knowledge base section of figure 1 can also be found in the lists underneath.

Foundations:

- Theories: formed by academic literature.
- Methods: literature review combined with requirements and interviews.

Methodologies (general):

- Interviews will be conducted with a pre-set list of questions tailored to the interviewees.
- Requirements will be constructed with the clients from HAN University of Applied Sciences.
- Chapter 4 evaluated specifically by interviews.
- All major points and subjects in chapters three, four and five have to be backed by at least one credible source that meets the academic literature criteria.

Academic literature criteria:

- Author (reputation on field, affiliation)
- Publisher (reputation, where, medium, format)
- Accuracy (references, citation, peer review, error-free, relevance)
- Currency (published date/ date matters?)
- Coverage (audience, depth of info)
- Point of view (bias? Info/fact/research outcome/analysis?)
- Editions/Revisions (update through time) and title of the journal

2.2 Interviews

To evaluate the choices that were made, three interviews with professionals were conducted. The interviewees have different professions in the field. All three interviews were conducted with pre-defined questions, to get answers to specific questions. These pre-defined questions can be found in appendix A. However, every interview had to be tailor-made during the interview itself, based on the interviewees knowledge and expertise. This resulted in three different interviews focused on different topics.

The interviews were conducted with: an enterprise architect, a data domain manager and a business consultant information, data & analytics. The data domain manager (interview 2) was from an external energy company, and the other two had strong connections with HAN University of Applied Sciences. The goal of these interviews was to validate choices made based on academic literature. The questions were focussed on achieving this goal. The interviews were conducted in an online meeting at about three quarters of the full project time. This way, the interviews could also influence the last quarter of the research. This spawned, for example, a part about the business case that needed further research. The interviews are given in table 1.

Interviewee job	Specialities	Date
Enterprise architect	Technical implementation	18-04-2023
Data domain manager	Context of the project	18-04-2023
Business consultant information, data & analytics	Business cases	19-04-2023

Table 1: Interview overview

The first interview with the enterprise architect was more about the technical side and the implementation of this technical side, because of his technical expertise. It resulted in a conversation about, jargon, metadata, thesauri, international standards and relevance for architects. He works for HAN University of Applied Sciences. He also works on the technical side and on the management side, which makes him a very suitable interviewee for this thesis.

The second interview with the data domain manager resulted in a conversation about the context, since he works for the company that posed the context in the first place. Which helps to clarify the context and the original intent. It also helps with verifying the goals set in this thesis. This helped to better place this thesis within the field, but the main takeaway was the clarification of the context, described in paragraph 1.1. He also mentioned that 100% certainty in the results was a must, which impacted the concept mining process heavily.

The third interview with the business consultant information, data & analytics was focussed on the implementation of the concept mining system and the implications for organizations. So this interview was more focussed on the business end. The main concepts that were posed in this interview were: the business case, data lakes, the support function and the benefits of such a system. This resulted in a further refinement of the business case and focus of the process. He also works for HAN University of Applied Sciences. This interview was again different to the others, because of his experience in business and data specifically.

The interviews were very diverse. Every interview had a different goal and a few different questions. They also complemented each other in terms of the vocal points and experiences of the interviewees. A lot of information was extracted. It resulted in a very positive impact on the rest of this thesis. The validation of information was the main goal, which was eventually met. It even resulted in more research, which the process benefitted from. Various quotes from all interviews can be found throughout the rest of the thesis. A full quotation report can be found in appendix B. The quotes in this appendix have been selected, because of their relevance to the overall thesis. The interviews were coded based on the subject of the quote. After which the most important quotes were selected and used throughout this thesis. In this process the professional tools: Amberscript and Atlas.ti were used for the transcriptions and the coding of the texts.

3 The definition of concept mining

This chapter is concerned with defining the term: "concept mining". It also looks at related concepts that have similar meanings to concept mining to position concept mining within the field and show the added value of this term. Another important part of this chapter are the requirements for a concept mining system.

3.1 Defining concept mining

The first important step is defining the main concept of this thesis: "concept mining". It is not easily defined in the field, since multiple sources use multiple definitions. It is also not a term that is used often. Only ten easily findable academic sources use concept mining in the title. This is very little, especially in such up-and-coming fields as data management or data engineering. These sources along with the insights from the client will be used to construct a new definition that can be used as a starting point for this thesis and future research into concept mining.

3.1.1 Definitions from literature

Four sources give an explicit definition of the term "concept mining". The first is the not always verifiable source, Wikipedia (Wikipedia, n.d.). They describe concept mining as: "an activity that results in the extraction of concepts from artifacts. Solutions to the task typically involve aspects of artificial intelligence and statistics, such as data mining and text mining. Because artifacts are typically a loosely structured sequence of words and other symbols (rather than concepts), the problem is non-trivial, but it can provide powerful insights into the meaning, provenance and similarity of documents." The second definition is given by Liu et al. (2019). This quite long definition is: "Concept Mining is used to search or extract the concepts embedded in the text document. These concepts can be either words or phrases and are totally dependent on the semantic structure of the sentence. When a new text document is introduced to the system, the concept mining can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts. In this way, the similarity measure is used for concept analysis on the sentence, document, and corpus levels."

The third source is a paper by Puri (2011). Their definition is linked to query matching: "Our objective of user-centred concept mining is to derive a word/phrase from a given user query which can best characterize this query and its related click logs at the proper granularity." This definition is less relevant for our context, since query matching is not the goal of this study. The last definition is from the Encyclopedia of information sciences and technology (Khosrow-Pour, 2020).

There, concept mining is defined as: "A kind of task derived from the knowledge discovery, focused to recognize patterns in large collections of unstructured data (that is, linguistic corpora), which contain relevant information associated to concepts. For solving such task, it is employed different methods based on natural language processing (e.g., tokenization, lemmatization, POS tagging, syntactic and semantic analysis, and others)." This definition seems to represent the initial idea of concept mining in this thesis the best.

3.1.2 Final definition

The final definition will be based on the definitions from the previous section. However, these definitions only focus on the functionalities of concept mining. They do not mention the goal of concept mining, which is to collect information to create or validate a domain model/conceptual model (like a knowledge graph for data management). This part should complement other concepts like: conceptual modelling or process mining. This part also differentiates concept mining from those similar concepts, so it has to be incorporated in the final definition.

The definition, that we propose, is the following: "Concept mining is an activity that focusses on the extraction of words or sequences of words from textual corpora (unstructured textual source), with as goal to extract information to aid, complement or guide the creation, evaluation and evolution of conceptual models." This definition could be extended to add some conditions to create the following extended version: "Concept mining is an activity that focusses on the extraction of words or sequences of words from textual corpora (unstructured textual source), with as goal to extract information to aid, complement or guide the creation, evaluation and evolution of conceptual models. Concept mining is at least partially performed by a system, with the goal to obtain important words or a sequence of words in order to identify the most important concepts related to the full textual corpora." These constructed definitions describe our goal as described in the introduction the best. However, the first (shorter) definitions will be the most commonly used definition.

This definition is a combination of previously given definitions from the literature, complemented by the set goal. Liu et al. (2019) determined that the form should be a word or phrase. To encompass more terms, this has been rewritten to: a word or a sequence of words. All sources agree on the fact that concept mining is an extraction process. This was best expressed by Wikipedia (n.d.): "Concept Mining is used to search or extract the concepts embedded in the text document." Since concept mining is a new term, it is justified to use Wikipedia here as a source. So the term extraction part was chosen from this source. According to the Encyclopedia of information sciences and technology, words are extracted from "Linguistic corpora" (Khosrow-Pour, 2020). Which is a good overarching term for unstructured documents, so this was also used in the final definition. However, it has been changes to textual corpora to better emphasize our intentions.

For the extended definition, the fact that a system has to be involved was added to emphasize that this is not a manual process. This was derived from Puri (Puri, 2011). The goal of concept mining was again derived from the Encyclopedia of information sciences and technology (Khosrow-Pour, 2020). When concept mining is mentioned from this point onward, these definitions will be kept in mind.

3.2 Requirements of a concept mining system

In this paragraph, we will try to identify possible requirements for a concept mining system. We will do this as systematically as possible. Every requirement will be described in detail, so they can be used to implement a system together with the process description found in chapter 5. These requirements were formed together with the client of the project. In total, there are ten requirements.

This chapter functions as a list of requirements with explanations for a future concept mining system. The scope of this project asks for a concrete research into the first four requirements. The other six requirements will also be explained, so they can be used for further research in the future. The extensive research into the first four requirements is given in chapter five of this thesis. The first four paragraphs of chapter five correspond with these requirements.

3.2.1 Requirements collection

The requirements, that are described, were established in a requirements collection meeting. The participants in this meeting were three employees from the research group model-based information systems at HAN University of Applied Sciences. They also function as the clients for this thesis project. This meeting acted as a brainstorm session. The results from this session were included in chapter 4. The meeting started with defining the users and stakeholders, because from this point you can determine their needs. This corresponds with the environment side of design science. Then the input and output of a future system were discussed. This can be found in chapter 4. The goal then was to think of steps that could go from the input to the output as efficiently as possible. This resulted in 10 steps or requirements. The fulfilment of the requirements were researched in this thesis according to academic standards. The session ended with the brainstorm of concepts that could be important for this research. The requirements were also used in the interviews that were conducted. This proves that the requirements have been a recurring theme in the research of the concept mining process.

3.2.2 Use of requirements

The requirements given in the next section are prioritized from most important to least important. This does not mean that all steps have to be done in this sequence specifically, but this order makes the most sense if you look at workload and importance. For the optimal system, we propose this order. The optimal system would also require all steps to be successfully completed, but not all steps are necessary to get a functioning system. The feasibility will have to be determined, according to academic research. The goal is to implement all requirements into one full concept mining system. Before formulating requirements, it is good to first look at the users/stakeholders of the system.

For the concept mining system, the users/stakeholders are:
- Data manager/Conceptual modeller (operator of the concept miner)
- Users of internal documentation (users of the output)
- Data managers/Data stewards (users of the output)
- Data engineers (users of the output)

There is only one user type that will operate the system. This user is the data manager or the conceptual modeller. These operators have to have inside knowledge of the documentation and the purpose of a concept mining system in general. The other three types of users benefit from the output of the system. They will not operate the system, but they may assist in determining the framework of concepts. So they can have an assisting role. The providers of the input (data) are stakeholders in the process. This is probably one of the output users.

3.2.3 Requirements list

This paragraph has a list of all ten requirements, together with explanations. These requirements will be mapped one-on-one to the concept mining system. That underlines the importance of this list. It will form a basis for the process and future research into concept mining. Step one, two, three and four will be researched in chapter five, additional explanation can be found there. The rest of the steps will have to be researched in future projects.

Requirement 1: Constructing a list of types from the documents
The goal of this first step is to identify types in the unstructured texts. Types are the category headers in, for example, a table or entity (?, ?). An example of a type is a car brand, which is a category header of cars. Constructing a list of those types is an abstraction problem. In abstraction levels, types are one level above instances (which will feature from step two onwards). Further examples of the types in a table can be found in table 2. The challenge lies in the Named Entity Recognition field. This is also where the solution can probably be found. The output of this step is a list of those types. It needs to be kept in mind that all steps will be executed in English and Dutch.

Requirement 2: List of instance examples per type

This is the last requirement that adds new, important terms to the knowledge base. Instances are the keyphrases/keywords that are one level lower in the abstraction level than types (Wu, Zhang, & Li, 2022). If the type is: car brands, the instances could be: Audi, BMW or Renault. The types will determine what instances are selected in this step. This creates a list with one type and multiple instances. The amount of instances is not known beforehand. This is the second step of the NER process. It could be iteratively done, together with the first two steps. It cannot be executed without step one.

Requirement 3: Ranking types in order of importance

The types that have been identified in the first step are going to be ranked in this step. The goal of this step is to determine the relevance of the identified types. This helps with further steps related to instances. The amount of types is unknown beforehand, which makes ranking even more important. If the amount of types is large, it could be decided that only the first x amount of types will be used. This step helps with determining the important concepts for a conceptual framework. The problem posed in this step can be solved by keyphrase ranking techniques. This will be the field of research for the final implementation of the concept mining system. The ranking could also be based on the instances of the type.

Requirement 4: Indication of potential homonyms and possible indication of definitions

This is the last step that will be fully described in this thesis. It consists of two steps. The first step is the indication of potential homonyms. A homonym is: "a word that sounds the same or is spelled the same as another word, but has a different meaning" (Cambridge-Dictionary, n.d.-a). These words are syntactically very important. If definitions were to be determined, these terms need to be annotated in order to define the right definitions. It is also useful to know what homonyms are, so people within an organization have the same understanding of a term. Homonyms will be annotated in this requirement in the knowledge base. The second part of this step is providing definitions for the types and instances from previous steps. The indication of these definitions will be difficult to research, but it will be crucial for the final execution.

Requirement 5: Indication of possible synonyms

The invert of a homonym is s synonym. A synonym is: "a word or phrase that has the same or nearly the same meaning as another word or phrase in the same language" (Cambridge-Dictionary, n.d.-d). The indication/annotation of synonyms will happen in the same way as with homonyms. They will be annotated in the knowledge base. These steps could be seen as improvement steps. They do not create new terms, but they improve upon them. The indication of synonyms will be based on their definitions. That is why this step happens after the indication of definitions in step four. It is also a possibility that the indication of definitions could also be executed after this step, if the specific use case calls for it.

Requirement 6: Indication of the choice for an entity or an attribute
The problem of identifying entities and attributes is also an abstraction problem, just like in previous requirements. An entity in this case is: "a class of objects about which a database owner has information. For example, a school might include student, course, professor etc. as entities in its information specification" (Johnson, Rosebrugh, & Wood, 2002). An attribute is, for example, students have a name, address, degree etc. (Johnson et al., 2002). The distinction between the two will be made in this step. This is especially relevant for entity-relationship models. Since our model will be a hybrid between a knowledge graph and an entity-relationship model, this step is also relevant for our use case. In terms of abstraction, the entity is one level above the attribute, but there is always a relation between the two.

