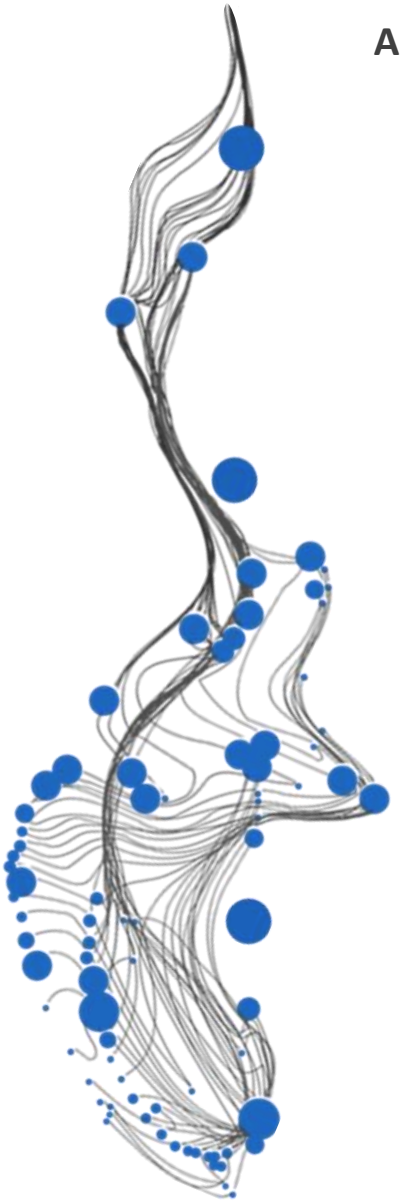
ORGANISING AND DISCLOSING RIVER KNOWLEDGE AT RIJKSWATERSTAAT

A recommendation to the Platform River Knowledge regarding a pilot website for knowledge disclosure



29-06-2020

Bachelor Thesis

T.J.A. Luyten – s1845640

University of Twente – BSc Civil Engineering

Supervisor: Lieke Lokin

Rijkswaterstaat – Water, Verkeer en Leefomgeving

Supervisors: Mirjam Flierman, Saskia van Vuren

UNIVERSITY OF TWENTE.

Rijkswaterstaat
Ministerie van Infrastructuur en Waterstaat

Preface

This thesis has been produced over the past ten weeks and is the final product in completing my Bachelor of Civil Engineering at the University of Twente. In this study I have set out to design a conceptual version of a pilot website for disclosing river knowledge, commissioned by Rijkswaterstaat.

It has been an interesting period that has not turned out as initially expected. Due to the regulations surrounding the COVID-19 pandemic, the entire research process was done from home. Unfortunately, I have not had the complete experience of working at Rijkswaterstaat among colleagues. Nevertheless, I have learnt a lot during this period, and I am glad that I am able to finish my bachelor's degree because of my time working for Rijkswaterstaat.

I would like to thank Lieke Lokin for the supervision and feedback on the academic part of this study. I would also like to thank Mirjam Flierman and Saskia van Vuren for the supervision within Rijkswaterstaat and the feedback on my report. Lastly, I would like to thank the other colleagues at Rijkswaterstaat that have helped me in the research process, particularly the people whom I have interviewed.

I hope you will enjoy reading my bachelor thesis.

Tim Luyten

Enschede, 29th June 2020

Abstract

As the main organisation that manages the river system of the Netherlands, Rijkswaterstaat (RWS) is closely involved in producing, managing, and disclosing river knowledge. It is important that this knowledge is easily accessible, but this is not always the case. The Platform River Knowledge (Platform Rivierkennis) is a community of practice at RWS that sets out to improve this. One of the goals of the Platform is to launch a pilot website for disclosing river knowledge. This study is a preliminary investigation on the topic and provides the Platform with recommendations on organising river knowledge, automatically categorising documents, and shaping a pilot website.

A literature review was carried out to explore different taxonomy forms, or knowledge structures, that could be used to organise river knowledge. Interviews were conducted with eight RWS employees working with river knowledge to discuss the results of the literature review and determine which categories to use for organising river knowledge. A machine learning model has been made to sort documents into predefined categories, based on the Naïve Bayes algorithm. Finally, the results were combined to create a conceptual version of the pilot website.

The chosen taxonomy form is a facet structure, which works by using multiple overlapping categories. Relevant categories are selected, and only items that belong to all selected categories remain in the search results. This is a useful way of organising river knowledge, as documents containing river knowledge often belong to multiple categories. The choice for a facet structure was approved by all interviewees. The categories proposed by interviewees have resulted in four alternative set-ups ranging from broad to specific. These alternatives should be treated as suggestions, as other combinations can be made.

The Naïve Bayes model has been applied to the simplest alternative which features two sets of categories and has been tested using eight documents. The model has correctly predicted the first category 6 out of 8 times and the second category 7 out of 8 times. However, it is unable to sort documents into multiple categories, which is needed if a document belongs to more than one category. The model is not fully functional.

The conclusion of this study is that the pilot website should use a facet structure, with the provided alternatives as suggestions for knowledge organisation. A Naïve Bayes model can be used to categorise documents, but the uploader should check to make sure that documents are labelled correctly. The results have been combined into a conceptual version of the pilot website, provided through visual examples.

The Platform is recommended to do further research into the demand for a website for disclosing river knowledge. Through the interviews, the demand turned out to be low among RWS employees. If a new website is introduced, interviewing a broader group of stakeholders is recommended to ensure a large support base among users.

Table of Content

1. Introduction

This section provides the background and motivation for this study. The objective and research questions are defined, providing the basis for the research. This gives an overall idea of the set-up and significance of the research.

1.1 Background

The Netherlands is characterised by the presence of water. The country is located in the river delta of the Rhine and the Meuse, which flow into the North Sea through different branches. The river delta of the Netherlands brings advantages, such as very fertile soil and the possibility of navigation over water. The river delta also brings challenges in the form of flood risk and high complexity of water management.

The intricate river system of the Netherlands must be managed adequately, especially in present times, under the increasing influence of climate change and socio-economic developments. The main organisation that manages the Dutch river system is Rijkswaterstaat (RWS), which is part of the Ministry of Infrastructure and Water Management.

The major tasks of RWS are to manage and develop the main roads, waterways, and water systems of the Netherlands. In order to efficiently fulfil these tasks, RWS is also involved in the production and management and disclosure of relevant knowledge. This means that RWS produces a large amount of knowledge within the domain of the river system. This knowledge is required to successfully fulfil the RWS roles of policy advisor, manager of the river system, and knowledge supplier.

It is important that river knowledge is easily accessible to accommodate the roles that RWS has in the river system. Currently, river knowledge at RWS is not always easily accessible. There is a central database where all knowledge documents at RWS can be stored, called ‘Kennisplein’. This database contains many documents conveying river knowledge, but it is not well-structured and therefore not easy to use. Because of this, knowledge is often not centrally shared but instead remains within the part of the organisation where it was produced (Rijkswaterstaat, 2020). This makes it difficult to find relevant knowledge from across and outside the organisation. It also causes managers to be asked the same questions by different people.

Due to these inconveniences, among other things, the board of RWS decided in 2017 to establish the ‘Platform River Knowledge’ (Platform Rivierkennis), which was launched on January 1, 2018. This platform is a community of practice at RWS which intends to improve the process of production, storage and disclosure of river knowledge (Rijkswaterstaat, 2020).

One of the goals of the Platform River Knowledge is to create a pilot website which will be used to disclose river knowledge in a clear and ordered manner. This study focuses on finding a method to systematically order documents, in a way that makes it easier both to share and to find river knowledge. This method will then be used to create a conceptual lay-out of the pilot website.

No prior research on this topic has been done, so this study serves as a preliminary investigation with recommendations to the Platform on how to organise river knowledge, categorise documents, and shape a pilot website for disclosing river knowledge.

The research is conducted at the department Hoogwaterveiligheid (flood safety) within WVL (Water, Verkeer en Leefomgeving), one of the national sections of Rijkswaterstaat. This department offers the main effort for the Platform River Knowledge.

1.2 Problem statement

The disclosure of river knowledge within RWS currently does not work as efficiently as desired. Employees in regional sections of RWS and other parts of the organisation do not always share documents containing river knowledge on Kennisplein. This means that relevant river knowledge cannot always be found by RWS colleagues and other parties, such as research institutes, universities, and consultancy companies. This is also the case for older documents that have not (yet) been disclosed. Additionally, Kennisplein lacks structure, so available documents are difficult to find. Because of this, there is currently no clear inventory of all available river knowledge at RWS.

Since there is currently no systematic method of categorising river knowledge, it is not yet possible to launch a pilot website where river knowledge can be publicly disclosed in a clear and ordered manner.

1.3 Research objective

The main objective of this study is to design a conceptual version of a pilot website for disclosing river knowledge. To accomplish this, a structure for organising river knowledge documents must first be set up by making informed choices based on scientific literature and available expertise within Rijkswaterstaat. A text sorting algorithm will be tested, to determine if this method can be used to automatically categorise documents containing river knowledge. This can help in making an inventory of already available river knowledge and make it easier to share new knowledge. The testing of this algorithm is meant to serve as a *proof of concept*, to illustrate on a conceptual level that this method has functional potential.

This research serves as a preliminary investigation on the subject and provides the Platform River Knowledge with recommendations regarding the organisation of river knowledge and the shaping of a pilot website for disclosing river knowledge.

1.4 Research questions

One main research question has been formulated and divided into three relevant sub-questions. The goal of this study is to answer these questions, which will contribute towards reaching the research objective. The research questions are:

How can a pilot website for disclosing river knowledge at RWS ideally be shaped?

1. ***How can river knowledge best be organised according to literature?***
   The goal of this research question is to find out what type of knowledge structure can best be used to organise river knowledge based on scientific literature. This forms a theoretical basis for the remainder of the study.

2. ***How should river knowledge be categorised according to RWS employees?***
   Answering this question will refine and expand upon the theoretical basis of *sub-question 1* by combining it with practical knowledge and experience of RWS employees. This ensures that the chosen set-up for the pilot website will be practically applicable.

3. ***Can documents containing river knowledge be sorted into predefined categories using a text categorisation algorithm?***
   The goal of this research question is to find out if a text categorisation algorithm can be successfully applied to documents containing river knowledge. This method can then be used to create a model that sorts documents into previously defined categories. Different text categorisation algorithms are discussed in ***Section 2.3***.

2. Theoretical Framework

This theoretical framework is meant to refine the scope of this study and provide background information and context to the research.

2.1 Key concepts

The first step is to define the key concepts which are central to this study. The key concepts of this study are **river knowledge**, **Platform River Knowledge**, **Kennisplein**, **(conceptual) pilot website**, and **knowledge taxonomy**.

River knowledge

All knowledge related to the river system of the Netherlands. In this study, only river knowledge managed by RWS is relevant. Physically, the scope of river knowledge is limited to the river system of the Netherlands, in the space shown in **Figure 2.1**. This is the space between the dike crests **(1)**, including the riverbed **(2)**, flood plains **(3)** and side channels **(4)**. For the part of the river Meuse in Limburg, the physical boundaries are determined by the bordering high grounds, instead of the dike crests (Rijkswaterstaat, 2020).

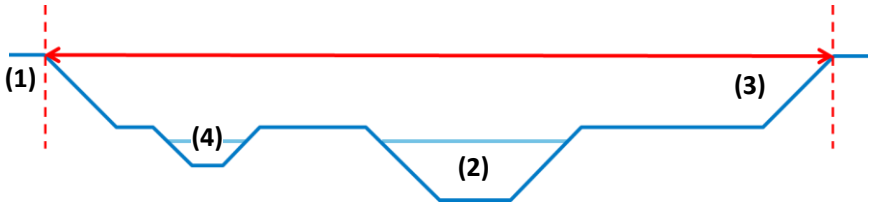


Figure 2.1 *– Physical boundaries that define the term 'river knowledge'* (Rijkswaterstaat, 2020)

Platform River Knowledge

A community of practice launched by RWS in 2018 in order to improve the knowledge exchange within RWS. The platform focuses on demand-based knowledge production and disclosure of river knowledge. This helps to ensure that knowledge gaps are identified, research is programmed and required river knowledge is produced. The scope of the platform is broad, and it covers every function of the river system of the Netherlands. This includes national functions such as water safety, navigability over water and freshwater availability. Other functions and regional interests include spatial planning, (urban) area development and agriculture (Rijkswaterstaat, 2020). This research is commissioned by the Platform.

Kennisplein

An internal database at RWS that is used to store and disclose knowledge documents within the organisation. Kennisplein has a search function but the database is not well structured, making it difficult both to share and to find documents through this database.

(conceptual) pilot website

A pilot website that RWS is considering launching, which will be used to disclose river knowledge. A conceptual structure of this pilot website will be the end product of this study.

Knowledge taxonomy

A structured classification scheme that is used to organise knowledge. A knowledge taxonomy will be constructed based on scientific literature and expertise within RWS. This knowledge taxonomy will serve as the basic structure for the conceptual pilot website.

2.2 Project context

River management in the Netherlands goes back to the Middle Ages. Because a considerable part of the country and its population are located near major rivers which were continuously shifting their paths, there was a constant risk of flooding land. It was not possible to sustain a growing population this way, so all main rivers in the Netherlands had been enclosed by dikes in the period between 1200 and 1400 AD (Van Baars & Van Kempen, 2009). The Dutch have continued to extend the flood protection systems up until today. A more recent project like the major Delta Works is a good example of this.

Every interference in the river system has its consequences. This makes it essential to have sufficient knowledge about the river system. As the main organisation responsible for managing the Dutch river system, Rijkswaterstaat is closely involved in the production, storage and disclosure of knowledge related to the river system. Because the process of production and disclosure of river knowledge was not working as desired, RWS decided in 2017 to introduce the Platform River Knowledge. An important function of the platform is the guidance of demand-based knowledge production; finding knowledge gaps and making sure that specific knowledge is produced to fill these gaps. The produced knowledge is then used to improve river management and create new policies.

One of the first important products of the Platform River Knowledge is the 'Story of the River' (Verhaal van de Rivier) and specifically the Story of the Meuse (2018) and the Story of the Rhine-Meuse estuary (2019). These stories focus on the history and challenges in managing the major Dutch rivers. The Stories of the River are written by experts who want to share their knowledge with RWS and parties that work with RWS in managing the Dutch rivers (Rijkswaterstaat, 2020). These parties include the national government, provinces, municipalities, water boards and market parties.

The Stories of the River also cover the main objectives that the river management is meant to serve, for which RWS is responsible. These are the core tasks of RWS in river management, as formulated in *''Beheer- en ontwikkelplan Rijkswateren 2016 (BPRW 2016)''* (Rijkswaterstaat, 2015):

- **Water safety**
- **Freshwater availability**
- **Navigability over water**
- **Water quality and nature**

An overarching theme used by the Platform River Knowledge in addition to the four main functions is **river morphology and sediment management**. Because these functions are essential in river management, the Platform River Knowledge needs to support RWS in serving these functions. This will be taken into account when deciding how to categorise river knowledge on the pilot website.

The sections of RWS which are most involved in producing and processing river knowledge are listed here:

- **The three 'river regions': Oost-Nederland (ON), Zuid-Nederland (ZN), West-Nederland-Zuid (WNZ)**
  These are the regional departments that contain the major rivers of the Netherlands. Each of these departments is responsible for the maintenance and construction of major roads and waterways in their respective regions.
- **Programma's, projecten en onderhoud (PPO)**
  PPO is a national department working on the maintenance of national roads, waterways, bridges, and other constructions. PPO collaborates with other sections of RWS and external parties.
- **Grote projecten en onderhoud (GPO)**
  GPO is a national department responsible for realising the major construction and maintenance projects of RWS, in cooperation with the regional departments.

- **Water, verkeer en leefomgeving (WVL)**
  WVL is a national department responsible for the main road network, main waterways, the river system, and their influence on the living environment.

The interviews in this study will mainly be conducted with employees from WVL and ON, as interviewing employees from every section will not fit into the timeframe of ten weeks. WVL is the section most involved with the Platform River Knowledge. ON is the river region with the largest area, covering the IJssel, Nederrijn and Waal. WVL focuses on the knowledge production and policy aspects of the river system, while ON focuses mostly on river management. Interviewing people from WVL and ON ensures that different perspectives on river knowledge are taken into account when deciding how to organise river knowledge.

The pilot website that RWS is considering launching for disclosing river knowledge will be used by different types of users. These user types include RWS employees, other government organisations like provinces and water boards, members of the scientific community and market parties. Each type of user will be looking for different types of information and will have different objectives. Interviewing people outside RWS is not within the scope of this study. Another thing to note is the three types of internal RWS users: river management, policy making, and knowledge development. People from these three groups will be interviewed, so their different perspectives can be taken into account when designing the conceptual pilot website.

2.3 Text categorisation algorithms

Some widely used text categorisation algorithms include Support Vector Machines (SVM), k-Nearest Neighbour (kNN), Linear Least Squares Fit (LLSF), Neural Network (NNet) and Naïve Bayes (NB) (Yang & Liu, 2020). These are machine learning models that categorise documents based on the text content found in training data. The training data in this case consists of documents that have been categorised correctly, so the model can assign new documents to the correct category after learning from the training data. The different models are briefly discussed here.

LLSF and NNet are complex methods compared to SVM, kNN and NB. Training NNet and LLSF models is more time consuming compared to the other methods and requires a larger amount of training data to function well (Yang & Liu, 2020). SVM also requires a long training time, although it does not require as much training data as LLSF and NNet (The Professionals Point, 2020). Due to the limited timeframe and training data available for this study, these methods have not been chosen. Out of the remaining methods, kNN and NB, the latter is the simplest and the easiest to implement. NB models also work well with a limited amount of training data (Simplilearn, 2020). Due to the limited time and training data available, the NB method is chosen to apply in this study.

The Naïve Bayes method rests on a few principles. The Naïve part of the NB method is the assumption that predictors, i.e. keywords, contribute independently to the probability that a document belongs to a certain category. In reality, keywords often contribute in conjunction with one another to the outcome value, i.e. category. This assumption makes models based on the NB algorithm more efficient because they require much less computation time. The downside of the NB method is that it is in many cases not as accurate as other methods. Despite this, NB models perform relatively well (Manning, Raghavan, & Schütze, 2008). An example of a problem in which NB models are often successfully applied is spam detection in e-mails (Pal, 2020).

3. Methodology

This section discusses the methods used to answer each research question in this study. **Figure 3.1** provides an overview of the steps that the research is divided into and the relation between the research questions.

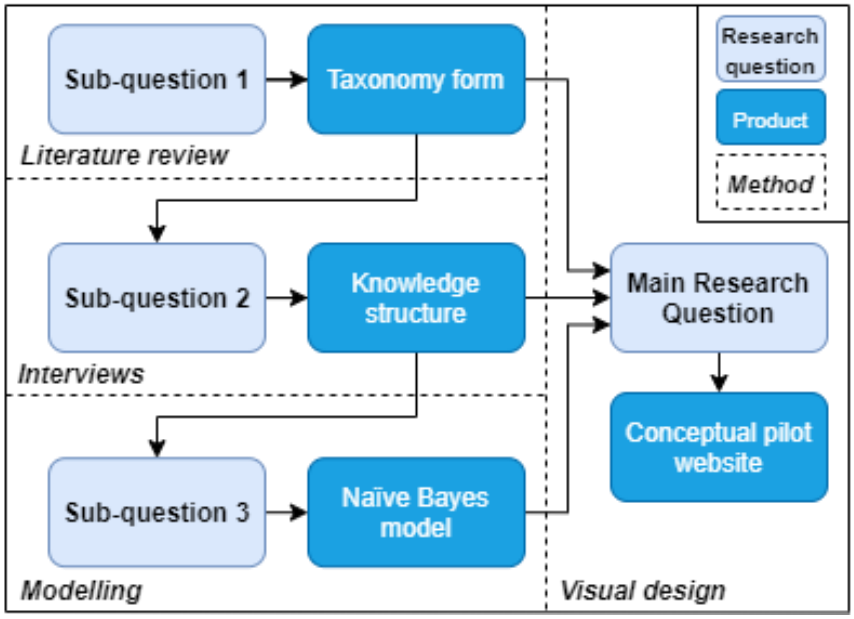


***Figure 3.1** – Overview of the research steps*

How can river knowledge best be organised according to literature?
This sub-question will be answered through a literature review. The literature review will focus on scientific literature that concerns organising and categorising knowledge in a broad sense. This will result in a chosen taxonomy form and an initial taxonomy set-up. This serves as a theoretical basis for the remainder of the research.

The main source for this literature review will be ''Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness'' (2007) by Patrick Lambe. This work is commonly cited within the field of knowledge management, with 257 citations (ScienceDirect, 2020). It goes in depth into the science of categorising knowledge in the form of taxonomies. This work conveys knowledge as a broad term, so the scope of this work is much broader than just river knowledge. However, the methods supplied by this source are still useful.

How should river knowledge be categorised according to RWS employees?
This sub-question will be answered by actively interviewing employees at different sections of Rijkswaterstaat involved in the use and development of river knowledge. The goal is to interview employees from the three different groups: river management, policy making and knowledge development. The main sections where employees will be interviewed are WVL and ON.

Interviewing RWS employees working with river knowledge is an effective way to find out how knowledge organisation can be improved. At RWS there are already ideas as to how river knowledge can be organised, so combining these ideas with the results of the literature review will result in a plan to categorise and organise river knowledge. The interviews will be held via an online connection like Skype. The interviewee will receive an introduction in advance, with the reasons for the interview and some

preparatory questions. The interviews will be in Dutch. After each interview session, a Dutch transcript and English summary will be made to process the information.

The interviews will have a semi-structured setting. This means that questions will be prepared in advance by the interviewer, while the interviewee has room to elaborate and add insights on subjects that the interviewer might not have thought of in advance. Barribal and While (1994) provide several reasons why a personal interview in a semi-structured setting is a useful method for collecting information:

- It has the potential to overcome the poor response rates of a questionnaire survey.
- It is well suited to the exploration of attitudes, values, beliefs, and motives.
- It provides the opportunity to evaluate the validity of the respondent's answers by observing non-verbal indicators.
- It can facilitate comparability by ensuring that all questions are answered by each respondent.
- It ensures that the respondent is unable to receive assistance from others while formulating a response.

The goal of the interviews is to first find out how familiar the interviewee is with the current knowledge disclosure methods and how extensively they use them. Then, the interviewee will be asked how they think the knowledge disclosure method could be improved and how to categorise river knowledge. The results of the literature review will be discussed with the interviewee. The combination of the interviews and the literature review will result in a knowledge taxonomy that forms the basic structure for the conceptual pilot website.

After a knowledge taxonomy has been set up, some time after the interview, the interviewees will be asked to categorise a number of documents. They will also be asked to point out keywords that are indicative of each category that a document can belong to. This will serve as training data for the Naïve Bayes model. The amount of training documents used will be at least 30, as this is the widely used minimum sample size for statistical inference (Webb, Boughton, & Wang, 2005).

Can documents containing river knowledge be sorted into predefined categories using a test categorisation algorithm?

A model will be made which sorts documents containing river knowledge into categories using the Naïve Bayes algorithm. The first step is to extract the occurrence of each keyword in each document of the training data. A MATLAB script will be made that is able to scan through each document and count the occurrences of every keyword. This serves as training data for the model.

The next step is to generate a probabilistic model based on the NB algorithm. This will be done in Excel using the training data and the number of occurrences of each keyword in the training documents. This enables the computation of the probability that a new document belongs to a certain category, given the occurrence of each keyword in the text of that document.

The last step is to test the accuracy of the model by using new documents as input. The categories these documents belong to will be predicted by the model, which is the model output. The predicted categories will then be compared to the correct categories as provided by the interviewees, to check if the model has assigned the documents correctly. If the model assigns the documents to the correct categories, this will serve as a proof of concept for the NB algorithm. This means that the model will be tested on a conceptual level, as opposed to creating a fully functioning model.