Requirement 7: Identify possible relations between selected types
This step is possibly the hardest to figure out, together with requirements eight and nine. Now that all types, instances, entities and attributes have been established, it is time to identify the relations between them. Types and instances, and entities and attributes already have a relation, but the relation between all types is not yet defined. This step is most likely an entity linking problem. This is the problem of linking entities to their versions in existing knowledge bases.

Requirement 8: Ranking of relations
Not all relations that have been identified in the previous step are 100% certain. That is why an additional ranking step has to take place. After this ranking, the best relations will form the data structure. Ranking of relations is, just like the previous step, a serious challenge.

Requirement 9: Indication of potential naming of relations
Identifying relations is important, but it is also important to establish the meaning of such relations. This step is concerned with naming these relations to realize an additional step of sense-giving. It will complete the data structure. This specification is important for viewing the relations and drawing conclusions from them. The relevance is especially high if there are relations that can be identified outside the text.

Requirement 10: Generate terminological definitions of types and relations
The definitions that were chosen in previous steps are possibly subjective or subjectively collected. It would be a good optimization to collect terminological definitions to secure objectivity and accurate terminology for the field, that the concept mining system is used in. The crucial part here is the genus differentia, or the dictionary definition of the term and definition. In this step, multiple types and identifiers can be given to the keyphrases.

3.3 The added value of concept mining

This section is concerned with defining the added value of a concept mining system for an organization that has unstructured data. It answers the why question of this thesis of: why would an organization use a concept mining system?

There are many opportunities when working with unstructured data, since most unstructured data has a lot of potential, but this untapped potential is rarely used (Blumberg & Atre, 2003). Many large corporations are now looking into using this untapped potential (Blumberg & Atre, 2003). The goal for businesses is to create more value from existing data. It also helps to clarify important concepts for different people within organizations and that helps with getting everyone on the same page, especially those users that were described in the requirements section, but also other stakeholders can benefit from such a system. Finding new patterns in existing data can create extra useful business value. All interviewees reported that a concept mining system would be highly relevant for their activities within their respected organizations. The data domain manager from the company that the problem originated reported that a high certainty is extremely important. In this use case, a high certainty can probably only occur with some manual intervention. Many data management systems are not perfect, but they work as a support system. A high certainty however is not impossible. This will be kept in mind when developing a new system.

The first point of added value is: creating understanding. This problem was mentioned in the interview with the enterprise architect: "Where it is difficult for the government is that we think we understand each other, but we just don't understand each other. This is how noise is created. In 9 out of 10 cases, I am also asking: what do you mean in this context? Because otherwise I just don't understand you." This could be solved by an unambiguous knowledge base. An unambiguous knowledge base or data structure could also solve the language barrier. Many companies have documents in multiple languages. In most Dutch companies, these languages are Dutch and English or as described by the enterprise architect: "This is the same problem that I encounter in daily practice. The official language at HAN is Dutch, but the world is English, and we are increasingly confronted with the need to translate things". Everyone who has a job in the IT sector can probably confirm this. This struggle is inherent to the IT business. Understanding each other and the language barrier are both social problems within organizations, but there are also technical advantages.

The first technical benefit is finding hidden patterns that were previously unknown. Especially, unstructured data has a lot of potential. This poses the following question from the data domain manager: "We know that a lot of things are hidden in documents. No one can connect them. How could you do something with that?" Concept mining could be the answer to this question.

If you go one step above that, there is the creation of insight in the data that flows through the organization, or as mentioned by the business consultant: "the essence is that an organization or institution gains better insight into which type flows through that organization." The various goals and applications create a technique that has loads of possibilities to add value to an organization, as proved by statements from professionals. This was perfectly summarized in the interview with the business consultant: "If you are able to filter out certain important concepts, important concepts for the organization and also give context to them, so also give a description. Yes, then I see a thousand and one possibilities for that."

3.4 Other relevant concepts

In order to further explore the field of concept mining, alternative concepts will be used to expand the knowledge about the field, because the field is very scattered in terms of concepts. There are three concepts that help answer the what question, that have not been mentioned yet, but are important to consider. This helps with expanding the knowledge from the field and positioning concept mining within that field.

3.4.1 Information extraction

Information extraction is the broadest term in this list. It acts as an umbrella term for multiple concepts related to the extraction of web data. Sarawagi (2008) defines information extraction as: "Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources" (Sarawagi, 2008). It is also used to link structured and unstructured data. Information extraction produces structured data, which is different to information retrieval, which concerns itself with relevant documents instead of entities (Chang, Kayed, Girgis, & Shaalan, 2008). Entities are instances of words in the form of data points. Entities include people, places, organizations etc. Information extraction shares similarities with concept mining in the retrieval of entities from unstructured sources. The difference lies in the scope. Concept mining is more specific on the data and the extraction process. Information extraction is also similar to knowledge extraction, which focusses on the creation of knowledge from structured or unstructured data (Chang et al., 2008).

3.4.2 Entity retrieval

Entity retrieval is very similar to information extraction. It is also an umbrella term for extracting information. It is defined as: "the task of retrieving relevant entities for search queries" (Unbehauen, Hellmann, Auer, & Stadler, 2012). It has become evident that this encompasses query matching. Query matching is less relevant for concept mining, since that stops at finding important terms in an unstructured text. However, these terms match in terms of relevance and relevance criteria. What is interesting about entity retrieval, is that in most sources it encompasses entity linking, like in (Unbehauen et al., 2012) and (Adafre, de Rijke, & Sang, 2007). This is could be a part of concept mining, so that is another match between the two.

3.4.3 Process mining

Process mining aims to: "discover, monitor, and improve real processes by extracting knowledge from event logs readily available in today's information systems" (van der Aalst & Stahl, 2011). There is some serious overlap with concept mining. Concept mining also aims to extract knowledge from some data source. The differentiator here is the type of data. Concept mining does not use event logs, but an unstructured textual source. If in the future, event logs could become a source of data or unstructured data. There are three types of process mining: discovery, conformance and enhancement (van der Aalst, 2012). The form that comes closest to concept mining is discovery. "A discovery technique takes an event log and produces a model without using any a priori information." This is almost identical to what concept mining tries to accomplish. The methods, however, are different because of the different data source and origin.

3.4.4 Conclusion

What is very noticeable about these similar terms, is that they share similarity in the basic problem of identifying important concepts. However, there is not a singular approach to this problem. From the used articles, it seems that there is no easy way to tackle this problem. Most differences lay in the type of text they analyse and the steps that are taken after language processing. Concept mining will use a combination of these techniques and their solutions. It has become evident that concept mining will act as a combination of various other terms.

4 Exploring the concept mining field

This chapter acts as an explorative study into the field of concept mining and the conditions that surround it. This helps with defining a process and the corresponding requirements, which will bring us closer to an implementation of a concept mining system. This chapter is based on explorative academic literature combined to tackle a few problems in the field and our context. It also acts as a preliminary study to chapter five of this thesis, because it looks at the input and output of the full concept mining system. This chapter is also the first step towards the answer to the question: how are we going to do concept mining?

4.1 Input

This paragraph deals with the input for the concept mining system. It encompasses everything that has to be done before the concept mining system can be used in terms of input. The paragraphs about the exploration into a possible concept mining system also has steps that can be performed before using the system as described by the requirements and chapter five.

4.1.1 Preliminaries

Before the concept mining process is described and defined, a starting point needs to be constructed. The "real" starting point is a collection of unstructured data. Unstructured data is plain text in various documents. An example of unstructured data came to light in the interview with the data domain manager. He mentioned: "we know that we have written down a lot of words and concepts in various wikis and documents." But this data does not come out of nowhere. To determine the actual starting point, the position in the environment needs to be considered. The environment in this case is any organization in the field. This means that the starting point can slightly vary depending on the organizations' setup. This is why the starting point of this version of concept mining starts at a data lake. A data lake is a collection of datasets (Nargesian, Zhu, Miller, Pu, & Arocena, 2019). The setup of these data lakes can vary a lot. A data lake is a collection of datasets with multiple tunable variables. These variables can be: the formats, the hosting systems and different formats to describe metadata (Nargesian et al., 2019). These variables can change over time, which makes a data lake a living entity. This can happen autonomously. The data lake makes for flexibility at the start of the process. A data lake can consist of raw, unstructured or multi-structured data (Nargesian et al., 2019). This corresponds with the focus of this study. Data lakes are known for having unrecognized potential. This study aims to exploit parts of that potential. An interview with a business consultant confirms that a data lake is an often used entity and can be used as the starting point for the input in this thesis. The potential data lakes that will be taken into account most likely have tens to hundreds of documents, but not thousands. These data lakes are often used, as confirmed by the business consultant in an interview for this thesis.

The interview with the enterprise architect once again proved that unstructured data and data lakes are highly relevant for our context: "We have an insane amount of documents in total fragmentation throughout our educational institution".

Managing a living entity, that is ever-growing, can be challenging. This creates the pitfall of a data swamp. A data swamp is a data lake from which nobody knows what is in it (Khine & Wang, 2018). This prevents measures against false data or inconsistencies. This also poses security challenges, which can be problematic for organizations and other stakeholders of the data (Khine & Wang, 2018). These challenges can be hard to solve, since they are rapidly expanding, just like data lakes themselves. This study helps with data management, so it can be helpful for managing data lakes and the information in them. The gathering of all documents in a data lake that are relevant is not included in the scope of this study. From now on, if unstructured text is mentioned, this corresponds to the unstructured documents in the data lake.

4.1.2 Pre-processing

The input for this preparation step is all unstructured documents that needs to be included in the concept mining process. The goal of this step is to output raw textual data. Raw textual data is plain text without any markup. While structured data is almost instantly ready for processing, unstructured and semi-structured data need some pre-processing to be ready for actual processing (Maedche & Staab, 2004). This is typically done by natural language processing methods, but as will appear later, most pre-processing will be incorporated in the natural language processing tools that are described in later chapters. This only leaves markup removal to be done before the next steps. It also eliminates the need for stop word removal, stemming and tagging (Zheng, Kang, & Kim, 2009).

Many documents contain textual elements that need to be filtered out before natural language processing. These elements include HTML headers, extra white spaces and extraneous control characters (Palmer, 2010). They are not considered content in data management, that is why they are useless in textual processing. They also do not provide any additional context or meaning to the actual content (Palmer, 2010). If the textual source is from the web, it can include additional elements like: images, advertisements, site navigation links, browser scripts, search engine optimization terms, and other markup (Palmer, 2010). They can also be filtered out without loosing context. The amount of mark up removal that is necessary depends on the specific document. The process of mark up removal can be done independently of the context or corpus and can be applied to all documents or sources within the data lake. However, it needs to be strictly controlled to prevent the removal of actual context, because every single word can be crucial in concept mining.

4.2 Exploration into a future concept mining system

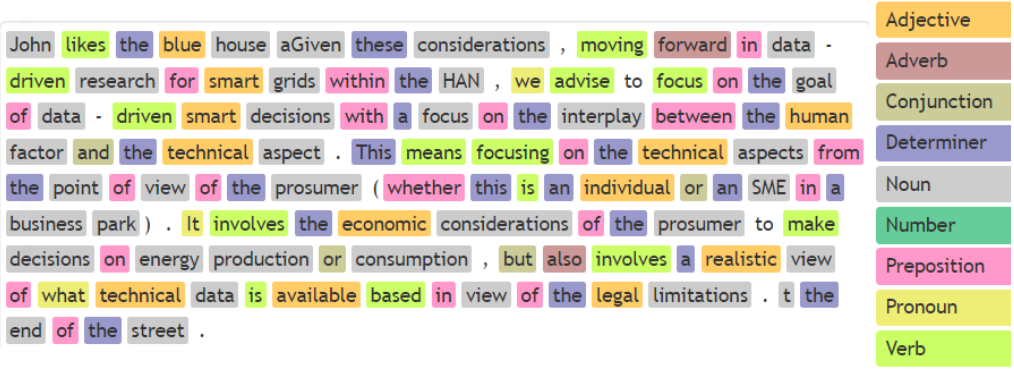
This paragraph describes some conditions that relate to the final system. It poses questions, problems and conditions that need to be solved before the full process is laid out. It also acts as an explorative study into general concept mining.

4.2.1 Keyphrases & word types

The question here is: what words should really be identified? In the field, the most important terms consist of multiple words. For example: concept mining, data management and knowledge graph. These are terms that appear frequently in this study and should be expected to appear frequently in the unstructured data within organizations. A concept mining system should be able to detect terms consisting of one or more words, just like stated in the concept mining definition. In relevant literature, the term used for a sequence of important words is a "keyphrase" (Turney, 2000). The "phrase" part in keyphrase does not mean that a full sentence from capital to period, but one or a few words. The length of a keyphrase depends on the implementation (Siddiqi & Sharan, 2015). An example of the importance of keyphrases is the term "data management." The words "data" and "management" separated have totally different meanings than "data management" in full. So in this study, it should be possible for a concept mining system to extract both keywords and keyphrases depending on the meaning of them. It is also important to keep differentiating the two and to keep in mind that every step should be able to identify both.

Sentences consist of words that can be classified into categories. These word types include: nouns, verbs, adjectives, adverbs etc. (Cambridge-Dictionary, n.d.-e). The vast majority of keyphrases are also noun phrases, and the vast majority of keywords are nouns (Li & Wu, 2006). The importance of nouns cannot be underestimated. In an example study, out of 56 different patterns, 51 contained a noun tag (a part of the term was a noun) (Hulth, 2003). Since nouns are overrepresented in keyphrases, a future system should take this into account.

The process of identifying these different word types is called: part-of-speech tagging. Part of speech tagging is the process of labelling corpus words to their corresponding part-of-speech tags (Asim, Wasim, Khan, Mahmood, & Abbasi, 2018). There are tools that do part-of-speech tagging independently, but a lot of concept mining techniques go one step further to also include part-of-speech tagging. In these independent systems, part-of-speech tagging is viewed as part of the pre-processing process (Asim et al., 2018). For this study, PoS is useful for the third step (ranking types). Systems that do use PoS to extract keyphrases report a 17.6% increase in performance compared to systems that do not use it (Asim et al., 2018). This emphasizes the importance of dissecting corpus' correctly using part-of-speech tagging. PoS is the first step towards identifying important keyphrases in the concept mining process. An example of part-of-speech tagging is given in figure 2.