How can a pilot website for disclosing river knowledge at RWS ideally be shaped?

After a knowledge taxonomy has been set up, along with a Naïve Bayes model to sort documents into the right categories, the main research question can be answered. This is done by providing a conceptual design of the pilot website for disclosing river knowledge. This will be realised in the form of visual examples of the website lay-out to provide a clear idea of what the website can look like. Visual examples of the website will be created using PowerPoint.

4. Knowledge Organisation

The results of the literature review and conducted interviews are presented in this section. This covers the different taxonomy forms that can be used and results in four alternative structures for organising river knowledge on the pilot website. This section provides the answers to **Sub-question 1** and **Sub-question 2**.

4.1 Initial taxonomy set-up

A literature review was carried out in order to determine the most suitable taxonomy form to be used on the pilot website for disclosing river knowledge. The main source for the literature review is *''Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness'' (2007)* by Patrick Lambe.

Lambe describes and discusses the different forms that a knowledge taxonomy can take. A brief description of each taxonomy form is provided in **Table 4.1**. A more detailed description and explanation of the different taxonomy forms can be found in **Appendix A**.

Table 4.1 – Overview of different taxonomy forms and their respective features

Taxonomy form	When to use	Downsides or issues
List	Lists can be used in very simple situations, with no more than 12-15 items. Items in a list must be connected by one common feature.	It is not possible to describe relationships between items. Lists are too simplistic to describe any conceptual structure or process.
Tree structure	Trees can be used when a list grows too long, and items can be clustered into logical subgroups. Trees are versatile because they can express different relationships between different items.	The versatility of the tree structure causes inconsistency. Items on the same level in a tree can represent different types of information. This can cause a loss of clarity, especially in larger structures.
Hierarchy	Hierarchies are a specific type of tree structure. They are effective when dealing with strictly defined items and classes. The consistency of the structure makes content easy to find. Hierarchies are most useful in hard sciences.	Hierarchies do not deal well with the complexity and ambiguity of the real world, especially outside of exact fields. This is because hierarchies do not acknowledge that categories might overlap or that terms can be ambiguous.
Polyhierarchy	Polyhierarchies can be used when hierarchies fail to represent items that belong to more than one category. This is done by cross-linking classes or entities to more than one superordinate class.	Polyhierarchies erase the rigidity of the hierarchy structure, which is its main strength. Cross-linking will very quickly cause the structure to become confusing to users and difficult to manage.
Matrix	Matrices are most effective when every entity can be organised along the same two or three dimensions. The content can easily be identified and compared due to the matrix lay-out. It is possible to add more dimensions through colour coding and restructuring of the matrix.	Diverse collections of knowledge cannot be expressed using a matrix because the entities cannot be described along the same few dimensions. Content that requires more than three describing dimensions is difficult to represent in a matrix.

Facets	Facets are useful for organising large amounts of content because items can be classed into multiple categories simultaneously. Search results can then be narrowed down by selecting all relevant categories. Facets are especially effective in digital knowledge databases, where items can easily be stored in multiple locations.	The use of facets requires a certain level of subject knowledge because the user must understand in which categories an item can be placed. This can be a problem when dealing with specialist knowledge presented to a general audience.
System map	System maps are a visual representation of a knowledge domain and the relations between item within the domain. This is useful when knowledge can be presented visually because they have a strong representational power. System maps can either be descriptive (representing a real-world domain) or conceptual (representing non-physical constructs).	System maps do not work well in complex situations. The more complex a representation becomes, the more difficult it is to convey information visually. It is also difficult to represent hierarchy using system maps.

Different taxonomy forms have been explored and their respective features are now known. Based on the literature, the most applicable taxonomy form for organising the pilot website can be picked.

The first thing to keep in mind is the four main functions *(water safety, freshwater availability, navigability over water, water quality and nature)* currently used by RWS to distinguish between different types of river knowledge. It is desirable to keep the use of the four main functions intact because these functions are central themes for RWS. The four main functions are taken as a base for the knowledge taxonomy, but this is subject to change after conducting the interviews.

The next thing to note is that there is often overlap between categories. Projects and reports can cover more than one theme. For example, the study *''Dealing with uncertainty of accelerated sea level rise''* covers water safety and navigability (Rijkswaterstaat, 2020). Thus, river knowledge documents cannot be organised well without the ability to represent overlapping categories.

Several taxonomy forms immediately appear unsuitable: lists, tree structures and hierarchies are unable to deal with overlapping categories. Polyhierarchies, matrices, facets and system maps remain:

- A polyhierarchy could be used but would become difficult to manage. Adding the required crosslinks between each item and category would become inefficient and unmanageable. This is the case especially when new content is added.
- A matrix is not suitable in this case because documents cannot be organised along only two to four dimensions. A matrix would not be able to sufficiently narrow down a search.
- Facets seem like an applicable method. This method is especially useful when dealing with a large body of knowledge in a digital environment. Presenting specialist knowledge to a general audience can be problematic, but that is not a goal of the pilot website.
- A system map would be unable to accommodate a large and growing amount of content. It is not possible to convey this visually. A concept map could be used as the underlying structure behind a website but like polyhierarchies, adding the required crosslinks would require much effort.

A ***faceted taxonomy*** is the most applicable choice in this case because we are dealing with a large body of knowledge in a digital environment. Tags and metadata (e.g. author(s), title, keywords, date of

publication, place of publication) can be used in order to further specify search results (Lambe, 2007). This makes a faceted taxonomy ideal for a website used to disclose knowledge. The challenge in building this faceted taxonomy lies in choosing the right facets, or sets of categories.

The main principle of facets is that items can belong to more than one category. Starting with all items available, search results are narrowed down by selecting the different categories the item belongs to. Using the example mentioned before, *''Dealing with uncertainty of accelerated sea level rise''* belongs to water safety and to navigability. The search can be narrowed down by selecting both categories. This process is generically illustrated in **Figure 4.1**.

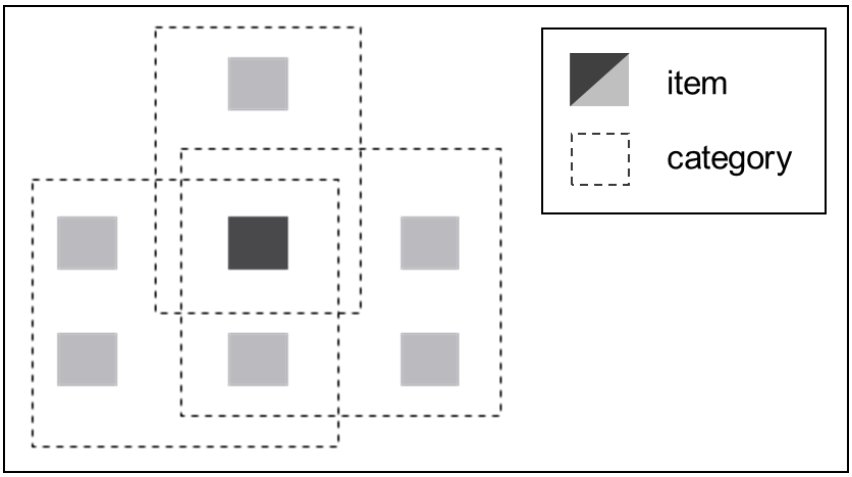


Figure 4.1 *– Graphical depiction of a faceted knowledge taxonomy* (Cloudset Solutions, 2020)

The initial set-up for the faceted taxonomy is based on the main functions used by RWS. The overarching theme of morphology and sediment management is left out, as it is not one of the core functions defined by RWS in BPRW 2016.

- **Water safety**
- **Freshwater availability**
- **Navigability over water**
- **Water quality and nature**

This is supplemented by distinguishing between different types of knowledge, as used by the Platform River Knowledge (Rijkswaterstaat, 2020). Miscellaneous has been added for documents that cannot be categorised into these knowledge types.

- **System knowledge**
- **Models and instruments**
- **Monitoring**
- **Measures and interventions**
- **Miscellaneous/other**

The different geographic areas involved can be used to further narrow down search results.

- **Meuse**
- **Rhine**
    - **Waal**
    - **IJssel**
    - **Nederrijn-Lek**

Finally, these tags and metadata can be applied on the search.

- **Title**
- **Author(s)**
- **Keywords**
- **Date of publication (period between A and B)**
- **Place of publication**

This is an initial set-up for the knowledge taxonomy. An overview is given in **Table 4.2**. It will be presented to RWS employees during interviews. Their understanding of the subject matter will enable to refine and improve this set-up.

Table 4.2 – Overview of initial taxonomy set-up

Main function	Knowledge type	Geographic area	Tags/metadata
Water safety	System knowledge	Meuse	Title
Freshwater availability	Models and instruments	Rhine - Waal - IJssel - Nederrijn-Lek	Author(s)
Navigability over water	Monitoring		Keywords
Water quality and nature	Measures and interventions		Place of publication
	Miscellaneous/other		Date of publication (DD-MM-YYYY – DD-MM-YYYY)

4.2 Interview results

In total, eight interviews have been conducted with employees from three different sections of Rijkswaterstaat. An overview is provided in **Table 4.3**. Summaries of each interview can be found in **Appendix B**.

Table 4.3 – Overview of interviews conducted

Name	Function	Section	Date
David Kroekenstoel	River consultant[1]	WVL	11-05-2020
Hendrik Buiteveld	River consultant	WVL	14-05-2020
Emiel Kater	River consultant	ON	15-05-2020
Margriet Schoor	Ecologist	ON	18-05-2020
Daniël van Putten	River consultant	ON	20-05-2020
Rien van Zetten	Project manager	GPO	25-05-2020
Ralph Schielen	River consultant	WVL	25-05-2020
Arjan Sieben	River consultant	WVL	27-05-2020

WVL focuses mostly on policy making and knowledge production in the river domain, while ON mainly works on maintenance and management of the river system in the region. The idea of interviewing people from WVL and ON was to gain knowledge from the perspective of each group. It was not a conscious decision to interview mostly river consultants. The interviewees are a combination of people already spoken to and those who were available. The interviewees working at ON were put forward by a contact person from ON. Because river consultants are overrepresented, interview results could be biased towards their preferences. Due to the limited time available for this study, the number of interviews has been limited to eight.

[1] Dutch: rivierkundige, rivierkundig adviseur

Every interview covered the same subjects, so that answers can be compared. The main goal of the interviews was to determine which categories to use on the pilot website. Other topics discussed are the current state of knowledge disclosure and the interviewees' thoughts on a new website for disclosing river knowledge. These subjects are not directly linked to the research questions but provide context to the answers given and aid in providing recommendations at the end of this study. The results of the interviews are discussed below.

Current state of knowledge disclosure

Four out of eight interviewees regularly share reports on Kennisplein, while one asks someone else to share it for them. Two also share river knowledge on *Helpdesk Water* in some cases. Helpdesk Water is a website maintained by RWS WVL, which is used to publicly disclose knowledge within the fields of water management and policy (Rijkswaterstaat, 2020). Out of the three people who do not disclose river knowledge, two work at ON and one at WVL. This could indicate that disclosing river knowledge is less common within ON than within WVL. This is possibly because WVL is more involved in the production of river knowledge, while ON is more involved in the management and maintenance of the river system. The interviewees who do not share river knowledge indicated that it is not clear to them when reports should or should not be disclosed.

Five out of eight interviewees regularly search for knowledge on Kennisplein and/or Helpdesk Water. It was pointed out that documents on Kennisplein are only findable if one has extensive knowledge about the subject matter. If this is not the case, Kennisplein yields irrelevant search results (Buiteveld, 2020), (Schoor, 2020). Furthermore, older reports taken from the paper archives have all been uploaded to Kennisplein but have often been scanned in low quality or have been labelled incorrectly, making them difficult to find (Kroekenstoel, 2020). This confirms that Kennisplein is currently not working as desired. Newer reports containing river knowledge are not always shared by colleagues (Schielen, 2020). Some prefer to search on themes rather than specific documents, which works better on Helpdesk Water (Kater, Interview Emiel Kater, 2020), (Van Putten, 2020).