John likes the blue house aGiven these considerations , moving forward in data - driven research for smart grids within the HAN , we advise to focus on the goal of data - driven smart decisions with a focus on the interplay between the human factor and the technical aspect . This means focusing on the technical aspects from the point of view of the prosumer (whether this is an individual or an SME in a business park) . It involves the economic considerations of the prosumer to make decisions on energy production or consumption , but also involves a realistic view of what technical data is available based in view of the legal limitations . t the end of the street .

Figure 2: Part-of-speech tagging example from (Schramm, n.d.)

4.2.2 Defining importance

Before looking at what important term identification should look like, it is important to look at what importance actually means. The question here is: when is a keyphrase important enough to be included in the important keyphrase list? According to the Oxford Dictionary, a keyword is: "a word or concept that is very important in a particular context" or "a word or phrase that you type on a computer or phone to give an instruction or to search for information about something" (Cambridge-Dictionary, n.d.-b). This means that a keyword has to capture the essence of a document or be an important part of the essence. That is what makes a keyword an actual keyword in a text. This definition also applies to keyphrases, so it is usable within our context. An important takeaway here is that a keyphrase can only be a keyphrase if it related to the context it is in. Therefore, the importance of a keyphrase can only be determined within its context.

Now, how do we measure if a term is important? Frequency is an often used criteria, but this can never be the sole measurement (Mihalcea & Tarau, 2004). Keyphrases can appear only once, but still be relevant enough to be included in the list. Frequency can be part of the determination, but never the only measurement. Frequency determinations have been found to lead to poor results in importance determination (Mihalcea & Tarau, 2004). For frequency measurements, tf-idf is the most used form (Qaiser & Ali, 2018). Tf-idf takes the term frequency and offsets it with the inverse frequency. The inverse document frequency punishes terms that appear more often (Qaiser & Ali, 2018). This is an improvement over just using term frequency, but in our case, the amount of times a term appears should not be leading. An important term can appear only once or very often. It should be taken into account. Tf-idf does not change this perspective.

Another approach could be to focus on jargon. This approach was proposed by an enterprise architect in one of the interviews. You could say that keyphrases mostly have jargon in them. Jargon is defined as "a set of vocabulary items, collocations, or formulaic constructions which are used among people holding similar opinions or sharing a common system of beliefs" (Llevadias Jané, Helmreich, & Farwell, 2005). Since a lot of important terms are specific for a certain domain, trying to extract these terms could be a way of defining important terms. The problem is that another method is needed besides identifying jargon. Also, searching for jargon seems problematic, because it does not extract all possible important terms (Llevadias Jané et al., 2005). It is not worth investing in an approach that is not precise enough or that does not extract every keyphrase necessary. Another problem with this approach is that it needs a substantial amount of training data, which seems to be a major problem for a lot of approaches. This is highly problematic within our context, so this problem will be discussed in the next paragraph.

4.2.3 Supervised vs unsupervised learning

The biggest differentiator in most concept mining techniques is supervised versus unsupervised learning. It is a recurring theme throughout all concept mining related papers. Most machine learning technologies in the concept mining field can be categorized into one of the two. Approaches that do not use machine learning are scarce and pretty much impossible to find. The reason of the choice for one of the two can be crucial in our use case, since we cannot count on any specific amount of data. No assumptions about this can be made. Supervised learning is simply stated: the idea of learning from examples in systems (Learned-Miller, 2014). Supervised learning does this by using labelled data. Labelled data is data that also has the answer to the supervised learning problem. This is how the system learns from examples. The labelling is often done via manual labour. The fact is that controlled learning results in high accuracy (Learned-Miller, 2014). However, the accuracy depends on the amount of training data and the quality of the labelling.

In contrast to supervised learning, there is unsupervised learning. Unsupervised learning does not need a labelled dataset to function. It uses unstructured data to learn from the context (Dayan, Sahani, & Deback, 1999). This also produces a model through an algorithm of interpretation and processing (Celebi & Aydin, 2016). The third variant is semi-supervised learning. It sits right between the two earlier approaches. It uses a small portion of labelled data and the rest is unstructured data (Zhu, 2005). Table 2 shows the fundamental differences between the three approaches.

Difference	Supervised learning	unsupervised learning	semi-supervised learning
Data	Labelled	Unlabelled	Partially labelled
Model	Classification/ Regression	Clustering/ Association	Both approaches
Manual work	High	Low	Medium
Feedback	Yes	No	Partially
Goal	Train model & predict	Find hidden patterns	Both

Table 2: Main differences in three approaches

The fundamental learning problem for this research lies in the labelling of data. Labelling data is a labour-intensive task. The task that would have to be done in the case of concept mining is making importance judgements. This entails the identification of important terms in a text. However, importance is highly subjective. In order to objectify this task, lots of labelling is needed. Most organizations outsource making these relevance judgements to reduce the amount of work. For example, by using crowdsourcing. This is also a much cheaper option. In our case, this will be a major problem, since internal documentation is sensitive information. Pre-trained models could be a solution to this labour problem, but the accuracy in contrast to unsupervised learning is certainly questioned. Another problem in the context is the use of multiple languages. English and Dutch both have to be supported. In supervised learning, two models are necessary. This also requires double the labelling work. This is clearly not an ideal situation. Unsupervised learning only learns from the document, so this method does not require extra work. Unsupervised methods often also support multiple languages. Both supervised, semi-supervised and unsupervised learning have pro's and con's. Which one to use depends on the implementation. All three could be feasible for concept mining, but the amount of data problem needs to be taken into consideration when implementing the process in chapter five.

But what does this mean for this thesis specifically? The decision for supervised-, unsupervised- or semi-supervised learning is based on two criteria. The amount of data available, the amount of manual labour and the amount of certainty. As mentioned by the business consultant: "A high degree of certainty is very important." The amount of certainty that unsupervised learning gives does not seem sufficient for the use case and the final system. Later on, it will show that supervised learning is inevitable for the final system. This brings some constraints in terms of the amount of data necessary and the amount of manual labour. This means that the smallest companies with little data will most likely not be able to easily use the system. To achieve a high level of precision, a serious amount of training data will need to be annotated, but this is a constraint that just needs to be accepted. An unsupervised add-on to the system could be an option for this problem in the future. Semi-supervised learning techniques are not developed well enough to be useful for our use case.

For now, we will focus on using supervised learning for the big tasks, with some support from unsupervised techniques.

4.2.4 Evaluation

Evaluation is important to determine the validity of a system. It is also useful in this case to measure the performance of the proposed system. The most basic form of evaluation is the counting of the amount of right identified keyphrases, like in the Topicrank paper by (2013). To do this in a more systematic way, precision, recall and f-measure are often used in the field. Precision is the number of rightly identified keyphrases divided by the total amount of keyphrases in the document pool (Buckland & Gey, 1994). It is calculated by the following formula:

$$Precision = \frac{\text{\# of retrieved documents that are relevant}}{\text{\# of retrieved documents}}$$

Recall is the amount of retrieved keyphrases divided by the total amount of keyphrases (Buckland & Gey, 1994). It is calculated by:

$$Recall = \frac{\text{\# of relevant documents that are retrieved}}{\text{\# of relevant documents}}$$

To combine these two metrics in the right way, the f-measure is used. This is the trade-off between precision and recall. That is why the f-measure is also called the "harmonic mean of recall and precision" (Ye, Chai, Lee, & Chieu, 2012). The f-measure is calculated by precision times recall multiplied by two. Then divide this by precision plus recall (Ye et al., 2012). This will give you a good indication of the methods' performance. It is also used to compare methods. The formula for F-measure is:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Many papers used in this study also use f-measure to draw conclusions about performance. All steps of the proposed concept mining system should be evaluated separately for the best result. Every step is fundamentally different, so every step calls for separate evaluation metrics and steps.

4.3 Output

This paragraph talks about the output of the whole concept mining system and step one until four of the requirements. The goal is to produce a model close or inspired by a knowledge graph or entity relationship model. An example of a knowledge graph can be found in figure 3. What steps need to be taken to get to such model has been mentioned by the previous chapter and the next chapter. The final product will be a basic form of data model. The research into requirement four until ten will dictate the exact look of this future model.

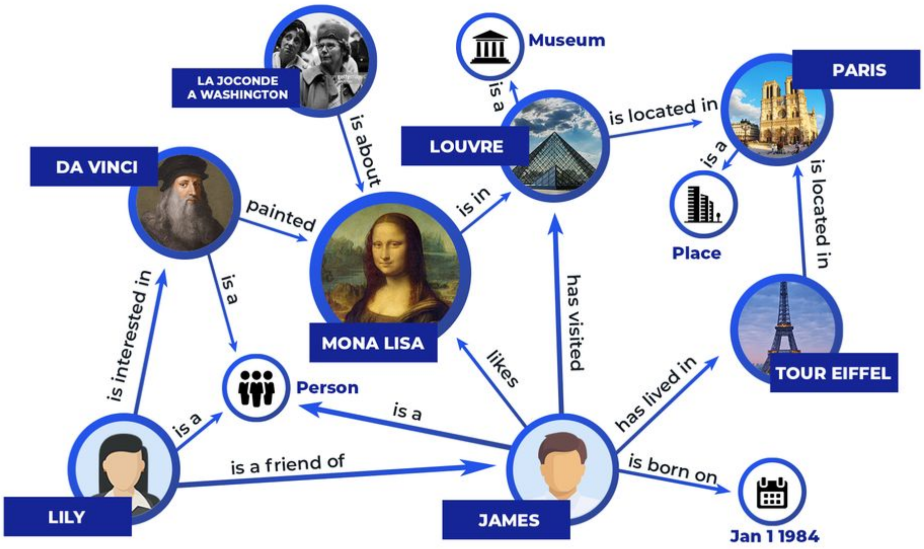


Figure 3: Knowledge graph example from (*Knowledge graph Atlassian.com*, 2023)

A knowledge graph is defined as: "a graph-structured knowledge base that stores knowledge in the form of the relation between entities" (Kertkeidkachorn & Ichise, 2017). Their purpose is to give real life context to abstract concepts (Lin, Zhao, Huang, Liu, & Pu, 2021). They have recently become very influential in data processing. The question is not if a knowledge graph can be made, but if a domain-specific knowledge graph can be made by an automatic system. This poses a serious challenge, not just for this study, but for the field in general. Mostly, because converting unstructured data to useful data for a knowledge graph has been described as "not straightforward" (Kertkeidkachorn & Ichise, 2017). If we look at the output from the elaborated steps in this thesis (one to four), the output will look different. The output will be a knowledge base with additional enhancements. The visualization of this knowledge base could be done in a table or list with annotated homonyms. An example of what this output could look like is given in table 2.

1. Software	2. Computing	3. People in organizations
Word	Client	Manager
Google Chrome	Desktop	Developer
Recycle bin	Motherboard	Client

Table 3: Proposed output step 1 to 4

Here, the instances that are homonyms are given in red. The types are given in blue, and the rest of the words/keyphrases are the instances. The types are ranked from the most important on the left until the least important on the right. The definitions of these terms could be added to the table, or they could be kept in a different list that is linked to this table.

4.4 Implementation

The implementation of a new system within an organization is an important part of software engineering. This was recognized in the interview with the business consultant: "How do you ensure that people actually use the information it generates on top of their already busy tasks? And how do you ensure that the people on the work floor, in this case teachers, researchers etc. really appreciate the value of this and act accordingly". The expectation is that a new concept mining system has a low burden on an organization, since it will only be operated by one person, one group or one division of the total organization. This in contrast to the total amount of users, which will be significantly higher. The balance of work and benefit will therefore in all likelihood be positive. This was also mentioned in section 3.3 of this thesis.

The demand could be high, since the overall interest in the data management field has skyrocketed in recent years. This could help make the organization future-proof. However, a concept mining will be an extra system on top of an already existing data architecture, which could make it hard to sell within an organization. This blends into the question: how do you make sure a concept mining system will actually be used in an organization? This question poses a challenge that could be answered with the help of issue selling.

Issue selling tackles the problem of changing organizational behaviour by peripheral experts (Dibenigno, 2019). Often people lower within the organizational structure will be the ones actually addressing problems and necessities. This could also involve concept mining. To be more accurate, issue selling has been defined as: "The process by which individuals affect others' attention to and understanding of the events, developments and trends that have implications for organizational performance" (Dutton & Ashford, 1993). The chance of successful issue selling depends on the moves of the issue seller, the collective interaction and the relational history with the recipient (Lauche & Erez, 2023).

Lauche and Erez (2023) have summarized the field of issue selling into three forms. The first form is: productive confrontation. This is an issue framed as an opportunity which ultimately invokes radical change. This form is characterized by:
- Explicit about strategic ambition, potential for improvement and resource issues.
- Discontent and personal involvement as a motivational force to fight for a solution, sense of urgency.
- Entire team engaged in a collective diagnostic.
- Recipients contribute by supporting and acknowledging.

The outcome is resolution (catharsis). This is the form that the issue seller should strive for. The second approach is: avoiding escalation (Lauche & Erez, 2023). This form is framed as a technical issue while avoiding upsetting those in power. This form is characterized by:
- Relational history: high power distance and underlying conflicts
- Rationality tactics - normally successful, here too implicit
- Foreshadowing, implicit fighting and bargaining for resources
- Recipients contribute by setting up dilemmas and deferring decisions

This approach will ultimately lead to the downfall of an issue. The third form is: collective moaning (Lauche & Erez, 2023). This issue selling form views issue selling as issues framed as avoiding errors from the past. This form is characterized by:
- Relational history: conflicts and concerns not taken seriously.
- Indirect alluding and appeasing.
- Concerns are aired but not acted upon.
- Recipients contribute by responding with indifference and leaving open.
The forms of issue selling are summarized in figure 4.