Interviewees agree that the current process of knowledge disclosure is not ideal. A recurring answer is that all produced knowledge should be available, which is currently not the case. Employees are not always aware of the knowledge already available, resulting in the same research being done twice (Van Putten, 2020). A better connection between the different sections of RWS can improve the situation, e.g. by presenting recently conducted studies to other sections. (Van Putten, 2020). It was acknowledged that not sharing river knowledge could be a 'cultural' problem, meaning that employees are not accustomed to it because people in their working environment are also not used to sharing river knowledge (Kater, Interview Emiel Kater, 2020). This may be part of the reason that river knowledge is not as frequently shared at ON as at WVL.

Facet structure

The interviewees agree that the facet structure is a useful method of conveying river knowledge. One interviewee proposed such a structure by themselves, while others tend to personally organise river knowledge using tree structures or without a specific structure in mind. After having the facet structure explained to them, interviewees unanimously agreed that it is useful. Especially since reports can belong to more than one category, the facet structure is applicable because it is able to represent overlap. An advantage of the facet structure is that it enables to search for knowledge using themes, rather than having to search for specific documents (Kater, Interview Emiel Kater, 2020), (Van Putten, 2020). Such a structure does require a certain level of discipline from the uploaders to consistently label documents correctly (Buiteveld, 2020), (Kroekenstoel, 2020).

Categories to use

The question which categories to use resulted in varying answers. Through the interviews, it became clear that everyone has their own preferences. River consultants tend to use physical attributes like vegetation, roughness, morphology, and hydraulics to categorise knowledge. The ecologist that was interviewed noted that not all documents can be categorised using the proposed categories (Schoor, 2020). For example, studies on sustainability cannot be categorised using just the main functions and geographic locations. This illustrates that people working in different fields have different interests regarding knowledge organisation. If more people with different backgrounds or people outside RWS had been interviewed, this likely would have resulted in different answers. Regardless of how many people are interviewed, it is inevitable to make compromises when deciding which categories to use when organising river knowledge.

Six out of eight interviewees agree on the use of the four main functions. These functions are already in use and form the core of river management in RWS (Kater, Interview Emiel Kater, 2020). Two interviewees prefer to have morphology & sediment management as a separate category. This is because studies are done with morphology & sediment management as the main subject (Schielen, 2020). Six out of eight interviewees think the distinction between geographic areas is a good way to categorise documents. This can be done by making the distinction between Meuse and Rhine branches. Two Rhine branches that should be added to complete the set-up are Bovenrijn and Pannerdens Kanaal (Kater, Personal conversation with Emiel Kater, 2020). Two interviewees proposed the distinction between upper and lower river regions to further specify geographic areas. The distinction between the upper and lower river region is that the water levels in the upper river region are not affected by the sea level, while the water levels in the lower river region are affected by the sea level (Rijkswaterstaat, 2020). The proposed knowledge types (system knowledge, models and instruments, monitoring, measures/interventions) currently used by the Platform River Knowledge are not clear to many interviewees, but they could be a useful way of narrowing down results (Schielen, 2020).

It can be difficult to serve the varying types of users (RWS employees, scientific community, market parties), as different users have different knowledge demands (Kroekenstoel, 2020). However, this can be countered in several ways. Labels indicating scientific research, policy documents and other reports can help to send different users in the right direction (Buiteveld, 2020). Adding research data in addition to the main report could be helpful to members of the scientific community (Van Putten, 2020). Providing a different entrance to the website for each user type was proposed as a solution, but this implies that a different structure and/or set of categories must be applied for each entrance and its respective user type (Kater, Interview Emiel Kater, 2020). Labels distinguishing knowledge institutes could also solve this problem, when users know which institute produced the knowledge they are looking for (Sieben, Interview Arjan Sieben, 2020).

The main trade-off in building a faceted taxonomy is deciding on the number of categories to be used. A small number of categories means that search results cannot be narrowed down extensively, while adding categories enables for more specific searching. The risk of using more specific categories is that over time these might change, or other categories may need to be added in order to maintain the same level of specificity (Blaas, 2020). This means that already labelled documents must be relabelled when categories are changed (Kater, Interview Emiel Kater, 2020). More specific categories work better in the short term, but only using broader categories ensures that these categories remain relevant in the future (Schoor, 2020).

If specific categories are no longer relevant, they could be simply removed from the website. However, this means that documents may no longer be found through the same way as before. Another downside

of using more categories is that applying the right labels requires more effort from the uploader, possibly discouraging them to disclose knowledge (Buiteveld, 2020). Even with a functioning text categorisation algorithm, the uploader will still need to check the categories predicted by the algorithm. An alternative to adding more categories is using keywords to identify documents (Van Zetten, 2020). The keywords can be predefined, with the uploader picking keywords relevant to content of the document (Schielen, 2020).

An overview of the categories proposed by the interviewees is given in **Table 4.4**.

Table 4.4 – Categories proposed by interviewees (indicated in blue)

	D.K.	H.B.	E.K.	M.S.	D.v.P.	R.v.Z.	R.S.	A.S.
Four main functions								
Morphology & sediment management								
Geographic area (Meuse & Rhine branches)								
Physical attributes								
Scientific research/policy documents/other								
Upper/lower river region								
Data								
Institutes								

Thoughts on new website

The interviewees agree that the current process of knowledge disclosure should be improved, but they would rather see a new structure implemented into an existing website than to have a new website introduced. People will be more inclined to disclose river knowledge if they can use a website they are already familiar with (Buiteveld, 2020). Furthermore, it can be beneficial to have all knowledge, broader than just the river domain, in one location (Van Putten, 2020), (Van Zetten, 2020). If every knowledge domain ends up getting its own website, that would become counterproductive (Kater, Interview Emiel Kater, 2020), (Van Putten, 2020). A new website should only be introduced if it has added value for most users (Kroekenstoel, 2020), and clear agreements must be made on when to share knowledge using this website (Kater, Interview Emiel Kater, 2020). Implementation of a new knowledge structure into an existing website is outside the scope of this study but is briefly discussed in **Section 9**.

4.3 Knowledge structure alternatives

Based on the literature review and the conducted interviews, four knowledge structure alternatives have been created. Each alternative is presented below. This section describes the set-up of the alternatives used; the conceptual pilot website with visual examples is provided in **Section 6**.

Alternative A

This alternative is the simplest. It features the main functions, geographic areas in the form of the Meuse and Rhine branches and tags and metadata to further narrow down search results. It has been chosen because most interviewees agree on the use of these categories. The set-up of this alternative is shown in **Table 4.5**.

Table 4.5 – Set-up of Alternative A

Main function	Geographic area	Tags/metadata
All	All	Title
Water safety	Meuse	Author(s)/team
Freshwater availability	Rhine - Bovenrijn - Waal - IJssel - Pannerdens Kanaal - Nederrijn-Lek	Keywords
Navigability over water		Place of publication
Water quality and nature		Date of publication (DD-MM-YYYY – DD-MM-YYYY)
Other	Other	

When searching, the website user can select one or multiple of the main functions and then the relevant river or river branch. If the user is not sure which main function is relevant to the document(s) they are looking for, they can select 'all', so that no documents will be excluded from the search results. If a document is not linked to one of the main functions, the user can select 'other'. The same principle applies to the geographic areas. The user can select 'Rhine' if they want to cover all Rhine branches or they can select individual branches.

In the tags/metadata column, the user can type the title of the document they are looking for. Similarly, they can search for the author(s) or the team that produced a report and the place of publication. They can choose from a list of predefined keywords relevant to the content of the document. The title, author, place of publication, and keywords are not used to narrow down search results like the categories, but can be used to make the most relevant search results appear first. The user can specify a period for the date of publication, which can be used to narrow down search results like the categories.

The advantage of this alternative is its simplicity. The main functions are already in use and the geographic areas are well-defined. This alternative demands less effort from the uploader, as they will have to label the documents with fewer categories than the other alternatives. The downside is that users cannot search as specifically compared to the other alternatives.

Alternative B

This alternative uses more categories than Alternative A, as can be seen in **Table 4.6**. The principle is the same, but the category 'morphology and sediment management' has been added. This is because studies are conducted with morphology and sediment management as the main subject, and users will want to search using this theme (Schielen, 2020). Note that morphology and sediment management is technically not a main function of the river system as defined in BPRW 2016, but an overarching theme that touches on every main function. This is why it has been added in a separate column to the main functions. Furthermore, a facet indicating different document types has been added.

Table 4.6 – Set-up of Alternative B

Main function		Geographic area	Document type	Tags/metadata
All		All	All	Title
Water safety	Morphology and sediment management	Meuse	Scientific research	Author(s)/team
Freshwater availability		Rhine - Bovenrijn - Waal - IJssel - Pannerdens Kanaal - Nederrijn-Lek		Keywords
Navigability over water			Policy documents	Place of publication
Water quality and nature				Date of publication (DD-MM-YYYY – DD-MM-YYYY)
Other		Other	Other	

This alternative allows for more specific searching than alternative A. The distinction between scientific research, policy documents and other types can help to serve the different user types looking for different types of information. Correctly labelling documents will require more effort from the uploader compared to alternative A.

Alternative C

This alternative uses an extra facet compared to Alternative B, as can be seen in **Table 4.7**. The geographic area can be further specified by distinguishing between the upper and lower river region.

Table 4.7 – Set-up of Alternative C

Main function		Geographic area		Document type	Tags/metadata
All		All		All	Title
Water safety	Morphology and sediment management	Meuse	Upper river region	Scientific research	Author(s)/team
Freshwater availability		Rhine - Bovenrijn - Waal - IJssel - Pannerdens Kanaal - Nederrijn-Lek			Keywords
Navigability over water			Lower river region	Policy documents	Place of publication
Water quality and nature					Date of publication (DD-MM-YYYY – DD-MM-YYYY)
Other		Other		Other	

The distinction between the upper and lower river region allows for more specific searching than alternatives A and B. The upper and lower river regions are often used within RWS but have no clearly defined borders (Rijkswaterstaat, 2020). For users outside RWS, this distinction might not be clear. Furthermore, correctly labelling documents will require more effort from the uploader compared to alternatives A and B.

Alternative D

This is the most elaborate alternative, as can be seen in **Table 4.8**. A facet indicating different knowledge institutes has been added. This distinction allows users to further narrow down search results. It can also help to serve different types of users, when they know which knowledge institute produced the knowledge they are looking for.

Table 4.8 – Set-up of Alternative D

Main function		Geographic area		Document type	Institute	Tags/metadata
All		All		All	All	Title
Water safety	Morphology and sediment management	Meuse	Upper river region	Scientific research	RWS	Author(s)/team
Freshwater availability		Rhine - Bovenrijn - Waal - IJssel - Pannerdens Kanaal - Nederrijn-Lek			Deltares	Keywords
Navigability over water					KNMI	Place of publication
Water quality and nature			Lower river region	Policy documents	Univer-sities	Date of publication (DD-MM-YYYY – DD-MM-YYYY)
Other		Other		Other	Other	

Note that this is only a selection of knowledge institutes. Consultancy companies that also produce river knowledge have not been included and can be found under 'other'. Further specification of RWS sections and individual universities has not been made, as this would become too specific (Sieben, Personal conversation with Arjan Sieben, 2020).

This alternative allows for the most specific searching, but labelling documents using this set of categories requires the most effort out of the four alternatives. Another disadvantage is that applying so many categories means the structure is the least flexible to future changes in the categories used.

General remarks

Although proposed by five out of eight interviewees, physical attributes such as vegetation, soil position and roughness, have not been used in the four alternatives. These attributes are a useful way of categorising knowledge for river consultants and have been proposed only by river consultants. However, these attributes are too specific to use on a website with a broad application. Furthermore, there are too many different physical attributes that could be used to categorise knowledge (Kroekenstoel, 2020).

The knowledge types (system knowledge, models and instruments, monitoring, measures/interventions) currently used by the Platform River Knowledge have not been used in the alternatives. This is because these categories are not clear to most interviewees and not relevant in many cases.

Specific examples of keywords that can be used have not been discussed with interviewees. This could have been a useful addition to the subjects covered. The idea to use title, author(s), place of publication and keywords to make relevant search results appear first is not demonstrated in this study. However, there are tested methods that have been proven to work for this purpose (Lambe, 2007). These methods are not covered in this research.