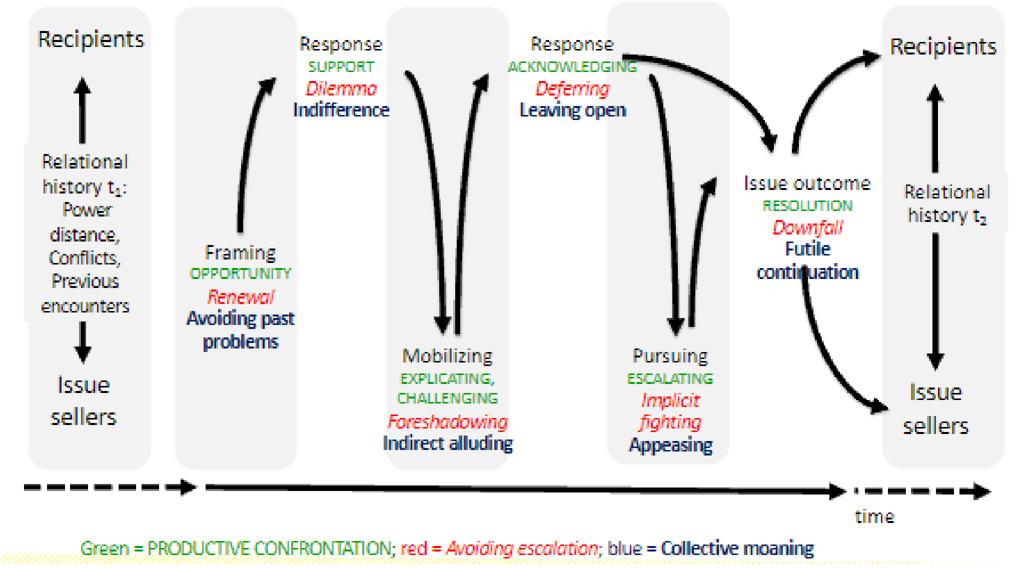


Figure 4: Issue selling approaches summarised

Issue selling is an important part of making the implementation of a new system actually work. What is important to take away about this subject is that many actors in an organization can initiate change. Peripheral experts, for example, can learn to build and maintain relationships with those in power. For these issue sellers, it is important to understand the specific genre of issue selling practices in the organization. The people in power should listen to these weak signals, because they serve as a precursor for change.

4.5 Summary

This chapter established a starting point. That starting point is unstructured data in a data lake. This could also offer a solution to a data swamp. Unstructured data needs pre-processing to get plain text out of it, including part-of-speech tagging. The extracted terms should not be just keywords, but should also include keyphrases. Such keyphrase is important when it resembles a part of the full document, just the amount of occurrences is not enough to define importance. To do the concept mining tasks, unsupervised and supervised learning are both feasible, even though they come with various conditions that need to be kept in mind. Supervised learning seams inevitable for the tasks we want our system to perform. All steps of the concept mining system should be separately evaluated using evaluation metrics as well. The output of the full system will be a data model like a knowledge graph or entity relationship model, while the first four requirements will produce a knowledge base with (ranked) types, instances and homonyms. To successfully implement a new system within an organization, a peripheral expert could use issue selling to convince people in higher layers of the organization to invest in the new system.

5 The concept mining process

This chapter describes the first part of the concept mining process. These first four steps can be executed in sequence. The last six steps will not be fully described, but some guidelines will be given in paragraph five. These steps are based on the requirements list of chapter three. The concept mining system will be a system for the support of the people involved with data within an organization. The goal is to improve data management as automatically as possible. Another goal is to find hidden patterns in data that are often overlooked and underused. The process can be implemented into one concept mining system that performs all steps. The fact that all steps run automatically does not rule out the chance of manual intervention. Systems are not perfect yet, which creates room for manual improvement of the proposed knowledge base. Figure 5 gives the process of the first four steps in the form of a diagram.

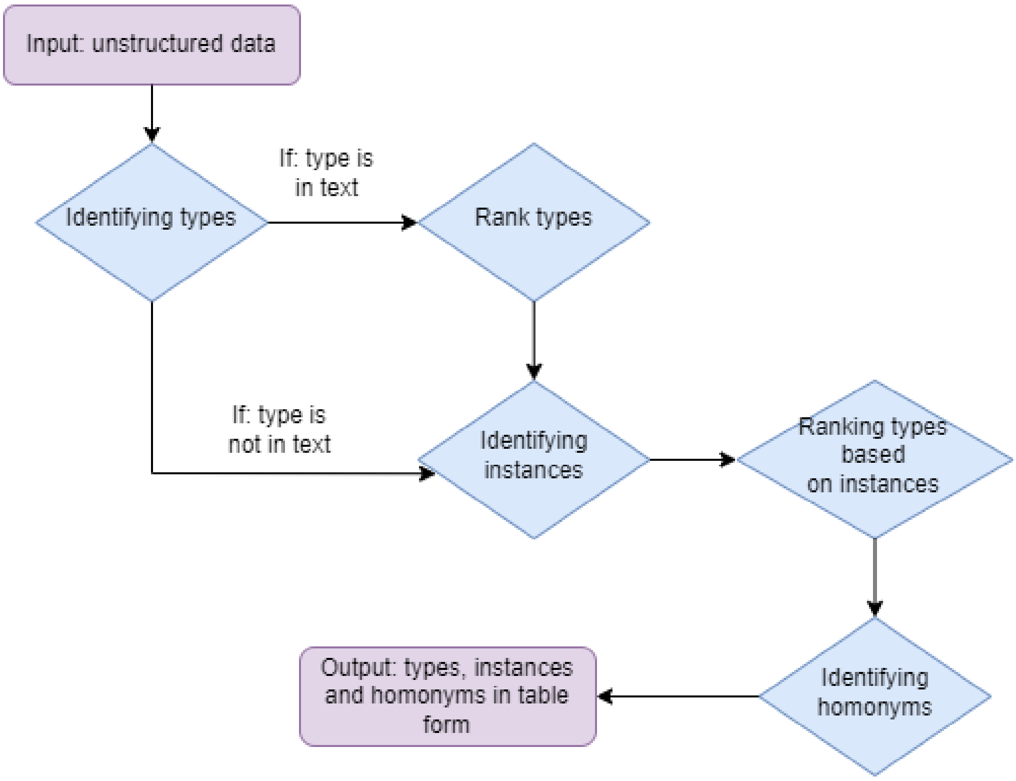


Figure 5: Process of the first four steps

5.1 Identifying types

The first step is the identification of types from the unstructured text. This step will be executed after the preparation mentioned in chapter four. The unstructured text from the data lake is the input for this step, and the output is a list of types. Extracting types could be seen as a text to table problem (Wu et al., 2022). In a table, you have types and instances. The category in the header of the table is called the type, and the population of the table are instances (Wu et al., 2022). An example of a type is the types: car brand. Examples of instances that correspond with this type are: Audi, BMW, Tesla and Renault. These types should be automatically identified in this step of the process.

5.1.1 Named Entity Recognition to identify types

The most useful field to look at for this step is the Named Entity Recognition (NER) field. Named entity recognition is concerned with identifying expressions that refer to people places, organizations and companies (Mansouri, Affendey, & Mamat, 2008). Named entity recognition (NER) is defined as: "a sub-problem of information extraction and involves processing structured and unstructured documents and identifying expressions that refer to peoples, places, organizations and companies" (Mansouri et al., 2008). This is an interesting concept, since it also uses categories to give meaning to important terms. The extraction of categories is also what we are trying to accomplish in this step. Since types are categories of instances. The major problem is that we are looking for other types than NER normally does. Some NER implementations use more categories than in usual approaches, but they never exceed fifteen to twenty categories. In our system, we are looking for specific categories (types) that can be identified from the text itself. Some sources call this domain-specific named entity recognition. NER techniques will be used, but in this use case, specific new types need to be constructed to cover the domain of the specific implementation.

5.1.2 Domain-specific NER Solutions

An often used technique in domain-specific NER is "BIO-tagging" or the "BIO Tagging Schema". It is used to detect entities (sequences) at sentence level (Fang et al., 2021). The beginning of an entity is tagged with a "B". The other token of the entity it tagged with "I" and the non-entity tokens are tagged by an "O". An often used technique in this very specific use case is: BERT. However, most approaches do not fully tackle our problem. Only two academic sources were found that actually tackle our problem.

Two solutions will firstly be highlighted before the actual chosen solution is given. The first solution is called HAMNER (S. Liu, Sun, Li, Wang, & Zhao, n.d.). It is easily usable, since it does not use annotated data, which limits the amount of manual labour. It is a difficult system, as these type-finding methods all are.

Its functionality is described as: "Given a sentence, HAMNER uses the trained model to predict types of all the possible spans subject to the pre-defined maximum number of words, and uses a dynamic programming based inference algorithm to select the most proper boundaries and types of entity mentions while suppressing overlapping and spurious entity mentions" (S. Liu et al., n.d.). The input is a dictionary of entities. This is a serious problem for our context, since constructing a dictionary or knowledge base is the sole purpose of the first four steps of the concept mining system. Another solution is called: "TEBNER" (Fang et al., 2021). This method also acknowledges the problem of manual labour, so this method also does not need large amount of annotated data, but it also needs a dictionary. This problem seems inescapable according to this academic literature.

The need for a corpus and dictionary to complete the identifying types step creates a paradox. We need a dictionary to identify types, but we need this step to complete creating a knowledge base that can act as a dictionary. To comprise a process that is capable of doing domain-specific NER, we need to look at practical implementations and use existing tools. A concrete application of custom NER comes from Microsoft Azure (*Custom named entity recognition - azure cognitive services*, n.d.). They have developed a tool for creating new labels for NER. Azures development lifecycle can be observed in figure 6.

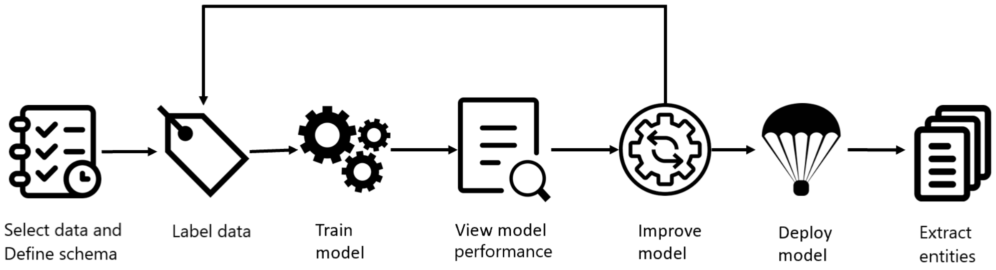


Figure 6: Custom NER according to Azure

From this, we can conclude that annotated data is needed to complete this step of the process. The annotated data that is needed, is for the identification of the custom types. We have not been able to find an automatic solution for the annotation of this data. However, Microsoft Azure offers a solution to handle this annotation. There are other tools available like the SpaCy NER annotation tool (*Custom named entity recognition - azure cognitive services*, n.d.). SpaCy is a python library that can also be used to build custom NER solutions (*spacy.io*, n.d.). There have been non-academic implementations of custom NER with the SpaCy library (Moonat, 2022).

We can conclude that type identification is an NER problem that does have an academic basis, but these solutions do not cater to our specific necessities. There are implementations available, however they require manual labour to annotate data in order to train a model to use custom NER that will also be beneficial for identifying instances. This will be a typical AI/machine learning process. Of training and improving a model for a specific domain before it is deployed. Azure and SpaCy are possible implementations of this process. The output of this step is a list of types and a NER model.

5.2 Identifying instances

The identification of instances is also a named entity recognition problem, just like the identification of types. When step one is performed correctly and the model is trained sufficiently, it should be possible to use that model for this step. When all domain-specific labels are known, identifying instances should be a lot easier than identifying types. The NER model can be used in for example SpaCy or Azure to identify instances. When these instances are identified, they should be grouped according to their corresponding types. The input of this step is the NER model and the identified types. The output of this step are lists of instances that correspond to the correct types. This could be formatted as a table.

5.3 Ranking types

The goal of the ranking types requirement is to determine what types are worth identifying instances for. This will determine the relevance of types and its instances for further steps. The input will be a list of types (and the unstructured text). The output will be a list of ranked types. It does not matter in this step that the keyphrases that we are ranking are types, since we just want to produce relevance judgement for these terms. This step is also iterative with the next step of identifying instances, since types with little or no instance are maybe also not relevant. This step will not be concerned with NER, since NER does not offer a solution in this field.

5.3.1 Possible implementations

This paragraph is concerned with different ranking techniques. For these techniques, we will look at the keyword extraction field, since that field also includes a ranking problem. This field does however offer solutions for this problem. An implementation that does offer a solution is KeyBERT. KeyBERT is an unsupervised form of the supervised method BERT that is specialized in extracting keywords and keyphrases using pre-trained models (Glazkova & Morozov, 2022). It also has a multilingual approach, which makes it automatically suitable to our context, since no models have to be trained to achieve the result needed. KeyBERT normally extracts keyphrases, but it also offers a functionality to give a list of words that you want relevance judgements for based on the unstructured texts they are in. This is a supervised method for obtaining relevance judgements based on the context the types are in.

This however does not keep in mind that we are using types. If this has to be taken into account, no ready-made solution is available for ranking specific types. Supervised keyphrase extraction ranking is the next best thing.

The second approach are the graph-based methods. The idea that graph-based methods are built upon is that of voting or recommendation (Mihalcea & Tarau, 2004). They use vector representation to model words. When one vertex links to another one, it is basically casting a vote for that other vertex. More votes equal a higher importance. Hence, the score associated with a vertex is determined based on the votes that are cast for it. The vertex that casts the vote is also important (Mihalcea & Tarau, 2004). Graph-based models are unsupervised methods visualized as graphs. Figure 7 shows a visualization of a query-based graph model in which the nodes represent sentences and the numbers the similarity between the sentences.