Only four alternatives have been presented, but other combinations can be made using these categories, and other categories may be used. It might be possible to compile a different set of categories that is more applicable for a knowledge disclosure website. Based on the limited research done, these alternatives are viable options, but may be biased towards the preferences of river consultants. An overview of the strengths and weaknesses of each alternative is provided in **Table 4.9**, based on four criteria.

Table 4.9 – Overview of strengths and weaknesses of each alternative

	A	B	C	D
Specific searching				
Easy labelling				
Likelihood of relevance in future				
Serves different user types				

Legend
Positive
Negative

5. Naïve Bayes Model

A probabilistic model has been created based on the Naïve Bayes algorithm. This model is applied to the categories defined in **Alternative A** to serve as a proof of concept. The theory, set-up and results of the model are discussed in this section, providing the answer to **Sub-question 3**.

Note: *the term 'keyword' is a central term in this section, but it has a different meaning in the NB model than in the alternative set-ups. In the previous section, a keyword is a word that can be connected to a document to make relevant search results appear first. In this section, a keyword is a word* ***occurring in the text*** *that can indicate a document belonging to a certain category.*

5.1 Theory of Naïve Bayes algorithm

A Naïve Bayes classifier is a probabilistic machine learning model that can be used for text categorisation. The goal of this model is to predict the probability that a document belongs to a certain category, based on the occurrence of keywords in the document text. This model can then be used to sort documents into the correct categories.

The model is based on Bayes' theorem (Gandhi, 2020). Applied to the problem context, the theorem is expressed as **Equation 5.1**.

$$P(C_j \mid w_1, \dots, w_n) = \frac{P(w_1, \dots, w_n \mid C_j)\, P(C_j)}{P(w_1, \dots, w_n)} \qquad \textbf{Equation 5.1}$$

Where:

- $P(C_j \mid w_1, \dots, w_n)$ *is the probability of a document belonging to category* C_j*, given the occurrence of keywords* $w_1, \dots, w_n$ *in the document*
- $P(w_1, \dots, w_n \mid C_j)$ *is the probability of keywords* $w_1, \dots, w_n$ *occurring in a document, given that it belongs to category* C_j
- $P(C_j)$ *is the probability of a document belonging to category* C_j
- $P(w_1, \dots, w_n)$ *is the probability of keywords* $w_1, \dots, w_n$ *occurring in a document*

Note that $P(w_1, \dots, w_n)$ is constant for all entries in the dataset, so the numerator can be left out when comparing the probabilities that a document belongs to a category. This results in the equivalence of **Equation 5.2** (Scikit learn, 2020). We are no longer speaking of probabilities, but prediction values.

$$P(C_j \mid w_1, \dots, w_n) \propto P(w_1, \dots, w_n \mid C_j)\, P(C_j) \qquad \textbf{Equation 5.2}$$

A Naïve Bayes model makes the 'naïve' assumption of conditional independence between predictors. This means that predictors are assumed to independently contribute to the outcome value, i.e. category (Gandhi, 2020). Because of the conditional independence assumption, $P(w_1, \dots, w_n \mid C_j)$ can be rewritten using the chain rule for probabilities (Scikit learn, 2020). This is expressed in **Equation 5.3**.

$$P(w_1, \dots, w_n \mid C_j) = P(w_1 \mid C_j) * P(w_2 \mid C_j) * \dots * P(w_n \mid C_j) = \prod_{i=1}^{n} P(w_i \mid C_j) \qquad \textbf{Equation 5.3}$$

Combining **Equation 5.2** and **Equation 5.3** results in the equivalence given in **Equation 5.4** (Scikit learn, 2020).

$$P(C_j \mid w_1, \dots, w_n) \propto P(C_j) \prod_{i=1}^{n} P(w_i \mid C_j) \qquad \textbf{Equation 5.4}$$

The main goal of a Naïve Bayes model is to find the category C_j with the highest prediction value, given the frequency of keywords $w_1, \dots, w_n$ occurring in a document. This is conveyed in **Equation 5.5** (Scikit learn, 2020)**.**

$$\hat{C} = argmax_{C_j} P(C_j) \prod_{i=1}^{n} P(w_i \mid C_j) \qquad \textbf{Equation 5.5}$$

Where:

- $\hat{C}$ *is the category that with the highest prediction value*
- $P(C_j)$ *is the probability of a document belonging to category* C_j
- $P(w_i \mid C_j)$ *is the probability of keyword* w_i *occurring in a document, given that it belongs to category* C_j

Since documents in this study can belong to more than one category, a threshold value is required to sort a document into the multiple correct categories. If a prediction value exceeds this threshold value, it will be sorted into the category linked to this prediction value. If such a threshold value can be determined, the function objective can be expressed as **Equation 5.6**. If a threshold value cannot be determined, the model will use **Equation 5.5** and only sort the document into one category.

$$\hat{C}_j = C_j \quad \leftrightarrow \quad P(C_j) \prod_{i=1}^{n} P(w_i \mid C_j) > m \qquad \textbf{Equation 5.6}$$

Where:

- $\hat{C}_j$ *are the categories with prediction values that exceed the threshold value*
- C_j *are the candidate categories*
- $P(C_j)$ *is the probability of a document belonging to category* C_j
- $P(w_i \mid C_j)$ *is the probability of keyword* w_i *occurring in a document, given that it belongs to category* C_j
- m *is the threshold value for categorisation*

Equation 5.5 and **Equation 5.6** show that in order to categorise documents using NB classifiers, $P(C_j)$ and $P(w_i \mid C_j)$ must first be computed for every category C_j and keyword w_i. These are the model parameters. This is where training data is required. The training data consists of documents that have been categorised correctly and a selection of keywords that indicate each category. The documents and keywords used for training the Naïve Bayes classifier have been identified by several interviewees. A list of the used documents and their respective categories can be found in **Appendix C.1**. All documents and keywords used are in Dutch.

5.2 Model set-up and training

A MATLAB script has been made which extracts text from the provided documents. The keywords are manually inserted into the script, after which MATLAB scans through each document for every occurrence of a keyword. The frequency of the keywords in every document is then written into a matrix, which is exported to Excel. This has been done for a total of 37 documents used as training data and 8 documents used to test the output of the model. The matrices containing keyword frequencies can be found in **Appendix C.2**. The MATLAB script with explanatory comments can be found in **Appendix C.3**.

The model analyses the categories of **Alternative A**: the four main functions and geographic areas. This section explains the set-up and training of the model for the four main functions. The same method is used to create the set-up for the geographic areas in the model. The four main functions are as follows:

- Water safety (WS)
- Freshwater availability (FA)
- Navigability over water (NW)
- Water quality and nature (WN)

The first objective is to compute the probabilities of a documents belonging to each category, i.e. $P(C_j) = P(WS), P(FA), P(NW), P(WN)$ based on the training data. This is done in Excel by dividing the number of documents belonging to each category by the total number of documents, in this case 37. The results are shown in **Table 5.1**.

Table 5.1 – Model parameters $P(C_j)$ based on training data

C_j	$P(C_j)$
WS	$P(WS) = 30/37 = 0.811$
FA	$P(FA) = 9/37 = 0.243$
NW	$P(NW) = 8/37 = 0.216$
WN	$P(WN) = 6/37 = 0.162$

Note that the cumulative probability $\sum P(C_j) > 1$. This is because some documents belong to more than one category, resulting in the total frequency of the categories being larger than the total number of documents. This does not affect the prediction capabilities of the model, since this is independent of the cumulative probability.

Also note that $P(WS)$ is significantly larger than the other values. This means that the model will be trained more accurately for documents belonging to category WS, since there is a larger amount of training data available for this category.

The next step is to compute every conditional probability $P(w_i \mid C_j)$, i.e. the probability of each keyword w_i occurring in a document, given that it belongs to category C_j. This problem features four different categories, making it a multinomial distribution: each trial has $k \geq 2$ possible outcomes (PennState Eberly College of Science, 2020). $P(w_i \mid C_j)$ can then be calculated using **Equation 5.7** (Synced, 2020).

$$P(w_i \mid C_j) = \frac{N_{i,j} + 1}{N_j + n} \quad (i = 1, \dots, n) \qquad \textbf{Equation 5.7}$$

Where:

- $N_{i,j}$ *is the total number of times that keyword* w_i *occurs in all documents belonging to category* C_j
- $N_j = \sum_{i=1}^{n} N_{i,j}$ *is the total frequency of all keywords* $w_1, \dots, w_n$ *occurring in all documents belonging to category* C_j
- n *is the number of different keywords*

The keywords used to train the NB model for the main functions are given in **Table 5.2**. The keywords used to train the model for geographic areas are provided in **Appendix C.1**. More keywords have been provided by the interviewees, but the keywords that do not occur in the training dataset have not been used in the model.

Table 5.2 – List of keywords used to identify main functions

WS		FA		NW		WN	
w_1	Hoogwater	w_{14}	Laagwater	w_{20}	Klasse	w_{26}	Droog
w_2	Risico	w_{15}	Afvoerverdeling	w_{21}	Vaste laag	w_{27}	Uiterwaarde
w_3	Overstroming	w_{16}	Drinkwater	w_{22}	Lading	w_{28}	Watertekort
w_4	Dijk	w_{17}	Verdringingsreeks	w_{23}	Modaliteit	w_{29}	Flora
w_5	Waterstand	w_{18}	Zout	w_{24}	Doorvaarthoogte	w_{30}	Fauna
w_6	Herhalingstijd	w_{19}	Waterverdeling	w_{25}	Vaardiepte	w_{31}	Vegetatie
w_7	Slachtoffers					w_{32}	Chloride
w_8	Schade					w_{33}	Temperatuur
w_9	Overstromingskans					w_{34}	Klimaat
w_{10}	Overschrijdingskans					w_{35}	Stuw
w_{11}	Frequentie						
w_{12}	Afvoer						
w_{13}	Faalkans						

Equation 5.7 has been used to calculate $P(w_i \mid C_j)$ for every keyword w_i and every main function C_j in Excel. The resulting values are given in **Table 5.3**. The same approach has been used to determine the model parameters for the geographic areas. These values can be found in **Appendix C.1**.

Table 5.3 – Model parameters $P(w_i \mid C_j)$ based on training data

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}
$P(w_i \mid WS)$	0.071	0.087	0.120	0.224	0.101	0.006	0.009	0.034	0.011	0.004	0.019	0.149
$P(w_i \mid FA)$	0.022	0.024	0.017	0.038	0.079	0.003	0.001	0.029	0.001	0.001	0.009	0.191
$P(w_i \mid NW)$	0.042	0.032	0.038	0.054	0.107	0.004	0.002	0.053	0.001	0.001	0.009	0.164
$P(w_i \mid WN)$	0.035	0.035	0.029	0.088	0.084	0.002	0.003	0.037	0.001	0.000	0.012	0.128
	w_{13}	w_{14}	w_{15}	w_{16}	w_{17}	w_{18}	w_{19}	w_{20}	w_{21}	w_{22}	w_{23}	w_{24}
$P(w_i \mid WS)$	0.009	0.003	0.011	0.004	0.001	0.002	0.001	0.007	0.000	0.000	0.000	0.000
$P(w_i \mid FA)$	0.000	0.033	0.002	0.069	0.007	0.038	0.011	0.005	0.000	0.002	0.001	0.001
$P(w_i \mid NW)$	0.001	0.022	0.007	0.017	0.005	0.011	0.004	0.011	0.014	0.003	0.002	0.000
$P(w_i \mid WN)$	0.001	0.013	0.002	0.019	0.006	0.012	0.004	0.009	0.001	0.002	0.001	0.000
	w_{25}	w_{26}	w_{27}	w_{28}	w_{29}	w_{30}	w_{31}	w_{32}	w_{33}	w_{34}	w_{35}	
$P(w_i \mid WS)$	0.000	0.012	0.012	0.002	0.001	0.002	0.006	0.000	0.011	0.061	0.018	
$P(w_i \mid FA)$	0.012	0.107	0.007	0.013	0.001	0.003	0.006	0.010	0.032	0.197	0.027	
$P(w_i \mid NW)$	0.003	0.054	0.014	0.008	0.002	0.005	0.003	0.001	0.031	0.243	0.032	
$P(w_i \mid WN)$	0.001	0.064	0.046	0.008	0.004	0.013	0.019	0.002	0.034	0.243	0.042	

The values of the model parameters in **Table 5.3** have been plotted in **Figure 5.1** in order to provide a visual overview of the relevance of each keyword to the main functions.