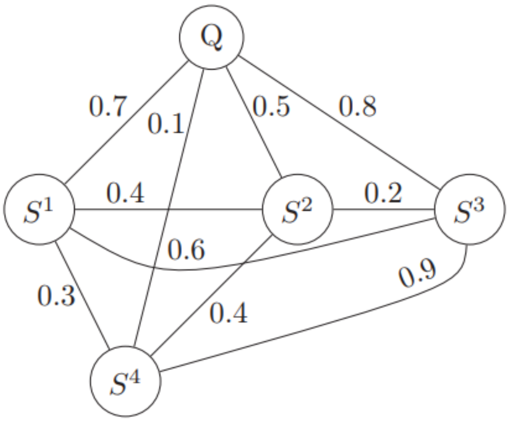


Figure 7: Graph-based model example from (Alhoshan & Altwaijry, 2022)

The most basic and widely used form of a graph-based model is TextRank. TextRank is a graph-based model specifically developed for keyword extraction (Mihalcea & Tarau, 2004). It is the text variant of the well known PageRank. This approach is very much usable within the context as described by Mihalcea and Tarau (2004): "it does not require deep linguistic knowledge, nor domain or language specific annotated corpora, which makes it highly portable to other domains, genres, or language." The third approach is topic modelling. Topic modelling views a document as a mixture of different topics or themes (Tong & Zhang, 2016). A topic is a hidden, to be estimated relation that links words in a vocabulary and the occurrences in documents (Tong & Zhang, 2016). The goal is to identify hidden themes in collections and to annotate these themes. Every word in the vocabulary is categorized into one of these topics. This is a way of viewing data from a different perspective (Tong & Zhang, 2016). The result of a topic modelling system are lists of similar words grouped according to similarity. TopicRank is another graph-based approach that also uses the concept of topic modelling.

TopicRank is a further developed version of TextRank. It is the best of both previously explained approaches. The initial paper about TopicRank defines the approach as: "an unsupervised method that aims to extract keyphrases from the most important topics of a document" (Bougouin et al., 2013). The steps that TopicRank takes are illustrated in figure 8. Here it shows that the graph-based ranking step of TopicRank is very much usable. It does not have an implementation in which a pre-defined list of words can be used as input, but it can be easily constructed by using the KeyBERT implementation of this part.

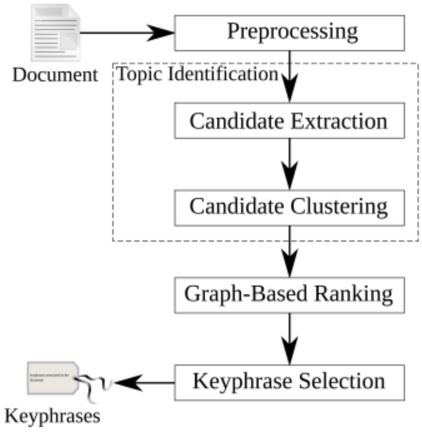


Figure 8: The TopicRank process as described by Bougouin, Boudin and Daille (2013)

Compared to TextRank, TopicRank has a higher F-measure, because TopicRank actually uses multiple rounds of TextRank. This means that it is more advanced than TextRank, which makes TopicRank automatically more effective. Also, the topic categorizing can help with future challenges.

The third graph-based method is CollabRank. CollabRank has the same characteristics as TopicRank and TextRank in that it also extracts keyphrases using a graph-based model (Wan & Xiao, 2008). The big differentiator here is that CollabRank uses all possible documents to identify keyphrases, where the other methods look at individual documents. In terms of f-measure, CollabRank scores higher than TextRank (Wan & Xiao, 2008). It has not been compared directly with TopicRank in terms of precision and recall. Since this study is to be kept practical, we have to look at actual implementations. TopicRank is used more often, since many implementations have been made ever since the original paper came out. CollabRank seems less popular within the field. TopicRank has the most advantages together with KeyBert in our evaluation.

The last question in this paragraph is: should TopicRank or KeyBERT be used? To determine this, a small experiment has been conducted with small example texts. The texts used were domain-specific texts about smart-grids in Dutch and English. This experiments purpose is mainly to determine the most effective method out of the two, not if the methods are precise enough to implement. That has been proven already. Firstly, we asked TopicRank and KeyBERT to extract the top 20 and top 40 most important keyphrases in the English text. This is immediately the first footnote. The original TopicRank and KeyBERT libraries need to have a top x amount of results parameter that needs to be given by the user. This makes the precision and recall the same, since the number of retrieved elements is the same as the number of relevant elements in the text. The table below has the results of the English test.

Method	Results @20	f-measure @20	Results @40	f-measure @40
TopicRank	8/20	0.4	15/40	0.25
KeyBERT	12/20	0.6	16/40	0.4

Table 4: Results English TopicRank vs KeyBERT experiment

The results are the rightly guessed keyphrases when asked to retrieve the top 20 and top 40 keyphrases. When asked to retrieve 40, both methods have similar results, but when asked to retrieve the top 20, KeyBert shows significantly better results. The f-measure from KeyBERT at 40 results is the same as the f-measure at 20 results from TopicRank. This shows that KeyBERT is the more precise method of the two in this case. What further showed in the keyphrase list from the two methods, is that the KeyBERT results were better formatted. TopicRank showed results like: transition energy instead of energy transition, grids smart instead of smart grids and efficient micro utilities questions instead of micro utilities. These were results when stemming was removed. With stemming on, the final letters of the keyphrases were gone, which made them hard to read. The KeyBERT results were way better formatted. Next, we have asked TopicRank and KeyBERT to extract the top 27 and top 15 keyphrases from a Dutch text. The table below has the results of this Dutch test.

Method	Results @15	f-measure @15	Results @27	f-measure @27
TopicRank	5/27	0.56	7/27	0.52
KeyBERT	5/27	0.56	7/27	0.52

Table 5: Results Dutch TopicRank vs KeyBERT experiment

The results from both methods were identical in the Dutch text. The real difference lies with the English texts. Here, in our opinion, KeyBERT is more precise. It has become apparent that KeyBERT produces the best results in this small experiment. The Dutch test showed similar results for both methods. So this test can be left out to determine the takeaways. The English test showed better results for KeyBERT. Also, the formatting of KeyBERT showed superior. KeyBERT was also easier to used and set-up. The current sentence transformers for English and Dutch are good enough. But it is not ruled out that additional models can be trained in the future that are for example domain specific. This means that KeyBERT and BERT in general are flexible enough for future endeavours.

The implementations thus far will only work if the type itself is findable in the text. But with the implementation of step one, it is possible that a type will not occur in the text, only its instances. That is why we propose an iterative process. Where first it is looked up if a type is in the text and if this is the case, a corresponding ranking is given. If this is not the case, the ranking of the type will be calculated by taking the average ranking of all its instances. The average ranking of types will be calculated for all types to keep the ranking fair. This makes the ranking step a two-part step. These steps can also be found in the full process of figure 5. That is why this step will be done after the identification of the instances.

5.4 Indication of potential homonyms & indication of definitions

This paragraph covers the fourth requirement: indication of potential homonyms & indication of definitions. The input of this step are all types and instances found, and the output are all types and instances with annotated possible homonyms.

5.4.1 Homonym detection

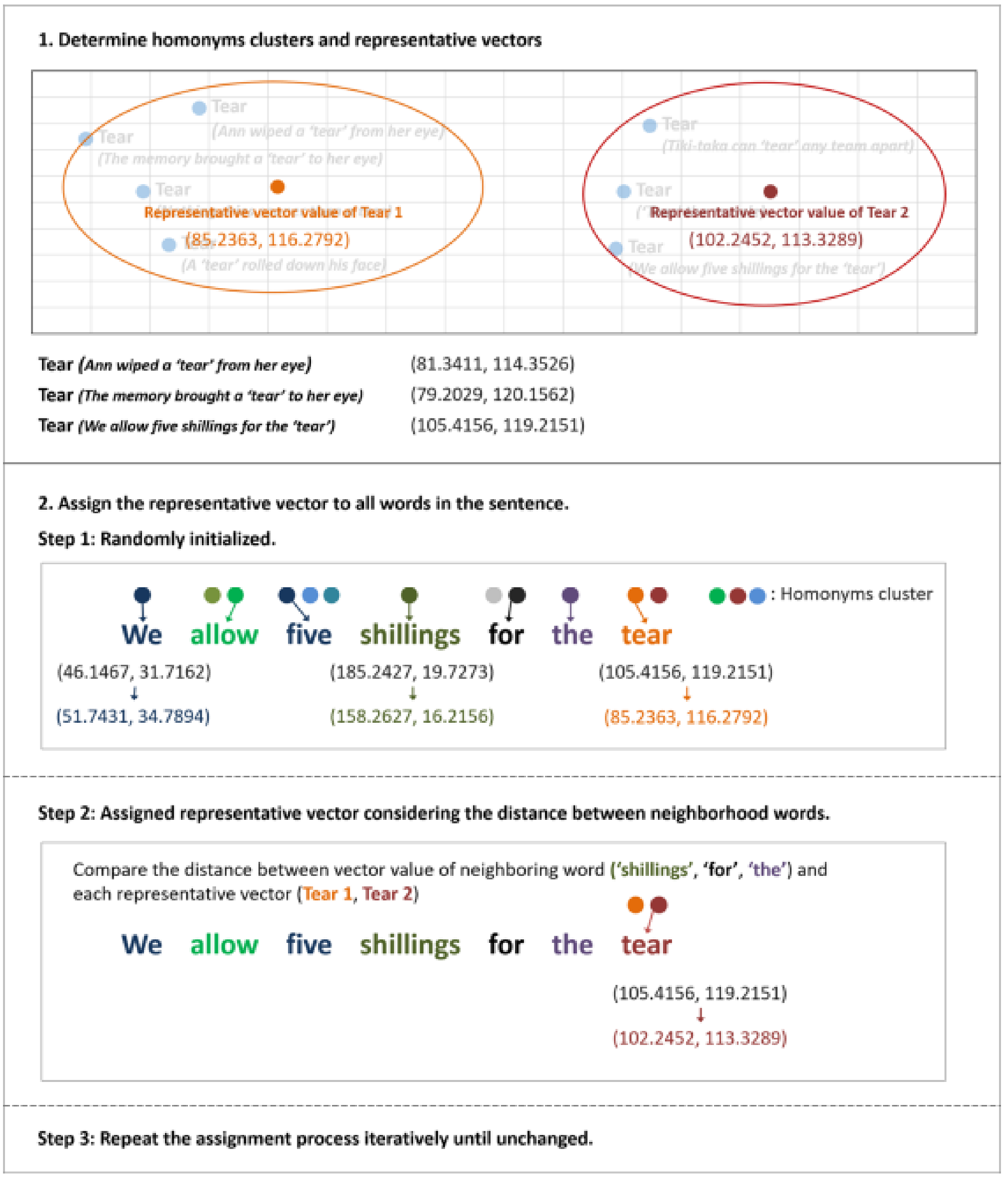
According to the Oxford Dictionary (n.d.-a), a homonym is: "a word that sounds the same or is spelled the same as another word, but has a different meaning." Polysemy is: "the fact of having more than one meaning" (Cambridge-Dictionary, n.d.-c). An example of a homonym is the word: "fly". It can mean to fly, in for example an aeroplane, or it can refer to the animal. Homonyms and polysemy refer to the same problem in this case. From now on, homonym detection will be used to refer to this problem. If a word has multiple meanings, it means that in a list of important terms one term might appear in different contexts. This is a major problem, since in knowledge graphs these terms have to be differentiated. A possible context was posed in the interview with the enterprise architect: "We have had a similar discussion. Participant in education, participant in research, participant in regulation. We have chosen to keep participant as an entity to the education domain en that with research, we are looking at a research subject. That is how we view homonyms." The goal of this step is to annotate homonyms to be able to make a more precise list of terms and maybe in the future their definitions. This will help in using the word in the right context within an organization.

Homonyms are super relevant for businesses in data management. This was confirmed in the interview with the business consultant: "There are a lot of homonyms and synonyms in the important terms, but how can we get to an unambiguous language that we speak reciprocally."

Two important concepts here are word embeddings and homonym detection. Word embeddings are defined as: "low-dimensional vector representations of vocabulary terms that capture the semantic similarity between them" (Zamani & Croft, 2016). Homonyms are identical words with different meanings. The first step towards handling homonyms has already been taken. Keyphrases help with problems related to synonymy and polysemy (Siddiqi & Sharan, 2015). Even though this is the first step, we are not nearly there yet. Another important thing to consider is that double words from different documents need to be kept in until after this step. Different words might have different meanings in multiple documents.

Academic literature described two approaches that do a similar thing with different methods. Homonym identification is only possible with supervised machine learning, as described by methods proposed by Saha (2021) and Lee (2021). This might deviate from approaches in earlier steps of the process. That means that this step might be harder and more time-intensive to do. Unfortunately, the consequence is that this step will not be possible in every use case in which the concept mining process will be used, since the amount of data is a major influence here. This was the same with the first step, so at this point it was accepted that manual model training is going to be an important part of the process. The concept mining process can be executed without this step, but it might be less precise, and it poses an extra challenge with knowledge graph construction. Word2Vec and Glove are two techniques that are often used in the word embeddings field (Lee, 2021). In Word2Vec and Glove, homonyms are embedded in the same vector, even when the meanings differ (Lee, 2021). This proves that these techniques are not sufficient for the homonym detection problem. The only possible solution is to do the annotation of homonyms fully manually until further research proves the feasibility of this task.

BERT is an often mentioned method. However, in a study by Saha (2021) it is stated that the accuracy is too low to actually be useful in this case. BERT and clustering methods in general are apparently not sufficient in their current state for homonym detection. This does not rule out that future research will reach a breakthrough in this field. This illustrates that just using word embeddings for this use case is not sufficient. A more realistic approach is ELMo. ELMo is a contextual word embedding approach (2021). It detects homonyms based on the context they are in. It assumes that if the same word is in a similar context, it has the same meaning (2021). Lee (2021) proposes a method that uses ELMo and vectors. This approach is best explained by a step-by-step guide given in figure 9.



1. Determine homonyms clusters and representative vectors

Tear
(Ann wiped a 'tear' from her eye)
Tear
(The memory brought a 'tear' to her eye)
Tear
Representative vector value of Tear 1
(85.2363, 116.2792)
(A 'tear' rolled down his face)
Tear
(Tiki-taka can 'tear' any team apart)
Tear
Representative vector value of Tear 2
Tear
(102.2452, 113.3289)
(We allow five shillings for the 'tear')

Tear *(Ann wiped a 'tear' from her eye)*	(81.3411, 114.3526)
Tear *(The memory brought a 'tear' to her eye)*	(79.2029, 120.1562)
Tear *(We allow five shillings for the 'tear')*	(105.4156, 119.2151)

2. Assign the representative vector to all words in the sentence.

Step 1: Randomly initialized.

We allow five shillings for the tear
● ● ● : Homonyms cluster
(46.1467, 31.7162)
(51.7431, 34.7894)
(185.2427, 19.7273)
(158.2627, 16.2156)
(105.4156, 119.2151)
(85.2363, 116.2792)

Step 2: Assigned representative vector considering the distance between neighborhood words.

Compare the distance between vector value of neighboring word ('shillings', 'for', 'the') and each representative vector (Tear 1, Tear 2)

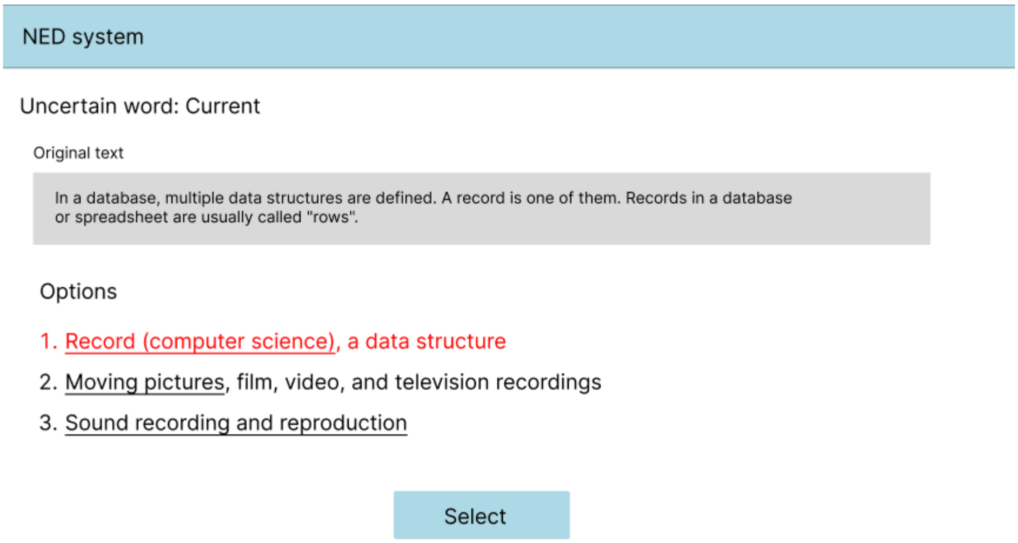
Step 3: Repeat the assignment process iteratively until unchanged.

Figure 9: Homonym detection described by Lee (2021)

This approach returns positive results in terms of improving F-measure. The main problem here is that additional research is necessary to tailor this method to our specific context. The main approach that shows promise is: Named entity disambiguation (NED), which will be discussed in the next paragraph.

5.4.2 Finding definitions

Named entity disambiguation is a practice focussed on linking a reference within a unit of text to its corresponding entity in some knowledge base. In a lot of NED approaches, Wikipedia is used to establish what words are similar. Wikipedia has various ways of giving meaning to homonyms like: redirect pages, disambiguation pages, hyperlinks and categories (Bunescu & Pasca, 2006). Other existing knowledge bases could also be used, according to the correct domain. This approach seems feasible, and it should be worth investing in this. The only problem here are the definitions of the terms from the domain-specific texts. In order to accurately do this, the definition of a term needs to be established, and it needs to be established that this word is the same as one of the definitions from Wikipedia. To implement NED into a concept mining system, we propose a system that uses NED to automatically detect homonyms and tries to find the right one according to the Bunescu & Pasca paper (2006) There are multiple academic papers that propose a similar solution like: Eshel et al. (2017) and Yamada et al. (2016). If the right word cannot be automatically chosen from the Wikipedia list, we propose a system in which the user chooses the definition according to the sentences around the word from the original text. A proposed functional design of such a system is given in figure 10. This uses the example of the word: "record".



Uncertain word: Current

Original text

In a database, multiple data structures are defined. A record is one of them. Records in a database or spreadsheet are usually called "rows".

Options

1. Record (computer science), a data structure
2. Moving pictures, film, video, and television recordings
3. Sound recording and reproduction

Figure 10: Functional design of a manual homonym detection system

This system, together with automatic homonym detection, should cover the largest part of possible homonyms on the conditions that homonyms are successfully detected. This application seems to be the most complete in the current point in time. Other automatic detection systems are still at a starting point. It is worth keeping an eye out for these methods in case they propose a more complete and automatic solution. NED can be used to find homonyms and to find definitions for keyphrases.

No other academic sources were found that describe the process of finding definitions in text from an existing list of keyphrases. The most ideal approach is to find the definition of a term according to the context it is in. However, no approach was found that properly tackles this problem. The system proposed in figure 10 shows a possible approach for this problem. The downside is the manual labour, but it does tackle multiple problems at once like: homonym/synonym finding, finding definitions and finding relations. Manual labour also comes with a higher likelihood of proneness to errors. It does not rule out different approaches in the future, but for now this seems like the most feasible solution to multiple limitations in the concept mining process. The data domain manager confirmed that a small amount of manual labour is fine for an organization: "that you have to do stuff manual, is not a problem, that is part of it".

5.5 Preparation for modelling & the rest of the process

This chapter is a preparation of the next requirements (5 to 10). It offers tools and advice for future research into these next steps. A next step that already has been described was the finding of synonyms, which is similar to finding homonyms.

5.5.1 Data structure input

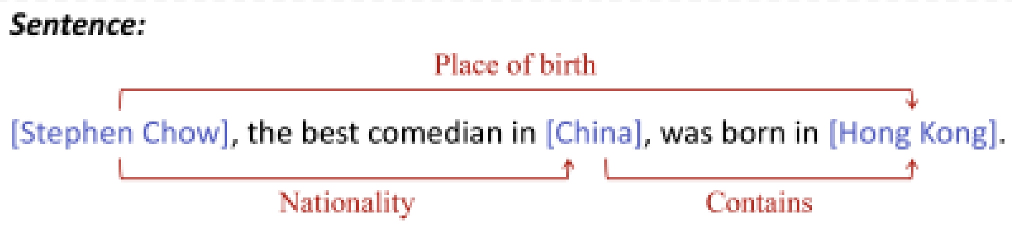
To link the concept mining process to the creation of a data structure, we need to look at what the input of a data structure is, so the knowledge base can be prepared for this step. If the knowledge base is of good quality, it can be mapped one-to-one to the data structure entities, since the entities of data structures are also keywords or keyphrases. The challenge here lies in doing as many steps as possible automatically, since not every step has a possible automized step. Another challenge is defining entities and attributes. This step has not been researched yet for this thesis.

5.5.2 Entity linking

The hardest part of knowledge graph construction is the extraction of relations. A possible widely used approach is entity linking. This is the task of linking entities (or keyphrases) to existing knowledge bases (Kertkeidkachorn & Ichise, 2017). These existing knowledge bases already know the links between entities, which makes knowledge graph construction a lot less complicated.

The problem that was described in the previous paragraph, is that definitions in organizations do not have to be the same as the definitions in these knowledge bases. Manual selection of definitions from these knowledge bases can be a solution, but what can be done about keyphrases that do not have a definition similar to one in any knowledge base known.

The only other possible way is to look at the context a keyphrase is in. This can be done via triples. A triple is a sequence of words between two entities (including the two entities) that need to be linked (Shang, Huang, Sun, Wei, & Mao, 2022). Figure 8 illustrates three triples.



[Stephen Chow], the best comedian in [China], was born in [Hong Kong].

Figure 11: Triples as described by Shang et al. (2022)

They are used to gather the relation between two terms. If you combine this with entity linking, you get a pretty high certainty of finding a relation. This system however would need to be built from scratch using these principles. There is no concrete example system that implements both in a manner that would be suitable for our context. The last problem is that, whenever both approaches are not suitable for a keyphrase, linking would have to be done manually. This is not an easy task, especially if a knowledge graph is large, since not all entities can be easily overseen.

5.6 Summary

This chapter was concerned with defining the first four steps of concept mining and giving direction to the other steps of the requirements. To identify types, custom NER needs to be performed. This will include a supervised learning model and a substantial amount of manual labour to teach that model. Existing tools like Azure custom NER or SpaCy could be used to tackle this problem. NER can again be used to identify instances, based on the types that have been previously established. Ranking types will be a step with two parts. If a type occurs in text, KeyBERT or another keyword ranking mechanism could be used to determine the importance of that keyphrase. If the type does not occur in the text, its ranking should be based on the combined ranking of its instances. Existing knowledge bases could be used to detect homonyms. This can be done partially automatic, based on the context of the keyphrase. The context then gives a definition of the term. If the context does not pose a concrete definition. A system could be constructed to do manual homonym detection, based on existing knowledge bases. For the other requirements, a valuable research direction is entity linking. It should be kept in mind that the posed data structure requires specific input.

6 Discussion

The first part of the discussion will look at limitation and a reflection on the results. The second part will specifically look at the validity of the findings.

6.1 Reflection & limitations

A recurring theme throughout this thesis was flexibility. It is an important part, since the process consists of ten steps that are not necessarily set in stone. A certain implementation can be given, but there are always other ways of achieving the same goals. We have tried to find the optimal implementation within the context, but this could change depending on the use case and context. Concept mining is a new term that could be used in many different forms, and that is a plus. It really is an asset to the data management field.

That flexibility also translates to the amount of data that is necessary. In the beginning, the goal was to try and cater to absolutely every organization in every context. However, it became apparent that a certain amount of data is necessary, because of the use of supervised methods. This poses a constraint on the amount of data and probably also the size of the organization. When conceptualizing this thesis, the manual labour that a future system should have was to be kept to a minimum. More labour or manual intervention has to be done in order to keep the concept mining program on the right track. With the training of supervised methods, it is almost impossible to do this fully automatic without manual intervention. This has been a limitation to the research. Also, the dependence on existing knowledge bases has been greater than anticipated. These external knowledge bases are reliable and are not expected to be stopped in the near future. Otherwise, different knowledge bases need to be found in future research.

Another point to take into account in the future is the support of multiple languages. For now, Dutch and English have been chosen since this is our target region. But in future endeavours more languages could be implemented, since most techniques and tools mentioned support even more languages. The input and output give room for interpretation. They have been kept intentionally vaguer, since this really depends on the organizational use case and context. A data lake and data model can take different forms, to take the need of the organization into account. This is also why concept mining requires specific demands for every organization. The process described so far has taken this into account. This should also be the case in the exact implementation. The keyword here is flexibility.

One of the major strengths of this thesis has been that it has used many practical tools that are pretty much instantly ready for implementation. This should help a lot with a future implementation. This means that it has been very practically applicable, which also was a demand from the clients that this research originated from. All the technologies that have been mentioned have been proven technologies as well. This means that we expect little problems when using these tools.

6.2 Validity of the thesis

Before doing the academic literature review, seven criteria for validity were defined. As far as we are concerned, these criteria were all met. Since the literature review is a huge part of this thesis, this was the most important part for the validity. The list of requirements was made based on our insights, other people could think differently. But there is no other way, since no concept mining systems have ever been constructed, to come to a complete list based on existing literature. This is a small constraint.

Interviews are also of course subjective, even though the interviewees had authority in their respected fields. This is the same situation as the requirements, however subjectivity is not a constraint in itself. Some things just cannot be explicitly proven in an academic way. This is why the requirements and interviews are valid ways of verifying and expanding the results. One specific part was not backed by academic literature, but by practical implementations. This is the identification of types (step one of the process). There was no other way to solve this problem, so this is an accepted breach in the methods. The validity of the overall thesis was not impacted. The overall validity conclusion is that the validity has been safeguarded, which makes the results academically valid. Especially within the total context.

7 Conclusion & further research

Concept mining turns out to be an accessible way for organizations to improve their data management. Even for people who have little understanding of software and systems. To summarize this thesis, firstly the answers to the five sub-questions will be given. These questions together answers the main research question of: How can concept mining be used to define and give meaning to important terms in an unstructured textual source for business-oriented data management?

1. How do we define "concept mining" based on existing literature?
Concept mining is defined in short as: "an activity that focusses on the extraction of words or sequences of words from textual corpora (unstructured textual source)."

2. What research is there to find about concept mining and other related concepts in the field?
Concept mining is not well-defined, which means that existing academic knowledge was found in other related concepts. This resulted in an explorative study that delves into questions like: what types of terms should be identified, when is a term important, should supervised or unsupervised learning be used, and how do you evaluate a concept mining system? This leads into the choice of using unstructured data as input and a data structure as output. The implementation was also a key factor in making sure that concept mining will be of value in an organization.

3. What can concept mining add to the data management field?
Concept mining creates new opportunities for better data management. It helps to create more value from existing data. It would also be highly relevant for many people within an organization. It also creates understanding between employees, so they can speak on the same level. This also relates to the use of multiple languages. So concept mining can create clarity for everyone. From a technical perspective, it can help with finding hidden patterns and creating new insights in the existing data.

4. Which requirements should a concept mining system have?
Ten requirements summarize what a concept mining system should look like. They have been summarized in chapter 3. Four of them have been fully worked out to create the first steps towards a future system.