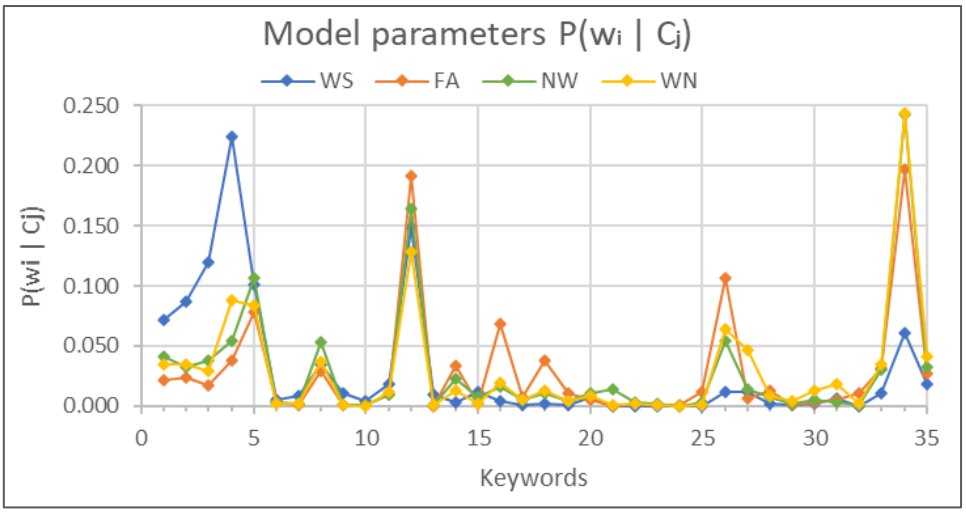


***Figure 5.1** – Visual overview of the relevance of each keyword to the main functions*

Table 5.3 and **Figure 5.1** show that there is significant variance in the relevance of keywords to the main functions. For example, w_{34} *(klimaat)* has a high probability of occurring in documents, especially of categories NW and WN, while w_{24} *(doorvaarthoogte)* has a very low probability of occurring in any category. Some keywords, such as w_5 *(waterstand)*, w_8 *(schade)* and w_{12} *(afvoer)* have a significant probability of occurring in a document, but the probabilities are relatively close for each category. This means that these keywords are not ideal for identifying main functions because they do not indicate strongly to any specific function.

Keywords w_4 *(dijk)*, w_{16} *(drinkwater)* and w_{26} *(droog)* are examples of ideal keywords because they strongly indicate towards a specific function and have a significant probability of occurring in a document. Ideally, only such keywords are used when creating a text categorisation model based on keywords. Due to the limited time available for this study, the provided keywords have been used instead of finding more ideal keywords.

In some cases, the model does not assign the highest probability to the expected category. For example, w_8 *(schade)* has the highest probability of occurring in NW according to the model, while it has been identified as a keyword for WS by the interviewees. Similarly, w_{26} *(droog)* has been assigned to FA, while it has been identified as a keyword for WN. This discrepancy is possibly caused by a lack of training data. In general, the model has assigned the highest probability to the same category as the interviewees.

All the observations mentioned above are based on a limited dataset and chosen keywords which are not ideal. This means that these observations are not conclusive and should not be taken as facts. However, it is important to keep these things in mind when developing or analysing a Naïve Bayes model or its results.

Now that the model parameters $P(C_j)$ and $P(w_i \mid C_j)$ have been computed, the model has been trained and new documents can be categorised using this model, as $P(C_j)\prod_{i=1}^{n} P(w_i \mid C_j)$ can be determined.

5.3 Results

Eight documents have been scanned using MATLAB for the frequency of each keyword occurring. These keyword frequencies have been used as input for the model. The documents and their categories are given in **Table 5.4**. This set of documents has been chosen so that every main function and geographic area occurs at least once in the dataset, and there is some overlap in the categories.

Table 5.4 – Documents used as input data

Nr.	Title	Function(s)	Geographic area(s)
1	Achtergrondreportage Beleidstafel Droogte	FA, WN	All
2	Verlaging van de waterstanden in het Albertkanaal juli 2011	FA	Meuse
3	Waterstandsanalyse in het kader van het onderzoek Optie Aa en Maas	WS	Meuse
4	Ruimte voor de Waal – Het plan	WS	Rhine
5	MIRT Onderzoek Duurzame Bodemligging Rijntakken Eindrapportage	NW	Rhine
6	Handreiking sedimentbeheer nevengeulen	WS, WN	Rhine
7	Kennisgeving M.E.R. Cortenoever, Tichelbeeksewaard en Voorsterklei	WS	Rhine
8	Masterplan Eilanden 3.0 Stadsblokken Meinerswijk	WS	Rhine

Note that for the geographic areas, only Meuse and Rhine have been specified, leaving out the individual Rhine branches. This has been done because the provided keywords indicating the different Rhine branches are not sufficiently specific for identifying these categories. In combination with the small amount of training data, this makes the model unable to predict geographic areas that include the Rhine branches. By only using the Meuse and Rhine as geographic areas, the model is able to make more accurate predictions, albeit less specific.

The model output is the prediction value $P(C_j)\prod_{i=1}^{n} P(w_i \mid C_j)$ for keywords $w_1, \dots, w_n$ and categories C_j. These prediction values have been determined for every document and category, and the category with the highest value is chosen for each document. The results for the main functions are shown in **Table 5.5** and the results for the geographic areas are shown in **Table 5.6**. Blue cells indicate the highest prediction value per document. Green cells indicate correct predictions, while yellow cells indicate partially correct predictions (where one of two categories is correctly predicted) and red cells indicate incorrect predictions.

Table 5.5 – Model output for main functions

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$P(WS)\prod_{i=1}^{n} P(w_i \mid WS)$	3.10* E-26	1.84* E-08	1.08* E-07	2.24* E-09	6.62* E-30	2.74* E-10	6.90* E-05	5.92* E-08
$P(FA)\prod_{i=1}^{n} P(w_i \mid FA)$	2.06* E-17	2.05* E-07	6.98* E-09	2.64* E-10	5.98* E-27	9.19* E-14	1.42* E-07	3.51* E-09
$P(NW)\prod_{i=1}^{n} P(w_i \mid NW)$	7.11* E-20	1.08* E-07	7.79* E-09	1.82* E-09	1.24* E-23	5.58* E-12	2.85* E-06	5.86* E-08
$P(WN)\prod_{i=1}^{n} P(w_i \mid WN)$	8.77* E-19	5.48* E-08	1.31* E-08	6.10* E-08	5.39* E-25	5.37* E-11	1.16* E-05	1.58* E-07
Correct category	**FA, WN**	**FA**	**WS**	**WS**	**NW**	**WS, WN**	**WS**	**WS**
Predicted category	**FA**	**FA**	**WS**	**WN**	**NW**	**WS**	**WS**	**WN**

Table 5.6 – Model output for geographic areas

	1	2	3	4	5	6	7	8
$P(WS)\prod_{i=1}^{n} P(w_i \mid Meuse)$	6.49* E-09	5.17* E-01	5.78* E-01	3.32* E-05	6.33* E-08	1.78* E-06	7.19* E-03	1.46* E-02
$P(FA)\prod_{i=1}^{n} P(w_i \mid Rhine)$	1.00* E-09	1.95* E-01	1.77* E-01	3.18* E-06	2.73* E-04	8.16* E-05	1.34* E-02	1.08* E-01
Correct category	**All**	**Meuse**	**Meuse**	**Rhine**	**Rhine**	**Rhine**	**Rhine**	**Rhine**
Predicted category	**Meuse**	**Meuse**	**Meuse**	**Meuse**	**Rhine**	**Rhine**	**Rhine**	**Rhine**

Table 5.5 and **Table 5.6** show that the model has incorrectly predicted the main function of two documents and the geographic area of one document. Because only the highest prediction value is taken for each category, the documents that belong to more than one category have been categorised partially correctly. Note that in each of these cases, the second highest value is connected to the other category the document belongs to. This means that the predictions are correct, but the model does not know when to sort documents into more than one category. A threshold value m as defined in **Equation 5.6** has not been found. This is because the prediction values per document differ vastly (see **Table 5.5** and **Table 5.6**), making it impossible to determine a single value above which the documents should be sorted into the connected categories.

Other approaches to enable categorisation into multiple categories have been considered. For example, if the second highest value falls within a certain range of the highest, the document could be sorted into the category with the second highest value as well. However, this does not produce useful results because some documents will be sorted into a category they do not belong to, while others are not sorted into a category they do belong to. For example, applying a factor 10 for highest to second highest value, **Document 2** in **Table 5.5** would be sorted into category NW because $\frac{2.05E-07}{1.08E-07} = 1.90 < 10$, while it does not belong to NW. In the same manner, **Document 1** in **Table 5.5** would not be sorted into WN because $\frac{2.06E-17}{8.77E-19} = 23.46 > 10$, while it does belong to WN.

It appears that the only way to enable automatic categorisation into multiple categories is through the use of a threshold value that scales with the prediction values connected to a document. A formula to determine such values has not been found. This might be due to the fact that although NB is a decent classifier, it is known to be a bad estimator (Scikit learn, 2020). This means that the prediction values produced by the NB algorithm are not very accurate. Because these values are inaccurate, it is difficult to do any statistically significant analysis using these values.

If the overlapping categories are disregarded, there are 6 out of 8 correct predictions for the main functions and 7 out of 8 correct predictions for the geographic areas. This suggests that Naïve Bayes can be a useful method to automatically categorise river knowledge. However, the uploader should always check if the suggested categories are correct to make sure that a document is labelled correctly.

Since this model has been trained using a training dataset of only 37 documents, the results may not be accurate. A larger set of training data is required to draw a more definitive conclusion about the accuracy of the NB model. If the pilot website is launched and documents are disclosed using the website, each time a new document is uploaded and the correct categories are selected, this will serve as an addition to the training dataset. The additional training data can then be used to improve the accuracy of the model and pick only the most useful keywords to be used in the model. Through this iterative process, the accuracy of the model can be continually improved.

In conclusion, the concept of a Naïve Bayes model has been tested and the results suggest that it can be a useful method for categorising river knowledge. Although the NB model is unable to sort documents into multiple categories, it appears to be a decent classifier. Because the NB method has only been applied to a simplified version of Alternative A and is based on a limited amount of training data, the proof of its functional potential is considered insufficient.

6. Conceptual Pilot Website

In this section, visual examples of the conceptual pilot website are provided. This is meant to provide an idea of what a pilot website for disclosing river knowledge might look like and how it works, providing the answer to the **Main Research Question**.

6.1 Searching for river knowledge

Figure 6.1 shows a visual example of the pilot website based on **Alternative A**, which can be used to search for river knowledge on the website. The pilot website features instructions on how to search for river knowledge, so new users will understand the method. The principle is the same for other alternatives.

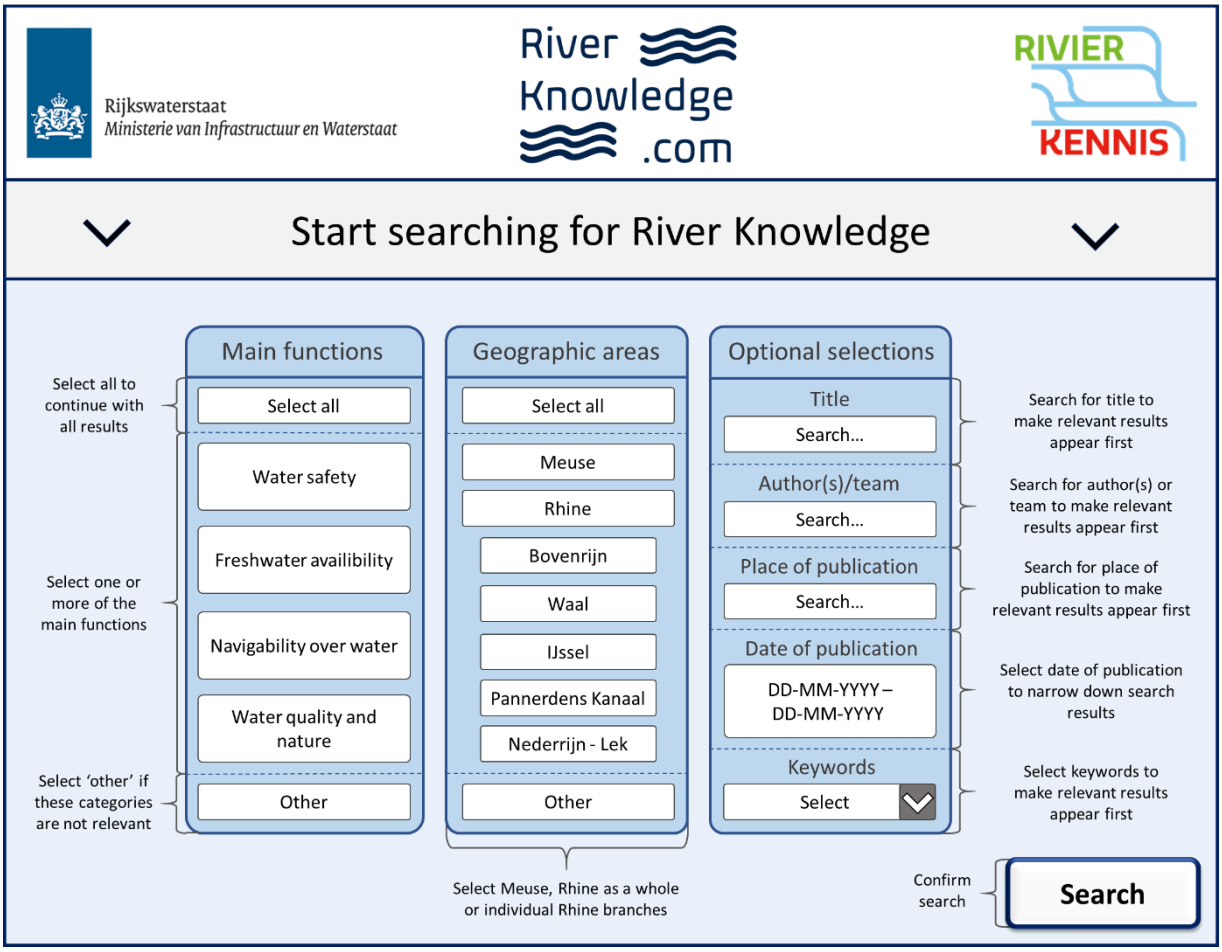


Figure 6.1** – Visual example of pilot website used to search for river knowledge, based on **Alternative A

The first two columns in **Figure 6.1** are mandatory to use when searching for river knowledge. In the first column, the user can select one or multiple main functions relevant to the knowledge they are looking for. If the knowledge they are looking for cannot be categorised using the main functions, they should select 'other'. If every main function is applicable, the user should choose 'select all'. This option can also be used when the user is not sure which categories are relevant, so that no documents are excluded from the search results. The second column features geographic areas and operates in the same way as the first column. The user can select Meuse, the Rhine as a whole, or individual Rhine branches. The options 'select all' and 'other' are used in the same way as in the first column.

The third column provides additional optional selections. Searching for title, author(s)/teams, and place of publication can help to make relevant search results appear first. A period can be specified within which the document was published, which will help to narrow down search results. Lastly, the user can select several keywords from a dropdown menu to make search results linked to these keywords appear before other search results. After the relevant selections have been made, the user can confirm their search.

6.2 Disclosing river knowledge

Figure 6.2 shows a visual example of the pilot website based on **Alternative A**, which can be used to disclose river knowledge on the website. The pilot website features instructions on how to label documents, so new users will understand the method. The principle is the same for other alternatives.

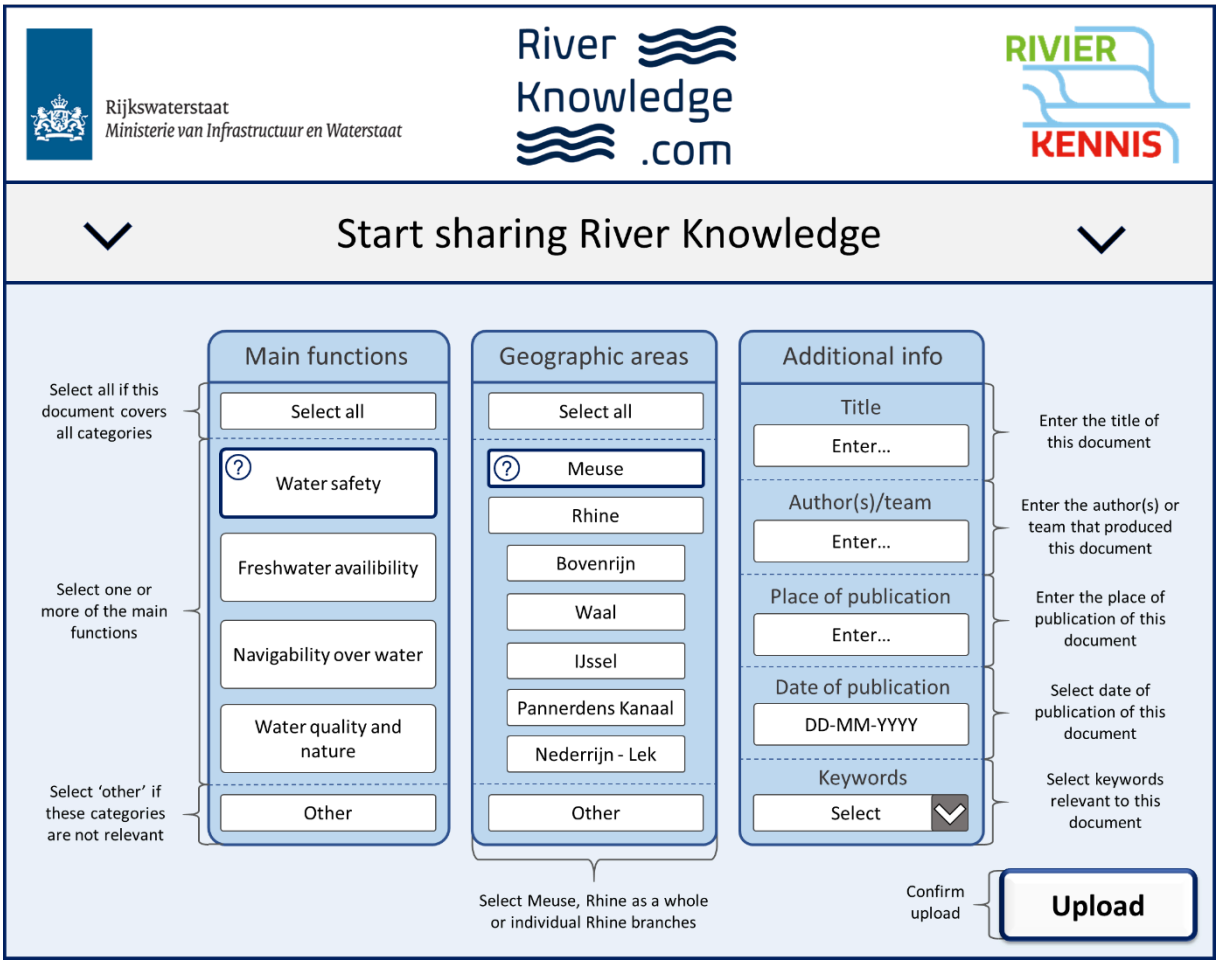


Figure 6.2** – Visual example of pilot website used to disclose river knowledge, based on **Alternative A

When disclosing river knowledge, the uploader needs to correctly label documents. This is done by selecting the relevant function(s), geographic area(s), and additional info. Keywords can be selected from a list of predefined keywords. When documents are uploaded, they are automatically categorised based on the NB model, which will serve as a suggestion to the uploader. In **Figure 6.2**, an example is given where the NB model provides 'water safety' and 'Meuse' as suggestions, indicated by a question mark icon.

7. Discussion

In this section, the results and the process of the research are reflected upon. This provides necessary context to the conclusions provided in the next section.

Literature review

A facet structure has been chosen as the taxonomy form to be used on the pilot website. This choice was made after conducting a literature review. The source for this literature review was mainly one work: ''Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness'' (2007) by Patrick Lambe. Using predominantly one work as the source for the literature review may have caused a bias in the options considered for organising river knowledge. More sources could have been consulted, which might have resulted in a different approach to organising river knowledge.

Although mainly one source was used, this work is often cited in the field of knowledge management, suggesting that it is a reliable source. In the end, using this work as the main source did not have any negative effects, as the interviewees unanimously agreed that the facet structure is a useful method for organising river knowledge.

Interviews and chosen alternatives

Eight interviews have been conducted in order to establish how RWS employees prefer to organise river knowledge. This has resulted in four alternative sets of categories which can be used to organise river knowledge on the pilot website. However, more combinations can be made using the proposed categories. As such, the given alternatives should be treated as suggestions and not as the only viable options.

The number of interviews conducted ideally would have been larger, so that alternatives could have been created with a wider support base. Since the pilot website is meant to be used by a target audience which is broader than just RWS, it would have been beneficial to interview other types of users about their preferences as well. This could have resulted in different categories being proposed. However, interviewing people outside RWS was not within the scope of this study.

Six out of the eight interviewees were river consultants. The homogenous character of the sample of interviewees may have caused a bias in the proposed alternatives. It was not a conscious decision to interview mainly river consultants, but a combination of people already spoken to, people who were available, and people put forward by a contact person. In hindsight, more care could have been taken in the decision who to interview, to ensure a more diverse group of interviewees. This could have partly prevented a possible bias in the interview results.

Due to the qualitative nature of the interview subjects, the interpretation of the results is not exact. In some cases, similar answers by different interviewees have been treated as the same in order to provide the visual overview in **Table 4.4** and the proposed alternatives. For example, different types of physical attributes have been proposed, and they have been treated as one category in the overview. Because the answers given by interviewees are subjective and based on personal experience, interpretation of the results is necessary in order to draw generalised conclusions.

The interviewees would rather see a new structure implemented into an existing website than to have a new website introduced for disclosing river knowledge specifically. This is because people will be more inclined to disclose river knowledge using a method they are already familiar with. It was pointed out by

interviewees that a new website should only be introduced if it has added value to most users. This means that further research into the demand for such a website must be done before it can be realised.

Naïve Bayes model

A probabilistic model based on the Naïve Bayes algorithm has been made which can sort documents into predefined categories based on keywords occurring in the text. The model created for this study was meant to serve as a proof of concept. More extensive modelling is required to create a fully functioning NB model. In particular, the MATLAB and Excel parts should be integrated into one model.

The created NB model is limited in a number of ways. Firstly, the model has been made using only documents and keywords in Dutch. Some reports at RWS are in English, so the model is unable to categorise these. Secondly, the model has been made using a limited amount of training data. This means that predictions made by the model may not be accurate. Furthermore, not all keywords used turned out to be relevant. Some are not indicative of any category, while others rarely occur in the documents used. With a larger training dataset, results will become more accurate and the most useful keywords can be picked.

Another important limitation of the NB model is that it can only assign a document to one category, while documents can belong to multiple categories. This is because the algorithm only chooses the category with the highest prediction value. Several ways to counter this problem have been considered, but no solution has been found. Further research is required to find a possible solution to this problem.

A sensitivity analysis on the NB model has not been done. This could have provided more insight into the accuracy of the predictions. A sensitivity analysis could have been done by removing certain keywords from the model or by changing the keyword frequencies in Excel and seeing how the predictions change compared to the initial predictions.

The choice for a Naïve Bayes model was made in the proposal phase of this study based on a brief literature review. This could have been done more thoroughly and might have resulted in a different algorithm being applied. However, considering the limited amount of training data available, using the simplest algorithm appears to have been the right choice.