5. What should the process of concept mining look like?
The concept mining process can be modelled after the ten requirements. The first four steps were worked out with the combination of existing techniques. These steps were: identifying types, identifying instances, ranking types and the indication of homonyms & indication of definitions. A preparation towards the rest of the steps was also given.

More research is necessary in order to fully define the project from start to finish. Requirements 5 to 10 have to be researched fully. A few tips have been given to coordinate future steps in the concept mining process. Of course, an implementation would also take some work in the future. It is important that every step is described before an implementation can even be on the horizon. Entity linking is probably the most important field to keep attention to for most steps of the process. We expect that finding entities and attributes could also be a challenge, just like finding types and instances was. However, we do have faith in the use case of concept mining. It could be a valuable combination of new techniques and advancements in the field. We also see that the requirements list could be valuable for more use cases in the data management field. This could be a good starting point for more researched focussed on other concepts that are not concept mining.

This thesis has served as a first step into the popularization of concept mining. Many organizations would benefit from a concept mining system. Data management in internal data will become increasingly important as the amount of data keeps rising in the near future. We hope that concept mining will be used and researched much more often. Let this thesis be the first step into that direction.

References

Adafre, S., de Rijke, M., & Sang, E. (2007). Knowledge Extraction from Structured Sources. *international conference recent advances in natural language processing proceedings*, 18-24. Retrieved from http://lml.bas.bg/ranlp2007/DOCS/RANLP2007.pdf#page=18

Alhoshan, M., & Altwaijry, N. (2022). Auss: An arabic query-based update-summarization system. *Journal of King Saud University - Computer and Information Sciences*, *34*(6, Part B), 3732-3743. Retrieved from https://www.sciencedirect.com/science/article/pii/S1319157820305565 doi: https://doi.org/10.1016/j.jksuci.2020.11.027

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. *Database*, *2018*. Retrieved from https://doi.org/10.1093/database/bay101 doi: 10.1093/database/bay101

Blumberg, R., & Atre, S. (2003). The problem with unstructured data. *Dm Review*, *13*(42-49), 62.

Bougouin, A., Boudin, F., & Daille, B. (2013). Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (ijcnlp)* (p. 543-551). Retrieved from https://hal.science/hal-00917969/

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, *45*(1), 12-19. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199401%2945%3A1%3C12%3A%3AAID-ASI2%3E3.0.CO%3B2-L doi: https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L

Bunescu, R., & Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. *11th Conference of the European Chapter of the Association for Computational Linguistics*, 9-16. Retrieved from https://aclanthology.org/E06-1002

Cambridge-Dictionary. (n.d.-a). *Definition homonym.* Retrieved from https://dictionary.cambridge.org/dictionary/english/homonym

Cambridge-Dictionary. (n.d.-b). *Definition keyword.* Retrieved from https://www.oxfordlearnersdictionaries.com/definition/english/keyword?q=keyword

Cambridge-Dictionary. (n.d.-c). *Definition polysemy.* Retrieved from https://dictionary.cambridge.org/dictionary/english/polysemy

Cambridge-Dictionary. (n.d.-d). *Definition synonym.* Retrieved from https://dictionary.cambridge.org/dictionary/english/synonym

Cambridge-Dictionary. (n.d.-e). *Word classes and phrase classes.* Retrieved from https://dictionary.cambridge.org/grammar/british-grammar/word-classes-and-phrase-classes

Celebi, M. E., & Aydin, K. (2016). *Unsupervised learning algorithms* (Vol. 9). Springer. Retrieved from https://link.springer.com/book/10.1007/978-3-319-24211-8

Chang, C., Kayed, M., Girgis, M., & Shaalan, K. (2008). A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1411-1428. Retrieved from https://ieeexplore.ieee.org/document/1683775 doi: 10.1109/TKDE.2006.152

Custom named entity recognition - azure cognitive services. (n.d.). Retrieved from https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/custom-named-entity-recognition/overview

Dayan, P., Sahani, M., & Deback, G. (1999). Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*, 857-859. Retrieved from https://web.math.princeton.edu/~sswang/developmental-diaschisis-references/dun99b.pdf

Dibenigno, J. (2019, 01). Rapid relationality: How peripheral experts build a foundation for influence with line managers. *Administrative Science Quarterly*, *65*, 1-41. Retrieved from https://www.researchgate.net/publication/330573075_Rapid_Relationality_How_Peripheral_Experts_Build_a_Foundation_for_Influence_with_Line_Managers doi: 10.1177/0001839219827006

Dutton, J. E., & Ashford, S. J. (1993). Selling issues to top management. *Academy of management review*, *18*(3), 397-428. Retrieved from https://journals.aom.org/doi/abs/10.5465/amr.1993.9309035145

Eshel, Y., Cohen, N., Radinsky, K., Markovitch, S., Yamada, I., & Levy, O. (2017). Named entity disambiguation for noisy text. *CoRR*. Retrieved from https://arxiv.org/abs/1706.09147

Fang, Z., Cao, Y., Li, T., Jia, R., Fang, F., Shang, Y., & Lu, Y. (2021). Tebner: Domain specific named entity recognition with type expanded boundary-aware network. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (p. 198-207). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.18 doi: 10.18653/v1/2021.emnlp-main.18

Glazkova, A., & Morozov, D. (2022). Applying transformer-based text summarization for keyphrase generation. Retrieved from https://arxiv.org/abs/2209.03791 doi: 10.48550/ARXIV.2209.03791

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105. Retrieved from http://www.jstor.org/stable/25148625

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (p. 216–223). Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/1119355.1119383 doi: 10.3115/1119355.1119383

Johnson, M., Rosebrugh, R., & Wood, R. (2002). Entity-relationship-attribute designs and sketches. *Theory and Applications of Categories*, *10*(3), 94-112. Retrieved from http://web.science.mq.edu.au/~mike/papers/40.pdf

Kertkeidkachorn, N., & Ichise, R. (2017). T2kg: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the thirty-first aaai conference on artificial intelligence.* Retrieved from https://eecs.csuohio.edu/~sschung/cis612/KnowledgeGraphTextProcessing_IAAA2017.pdf

Khine, P., & Wang, Z. (2018). Data lake: a new ideology in big data era. *ITM Web Conf.*, *17*. Retrieved from https://doi.org/10.1051/itmconf/20181703025 doi: 10.1051/itmconf/20181703025

Khosrow-Pour, M. (2020). *Encyclopedia of Information Science and Technology, Fifth Edition.* IGI Global. Retrieved from https://www.igi-global.com/book/encyclopedia-information-science-technology-fifth/242896#table-of-contents

Knowledge graph Atlassian.com. (2023). Retrieved from https://community.atlassian.com/t5/Confluence-questions/Knowledge-graph/qaq-p/1565284

Lauche, K., & Erez, M. (2023). The relational dynamics of issue-selling: Enacting different genres for dealing with discontent. *Academy of Management Journal*, *66*(2), 553-577. Retrieved from https://doi.org/10.5465/amj.2020.1484 doi: 10.5465/amj.2020.1484

Learned-Miller, E. G. (2014). Introduction to supervised learning. *I: Department of Computer Science, University of Massachusetts*, 3. Retrieved from https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf

Lee, Y. (2021). Systematic homonym detection and replacement based on contextual word embedding. *Neural Processing Letters*, *53*(1), 17-36. Retrieved from https://link.springer.com/article/10.1007/s11063-020-10376-8

Li, Q., & Wu, Y.-F. B. (2006). Identifying important concepts from medical documents. *Journal of Biomedical Informatics*, *39*(6), 668-679. Retrieved from https://www.sciencedirect.com/science/article/pii/S1532046406000219 doi: https://doi.org/10.1016/j.jbi.2006.02.001

Lin, J., Zhao, Y., Huang, W., Liu, C., & Pu, H. (2021). Domain knowledge graph-based research progress of knowledge representation. *Neural Computing and*

Applications, *33*(2). Retrieved from https://doi.org/10.48550/arXiv.1206.4625 doi: 10.1007/s00521-020-05057-5

Liu, 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. *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 1831-1841. Retrieved from https://doi.org/10.1145/3292500.3330727 doi: 10.1145/3292500.3330727

Liu, S., Sun, Y., Li, B., Wang, W., & Zhao, X. (n.d.). Hamner: Headword amplified multi-span distantly supervised method for domain specific named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6358 doi: https://doi.org/10.1609/aaai.v34i05.6358

Llevadias Jané, J., Helmreich, S., & Farwell, D. (2005). Identifying jargon in texts. *Procesamiento del Lenguaje Natural*, *35*. Retrieved from https://www.researchgate.net/publication/28167432_Identifying_jargon_in_texts

Maedche, A., & Staab, S. (2004). Ontology learning. In *Handbook on ontologies* (p. 173-190). Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-540-24750-0_9 doi: 10.1007/978-3-540-24750-0_9

Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. *IJCSNS International Journal of Computer Science and Network Security*, *8*(2). Retrieved from http://paper.ijcsns.org/07_book/200802/20080246.pdf

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404–411). Association for Computational Linguistics. Retrieved from https://aclanthology.org/W04-3252

Moonat, D. (2022). *Custom named entity recognition using spacy v3.* Retrieved from https://www.analyticsvidhya.com/blog/2022/06/custom-named-entity-recognition-using-spacy-v3/

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data Lake Management: Challenges and Opportunities. *International Journal of Advanced Computer Science and Applications*, *12*(12), 1986–1989. Retrieved from https://doi.org/10.14778/3352063.3352116 doi: 10.14778/3352063.3352116

Palmer, D. (2010). Text Preprocessing. In *Handbook of natural language processing (second edition)* (p. 35-56). Taylor and Francis Group. Retrieved from http://nozdr.ru/data/media/biblio/kolxoz/Cs/CsNl/Indurkhya%20N.,%20Damerau%20F.J.%20(eds.)%20Handbook%20of%20natural%20language%20processing%20(2ed.,%20CRC,%202010)

(ISBN%209781420085921)(O)(692s)_CsNl_.pdf#page=35

Pradeep, W. N., R M M. (2019, 12). Literature evaluation criteria. Retrieved from https://www.researchgate.net/publication/338228064_Literature_Evaluation_Criteria doi: 10.13140/RG.2.2.23148.51840/1

Puri, S. (2011). A Fuzzy Similarity Based Concept Mining Model for Text Classification. *International Journal of Advanced Computer Science and Applications*, *2*(11), 115-121. Retrieved from https://doi.org/10.48550/arXiv.1204.2061 doi: 10.48550/ARXIV.1204.2061

Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, *181*(1). Retrieved from https://www.researchgate.net/publication/326425709_Text_Mining_Use_of_TF-IDF_to_Examine_the_Relevance_of_Words_to_Documents doi: 10.5120/ijca2018917395

Saha, R. (2021). Homonym identification using bert–using a clustering approach. *arXiv preprint arXiv:2101.02398*. Retrieved from http://rgdoi.net/10.13140/RG.2.2.29120.07681 doi: 10.13140/RG.2.2.29120.07681

Sarawagi, S. (2008). Information Extraction. *Foundations and Trends in Databases*, *1*(3), 261-377. Retrieved from http://dx.doi.org/10.1561/1900000003 doi: 10.1561/1900000003

Schramm, S. (n.d.). *Parts-of-speech.Info.* Retrieved from https://parts-of-speech.info/

Shang, Y.-M., Huang, H., Sun, X., Wei, W., & Mao, X.-L. (2022). Relational triple extraction: One step is enough. *arXiv preprint*. Retrieved from https://doi.org/10.48550/arXiv.2205.05270

Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computer Applications*, *109*(2), 18-23. Retrieved from https://www.researchgate.net/publication/272372039_Keyword_and_Keyphrase_Extraction_Techniques_A_Literature_Review doi: 10.5120/19161-0607

spacy.io. (n.d.). Retrieved from https://spacy.io/

Tong, Z., & Zhang, H. (2016). A text mining research based on lda topic modelling. In *International conference on computer science, engineering and information technology* (p. 201-210). Retrieved from https://www.researchgate.net/publication/303563965_A_Text_Mining_Research_Based_on_LDA_Topic_Modelling doi: 10.5121/csit.2016.60616

Turney, P. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, *2*, 303–336. Retrieved from https://doi.org/10.1023/A:

1009976227802

Unbehauen, J., Hellmann, S., Auer, S., & Stadler, C. (2012). Knowledge extraction from structured sources. In *Search computing: Broadening web search* (p. 34-52). Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-34213-4_3 doi: 10.1007/978-3-642-34213-4_3

van der Aalst, W. (2012, jul). Process mining: Overview and opportunities. *ACM Trans. Manage. Inf. Syst.*, *3*(2). Retrieved from https://doi.org/10.1145/2229156.2229157 doi: 10.1145/2229156.2229157

van der Aalst, W., & Stahl, C. (2011). *Modeling business processes: A petri net-oriented approach*. The MIT Press. Retrieved from http://www.jstor.org/stable/j.ctt5vjqff doi: https://doi.org/10.7551/mitpress/8811.001.0001

Wan, X., & Xiao, J. (2008). Collabrank: Towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd international conference on computational linguistics* (p. 969-976). Coling 2008 Organizing Committee. Retrieved from https://aclanthology.org/C08-1122

Weigand, H., Johannesson, P., & Andersson, B. (2021). An artifact ontology for design science research. *Data Knowledge Engineering*, *133*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169023X21000057 doi: https://doi.org/10.1016/j.datak.2021.101878

Wikipedia. (n.d.). *Concept Mining.* Retrieved from http://en.wikipedia.org/w/index.php?title=Concept%20mining&oldid=1136019394

Wu, X., Zhang, J., & Li, H. (2022). Text-to-table: A new way of information extraction. *arXiv preprint arXiv:2109.02707*. Retrieved from https://arxiv.org/abs/2109.02707

Yamada, I., Shindo, H., Takeda, H., & Takefuji, Y. (2016). Joint learning of the embedding of words and entities for named entity disambiguation. *CoRR*. Retrieved from http://arxiv.org/abs/1601.01343

Ye, N., Chai, K. M. A., Lee, W. S., & Chieu, H. L. (2012). Optimizing f-measure: A tale of two approaches. *CoRR*. Retrieved from https://doi.org/10.48550/arXiv.1206.4625

Zamani, H., & Croft, W. B. (2016). Embedding-based query language models. In (p. 147–156). Association for Computing Machinery. Retrieved from https://doi-org.ru.idm.oclc.org/10.1145/2970398.2970405 doi: 10.1145/2970398.2970405

Zheng, H.-T., Kang, B.-Y., & Kim, H.-G. (2009). Exploiting noun phrases and semantic relationships for text document clustering. *Information Sciences*, *179*(13), 2249-2262. Retrieved from https://www.sciencedirect.com/science/article/pii/S0020025509001108 doi: https://doi.org/

10.1016/j.ins.2009.02.019

Zhu, X. J. (2005). Semi-supervised learning literature survey. Retrieved from https://minds.wisconsin.edu/bitstream/handle/1793/60444/TR1530.pdf?sequence=1&isAllowed=y

Appendices

Appendix A: Interview questions

General questions

- Have you heard of concept mining before?
- What do you think of the current definition that we have defined?
- Do you miss anything in the requirements?
- Which techniques could be used in a concept mining system?
- What do you think about making knowledge graphs or entity relationship models?
- What do your day-to-day activities look like?
- Would concept mining be a benefit to your working activities?

Specific questions interview enterprise architect

- How do you think we could identify types and instances?
- How do you think we could rank keyphrases?
- How do you think about the identification of homonyms and synonyms?
- How do you think definitions could be extracted from texts?
- What do you think about the relevance of entities and attributes?

Specific questions interview data domain manager

- What was the end-goal of the posed case?
- Do the requirements correspond with what you had in mind?
- Which data should be the input of the system?
- What do you want to be the output?
- What should the conceptual framework look like?
- Should the conceptual framework include definitions for all terms?

Specific questions interview business consultant information, data & analytics

- What does the data lake at Han University of Applied Sciences look like?
- Would HAN benefit from concept mining?
- Which types of data flow through the organizations that could be used?
- Who could benefit from the concept mining system?
- What should the management of a new system look like?
- What does the current conceptual framework at HAN look like?

Appendix B: Quotation report

A selection of 56 quotes from the three interviews in Dutch (the original language).

Interview 1: enterprise architect

Quote 1: Binnen onze architecturen hebben wij ook documenten, bijvoorbeeld opgedeeld in documenttypes. Ik kan een document op allerlei manieren classificeren.

Quote 2: De archivaris heeft een thesaurus opgesteld, een digitaal structuurplan, waarbij wij allerlei documenten in types opgedeeld hebben.

Quote 3: Je moet je daarbij voorstellen: is dit document een besluit, is dit document een overeenkomst, is dit document een advies, en dat helpt enorm las je dat soort dingen kan analyseren uit je bak ongestructureerde data.

Quote 4: Je gaat sowieso je documenten analyseren, maar als je daarin de verbinding kan leggen met documenttypes, maar dan gestructureerd volgens een bepaalde lijst, dan gaat je dat helpen denk ik.

Quote 5: Het tweede is van het type document, ook vaak weer gerelateerd aan een proces, heeft een aantal zaken vanuit de archiefwet uiteraard. Dat betekent dat er andere betwaartermijnen bij horen.

Quote 6: Dus het proces bepaalt uiteindelijk een soort van ranking en uiteraard ook in termen van privacy en security. Daar zit ook een soort van ranking in. Dat gaat dan over beschikbaarheid, integriteit, vertrouwelijkheid, die rankings. En dat bepaalt ook hoe je daarmee om moet gaan. Er zijn meer manieren om je documenttypes te ranken.

Quote 7: Of je nou vanuit security kijkt of vanuit archivering of vanuit architectuur of vanuit privacy of whatever, je komt altijd bij hetzelfde proces uit en iedereen heeft daar zijn eigen ranking in. En dat gezamenlijk kun je als organisatie van zeggen, dat bepaalt eigenlijk weer hoe belangrijk we dat vinden.

Quote 8: Daar hebben wij voor de belangrijkste entiteiten keuzes gemaakt van wie is eigenaar van die entitieit?

Quote 9: We hebben ook zo'n disussie gehad. Deelnemer aan het onderwijs, deelnemer aan het onderzoek, deelnemer aan een regeling. We hebben er voor gekozen de deelnemer entiteit voor te behouden aan het onderwijs domein en dat we het bij onderzoeken over een onderzoekssubject hebben. Zo hebben we gekeken naar homoniemen.

Quote 10: Ja, synoniem, afstudeerscriptie, afstudeer thesis, afstudeer enz. dat hebben we in die thesaurus ook vastgelegd. We hebben er voor 1 gekozen die in de business het meest gangbaar is en die andere hebben we genoteerd als gegeven. De thesaurus heb je gewoon nodig om alles bij elkaar te brengen.

Quote 11: Entiteiten of attributen of documenten hebben een relatie met elkaar.

Quote 12: Elke entiteit heeft z'n eigen attributen. Maar ik kom toch ook steeds vaker tot ontdekking dat een attribuut ook weer op zichzelf een entiteit kan zijn.

Quote 13: Maar we hebben krankzinnig veel documenten in totale versnippering over de hele HAN staan.

Quote 14: Dat betekent dat je daar afspraken of conventies voor moet maken. Dus concept mining afspraken of algoritme afspraken, zodat als je weet dat je een bepaalde vraag stelt, dat je ook de juiste antwoorden krijgt.

Quote 15: Op de middelbare school heb je de vakken aardrijkskunde en geschiedenis. Nou op het HBO en de universiteit komen die vakken helemaal niet meer voor, maar hoe noem je ze dan wel? Nou wat blijkt, er is geen eenduidig lijstje van het genereren van welke dat zijn.

Quote 16: Het is een beetje Engels, een beetje Nederlands. Het is alles door elkaar heen. Iedereen gebruik daar zijn eigen jargon weer in. Dus als we daar in een vorm gestructureerd naar kunnen metadateren en terugvinden zou dat het terugvinden naar spullen enorm vereenvoudigen.

Quote 17: Als jij student bent en je oriënteert je op het opleidingsaanbod van het hoger onderwijs en je kent her jargon nog niet, want je komt van de middelbare school, hoe zoek je daar dan op?

Quote 18: Kijk naar standaarden, misschien is dat een suggestie?

Quote 19: Dat is dezelfde problematiek waar ik in de dagelijkse praktijk ook tegen aanloop. De voertaal bij de HAN is Nederlands, maar de wereld is Engels en we lopen er steeds vaker tegenaan dat we dingen moeten vertalen.

Quote 20: Waar het bij de overheid moeilijk loopt is dat we elkaar denken te begrijpen, maar we elkaar gewoon niet begrijpen. Zo ontstaat er allemaal ruis. In 9 van de 10 gevallen ben ik ook bezig met de vraag te stellen: wat bedoel je in deze context? Want anders begrijp ik je gewoon niet.

Interview 2: data domain manager

Quote 1: We weten dat we heel veel woorden, concepten hebben opgeschreven inde verschillende wiki's en documenten.

Quote 2: Maar eigenlijk weten we niet waar het staat of wat het is, dus dat je het niet terug kunt vinden.

Quote 3: Het idee was om dat allemaal bij elkaar te krijgen door middel van een soort knowledge graph.

Quote 4: Van de andere kant ga je ook zien dat je verschillende interpretaties hebt over vraagstukken. Daar zit wel een aanvullende vraag achter, namelijk: wat is dan wat?

Quote 5: Ik zit daar wat strikter in. Sommige dingen zijn gewoon niet waar, want iedereen kan wel vinden dat het zo is, maar er is ergens wiskundig of natuurkundig vastgesteld dat het zo is.

Quote 6: We weten dat heel veel dingen verborgen zijn in documenten. Niemand kan ze verbinden. En hoe zou je daar iets mee kunnen doen?

Quote 7: Ik denk dat daar het meest uitdagende stuk in zit. Wat is nou het concept dat we willen beschrijven. Die is inderdaad wat abstracter en concptueler van aard. Maar het is in feite, hoe kijken we nou tegen een mens aan? En je zult zien dat als je in de documenten gaat rondstruinen, dat iedereen daar een eigen beeldvorming bij heeft.

Quote 8: Onze grootste uitdaging zit, dat we geen besef hebben van wat het begrip is.

Quote 9: Dan heb je het gewoon over documenten en dat is eigenlijk gewoon systeembouw termen en het mooiste zou dan zijn: een soort drag-and-drop achtig systeem, Je pleurt er een word document in, en dan scant en leest hij hem.

Quote 10: Rond een begrippenkader gaan verschillende beelden overheen. Maar ik zou willen kijken, je hebt synoniemen en homoniemen, maar ergens zit het beginpunt zegmaar.

Quote 11: Waarschijnlijk dat je heel veel homoniemen en synoniemen gaat hebben. Maar ergens moet je zeggen, dit is de definitie die wij als vertrekpunt kiezen.

Quote 12: Dat je handmatig iets moet doen is niet erg, dat hoort erbij.

Quote 13: Een hoge mate van zekerheid is heel belangrijk.

Quote 14: Dus het is niet alleen, ik heb wat gevonden, maar je maakt ook een conclusie daar duidelijk uit.

Quote 15: Tegenwoordig zou je ook de link kunnen leggen met fakenews, dat zou je langs die lat op kunnen leggen en kunnen uitsluiten, kijk dit is fakenews.

Interview 3: business consultant information, data & analytics
Quote 1: Daar is vorig jaar een onderzoek naar gedaan. Hoe kan je data management inrichten voor die ongestructureerde informatie.

Quote 2: Dus dan nogmaals, dan het je het over archivering van ongestructureerde informatie in een vorm van documenten of in welke vorm dan ook die wij aanduiden en archiveren.

Quote 3: Wat slaan wij nu eigenlijk op voor een type data, zowel aan gestructureerde zijde als aan ongestructureerde zijde. En vooral aan ongestructureerde zijde aan wat voor type slaan we op en waar kan ik bepaalde informatie vinden op basis van sleutelwoorden. Ja, dat heeft absoluut nut.

Quote 4: Als je vanuit ongestructureerde data een aantal kenmerken bijvoorbeeld uit zou kunnen halen die de vindbaarheid van de ongestructureerde data beter maken. Ja, dat heeft absoluut nut.

Quote 5: Want volgens mij is de essentie dat een organisatie of instelling steeds beter inzicht krijgt in welke type stroomt er dan allemaal door die organisatie.

Quote 6: Ja dan weten we wat er door een organisatie stroomt, maar dat heeft vaak een ondersteunende functie en er is natuurlijk veel meer data. Dus hoe meer zicht je daarop krijgt, op welke type informatie er allemaal is in je organisatie en wat je daarmee zou kunnen. Hoe beter je die waarde die in die data zit kan exploiteren.

Quote 7: Als ik denk aan patronen, denk ik ook aan process mining, om te kijken van welke patronen heb je dan.

Quote 8: Als jullie in staat zijn om bepaalde belangrijke begrippen, om die er uit te filteren, belangrijke begrippen voor de organisatie en daar ook context aan geeft, dus ook een beschrijving aan geeft. Ja, dan zie ik daar duizend en één mogelijkheden voor.

Quote 9: Alleen al het feit van waar moet je over praten. Moet je praten over student of over deelnemer. Want als je praat over student, sluit je een cursist uit. Maar wij geven natuurlijk ook onderwijs, cursussen en dus hebben wij ook cursisten Dus moet je dan praten over een deelnemer, die student en deelnemer omvat?

Quote 10: Daarnaast als je gaat kijken naar een onderwijsinstelling. Daar gaat zoveel ongestructureerde data rond en iedereen voert daarin z'n eigen dialect.

Quote 11: En zeker bij ongestructureerde informatie helemaal. Daar zit een eigen interpretatie en betekenis in.

Quote 12: Dit zijn klaarblijkelijk belangrijke begrippen die wij hanteren in deze organisatie, maar daar zitten veel homoniemen en synoniemen in etc. Hoe kunnen we nou het beste komen tot een eenduidige taal die we onderling spreken?

Quote 13: Als je daadwerkelijk in de praktijk een code gaat inbrengen. Ja dan wordt er ook weer een eigen interpretatie aan gegeven en een eigen kleur aangegeven. Dus die coderingen gingen enorm van elkaar verschillen en dat is gewoon een gigantische opgave geweest om te komen tot een eenduidig, en in de praktijk ook, eenduidig coderingsstelsel met enorm veel weerstanden.

Quote 14: Als ze het hebben over datamodellen, kunnen we kijken naar die modellen die in die transactionele systemen zitten. Die op een bepaalde manier gemodellerd zijn. Je hebt ook weer data modellen in data warehouses.

Quote 15: 15. Daarnaast heb je ook weer data lakes die dan weer veel ongestructureerde data hebben.

Quote 16: Hoe positioneer je zou zo'n datamodel in het geheel van de verschillende modellen die een organisatie al bezit.

Quote 17: Die systemen moeten ook beheerd en verbeterd worden en dergelijke, dus hoe ga je daarmee om?

Quote 18: En wat voor extra belasting is dat dan voor een organisatie naast die andere beheertaken die ze al hebben op die andere modellen.

Quote 19: Waar komt het systeem terecht? En wie is daar voor verantwoordelijk? En om hoeveel werk gaat dat? En hoe ga je dat opleveren?

Quote 20: Een business case is natuurlijk één ding. Laten we zeggen, meer de economische waardering, min of meer om die scherp te krijgen en die inversteringsverantwoording. Maar het gaat natuurlijk uiteindelijk om het gebruik van de informatie die het oplevert en de acceptatie daarvan en de implementatie daarvan. Dat is ook een verhaal apart.

Quote 21: Hoe zorg je nou dat mensen die informatie die het oplevert ook daadwerkelijk gaan gebruiken bovenop hun al drukke takenpakket? En hoe zorg je er nou voor dat ook de mensen op de werkvloer, in dit geval docenten, onderzoekers en dergelijken de echt de waarde daarvan inzien en daar ook naar handelen.