COVID-19 regulations

Due to the regulations surrounding the COVID-19 pandemic, the research could not be done as initially planned. The Rijkswaterstaat offices and University of Twente remained closed over the course of the research, so the entire study was done from home. This impacted the research in several ways.

All contact was done through e-mails, phone calls and online meetings. This made it more difficult to become acquainted with colleagues at the department and with Rijkswaterstaat as an organisation. Questions were not always quickly answered through e-mails because of the delayed response and Skype meetings did not always fit into people's schedules. Phone calls were usually the best option for brief questions. Although the situation sometimes made contact more difficult, colleagues at RWS were generally very helpful.

Not being physically present at the organisation has possibly made it more difficult to acquire an understanding of how this research fits into the broader goals of RWS. Working from home did not allow for easy conversation with colleagues about their views surrounding the research topic. The only way in which direct views on the pilot website were obtained was through the interviews conducted. Nevertheless, it was possible to answer every research question.

8. Conclusion

The main goal of this study was to set up a conceptual version of a pilot website for disclosing river knowledge. This has been done by determining which knowledge taxonomy form is most suitable for such a website, how river knowledge should be categorised according to RWS employees, and by creating a Naïve Bayes model that is able to categorise documents based on their content text.

Based on literature, a facet structure was determined to be the most applicable taxonomy form to use when organising river knowledge. This is because facets work well with large amounts of content and with overlapping categories, which is often the case when dealing with river knowledge. Furthermore, a facet structure is especially effective in a digital environment, where items can easily be stored in multiple locations (i.e. categories) and the use of tags and metadata to further specify search results can be accommodated. The choice for a facet structure was approved by interviewees, who unanimously agreed that it is a useful method for organising river knowledge.

Four alternative set-ups have been made which can be used to organise river knowledge, based on interviews conducted with RWS employees. These set-ups vary from broad to specific, each with different advantages and disadvantages. The set-ups that only use broader categories require less effort from the uploader to correctly label and are more likely to remain relevant in the future because the categories are broader themes. The downside is that this does not enable specific searching and the broad categories may not serve every different user type. The reverse is the case for set-ups with increasingly specific categories. Other combinations of categories can be made, so the given alternatives should be treated as suggestions.

A text categorisation model has been made based on the Naïve Bayes algorithm. This model has been applied to a simplified version of the first alternative (four main functions and the distinction between Meuse and Rhine). For an input of 8 documents, the model has correctly predicted the main function 6 out of 8 times and the geographic area 7 out of 8 times. However, it is unable to assign documents to multiple overlapping categories. Due to the small amount of training data available and the simplicity of the categories used, the proof of the method's functional potential is considered insufficient. This means that it can be beneficial to use this method to automatically categorise river knowledge, but the uploader should always check to make sure that documents are labelled correctly.

Although not covered by the research questions, it became clear during the interviews that the demand for a pilot website exclusively for disclosing river knowledge is not high among RWS employees. People will be more inclined to share river knowledge if they can use a method that they are already familiar with and it is preferable to have all knowledge, broader than only river knowledge, in one location. However, the interviewees agreed that the current state of knowledge disclosure should be improved.

To conclude, the main research question has been answered by providing a conceptual version of the pilot website for disclosing river knowledge. This has been done for the perspective of searching for river knowledge and the perspective of disclosing river knowledge. The website set-up is based on a facet structure, categories proposed by interviewees and a Naïve Bayes text categorisation model.

9. Recommendations

Based on the conclusions and the limitations of this research, some recommendations to the Platform River Knowledge can be made regarding potential follow-up research on this topic. The recommendations are provided in this section.

- The demand for a website for disclosing river knowledge should be further investigated before continuing development. Such a website only works well if it is widely used, which will only happen if it has added value to most users. If there is insufficient demand for a new website among stakeholders, a launch may turn out unsuccessful.
- Implementation of a new structure into an existing website like Kennisplein or Helpdesk Water is a preferred solution by most interviewees. Exploring the possibilities in this regard may be worthwhile.
- Whether a new website is launched or a new structure is implemented into an existing website, using a facet structure is recommended. The consulted literature and the interviewees agree that this taxonomy form is a useful way of organising river knowledge.
- More extensive interviewing is advised before introducing a new structure for disclosing river knowledge. The alternatives provided in this study are based on a limited sample of interviewees skewed towards river consultants. Interviewing a broader group of stakeholders, ideally including all user types, is recommended. This may result in other categories being proposed, but ensures broader support among users.
- Further modelling and data collection are required to create a functioning Naïve Bayes model that can be used to automatically categorise documents containing river knowledge. Finding a method that enables categorisation into multiple categories will be especially valuable. Research into other more accurate text categorisation methods should be considered as well.
- It is important that river knowledge is consistently shared. RWS employees dealing with river knowledge could be more actively encouraged to start sharing river knowledge. This is already known within the Platform, but launching a new website may not be useful before river knowledge is more actively shared.

10. References

Asselman, N., Barneveld, H., Klijn, F., & Van Winden, A. (2018). *Het verhaal van de Maas.* Rijkswaterstaat.

Barriball, K., & White, A. (1994). Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced nursing*, 329.

Bhardwaj, A. K. (2020, April 28). *Code Project*. Retrieved from https://www.codeproject.com/Articles/614898/Tree-Organization-structure

Blaas, M. (2020, May 15). Personal conversation with Meinte Blaas. (T. Luyten, Interviewer)

Buiteveld, H. (2020, May 14). Interview Hendrik Buiteveld. (T. Luyten, Interviewer)

Cambooth.net. (2020, April 29). *New Moscow Metro Diagram*. Retrieved from cambooth.net/new-moscow-metro-map/

Cloudset Solutions. (2020, April 29). *Filterable Non-hierarchical Knowledge bases*. Retrieved from https://www.cloudset.net/hc/en-us/articles/420212-Faceted-Knowledge-Base

Deltares. (2017). *Het Verhaal van de Rivier, een eerste versie.* Rijkswaterstaat.

Gandhi, R. (2020, June 15). *Naive Bayes Classifier*. Retrieved from Towards data science: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Kater, E. (2020, May 15). Interview Emiel Kater. (T. Luyten, Interviewer)

Kater, E. (2020, June 11). Personal conversation with Emiel Kater. (T. Luyten, Interviewer)

Kroekenstoel, D. (2020, May 11). Interview David Kroekenstoel. (T. Luyten, Interviewer)

Lambe, P. (2007). *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness.* Oxford: Chandos Publishing.

Manning, C. D., Raghavan, P., & Schütze, H. (2008, April 2). *Introduction to information retrieval*. Retrieved from https://nlp.stanford.edu/IR-book/html/htmledition/choosing-what-kind-of-classifier-to-use-1.html

Pal, A. (2020, June 25). *Spam Detection and filtering with Naive Bayes Algorithm*. Retrieved from Medium: https://medium.com/secure-and-private-ai-math-blogging-competition/spam-detection-and-filtering-with-naive-bayes-algorithm-f6c2ac181174

Pellini, A., & Jones, H. (2011). *Knowledge taxonomies: A literature review.* London: Overseas Development Institute.

PennState Eberly College of Science. (2020, June 18). *The Multinomial Distribution*. Retrieved from Stat 504 | Analysis of Discrete Data: https://online.stat.psu.edu/stat504/node/40/

QNT. (2020, April 28). *A new approach to classification of large diverse data sets*. Retrieved from Quasinewtonian: http://www.quasinewtonian.com/popular.html

Rijkswaterstaat. (2015). *Beheer- en ontwikkelplan voor de rijkswateren 2016-2021.* Ministerie van Infrastructuur en Milieu.

Rijkswaterstaat. (2019). *Het verhaal van de Rijn-Maasmonding.* Rijkswaterstaat.

Rijkswaterstaat. (2020, June 4). *Bovenrivierengebied - Beschrijving van het rivierengebied t.b.v. ontwerpbelastingen*. Retrieved from Helpdesk Water: https://www.helpdeskwater.nl/onderwerpen/waterveiligheid/primaire/technische-leidraden/zoeken-technische/@192704/bovenrivierengebied/

Rijkswaterstaat. (2020). *Concept - Plan van aanpak Platform Rivierkennis.* Rijkswaterstaat, Verkeer en Leefomgeving, afdeling Hoogwaterveiligheid.

Rijkswaterstaat. (2020, March 19). *Helpdesk Water*. Retrieved from https://www.helpdeskwater.nl/onderwerpen/waterveiligheid/programma-projecten/rivierkennis/verhaal-rijn/

Rijkswaterstaat. (2020, June 11). *Helpdesk Water*. Retrieved from https://www.helpdeskwater.nl/

Rijkswaterstaat. (2020). *Kennisvragenoverzicht sheets.* Platform Rivierkennis.

Schielen, R. (2020, May 25). Interview Ralph Schielen. (T. Luyten, Interviewer)

Schmidt. (2020, 4 28). *Tree data structures*. Retrieved from people.cs.ksu.edu/~schmidt/300s05/Lectures/Week7b.html

Schoor, M. (2020, May 18). Interview Margriet Schoor. (T. Luyten, Interviewer)

ScienceDirect. (2020, June 24). *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*. Retrieved from ScienceDirect: https://www.sciencedirect.com/book/9781843342274/organising-knowledge

Scikit learn. (2020, June 15). *Naive Bayes*. Retrieved from scikit learn: https://scikit-learn.org/stable/modules/naive_bayes.html

Sieben, A. (2020, May 27). Interview Arjan Sieben. (T. Luyten, Interviewer)

Sieben, A. (2020, June 10). Personal conversation with Arjan Sieben. (T. Luyten, Interviewer)

Simplilearn. (2020, June 25). *Understanding Naive Bayes Classifier*. Retrieved from Simplilearn: https://www.simplilearn.com/tutorials/machine-learning-tutorial/naive-bayes-classifier#:~:text=Advantages%20of%20Naive%20Bayes%20Classifier&text=It%20doesn't%20require%20as,to%20make%20real%2Dtime%20predictions

Synced. (2020, June 18). *Applying Multinomial Naive Bayes NLP Problems: A Practical Explanation*. Retrieved from medium.com: https://medium.com/syncedreview/applying-multinomial-naive-bayes-to-nlp-problems-a-practical-explanation-4f5271768ebf

The Professionals Point. (2020, June 25). *Advantages and Disadvantages of SVM in Machine Learning*. Retrieved from The Professionals Point: http://theprofessionalspoint.blogspot.com/2019/03/advantages-and-disadvantages-of-svm.html

Van Baars, S., & Van Kempen, I. (2009). *The Causes and Mechanisms of Historical Dike Failures in the Netherlands.* Delft: European Water Association (EWA).

Van Putten, D. (2020, May 20). Interview Daniël van Putten. (T. Luyten, Interviewer)

Van Zetten, R. (2020, May 25). Interview Rien van Zetten. (T. Luyten, Interviewer)

Webb, G. I., Boughton, J. R., & Wang, Z. (2005). *Not So Naive Bayes: Aggregating One-Dependence Estimators.* Springer Science + Business Media, Inc.

Wikimedia Commons. (2020, April 28). *File:Simple Periodic Table Chart-en.svg*. Retrieved from https://upload.wikimedia.org/wikipedia/commons/2/2e/Simple_Periodic_Table_Chart-en.svg

Yang, Y., & Liu, X. (2020, April 2). *A re-examination of text categorization methods.* Retrieved from https://dl.acm.org/doi/pdf/10.1145/312624.312647

Appendix A – Literature review

This section covers the full results of the literature review. The source of the information is *''Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness'' (2007)* by Patrick Lambe, except where noted.

Lambe covers different taxonomy forms that can be used to organise knowledge. This literature review describes and discusses each taxonomy form in order to create an understanding of the strengths and weaknesses of each. The results are covered below.

Lists

This is the most basic taxonomy form and the foundation for most of the more complex taxonomy forms. A list is in essence just a collection of related things. This taxonomy form is only useful in very simple situations because it is not possible to convey sub-categories or properties for each item in a list. In fact, more complex taxonomy forms like tree structures and system maps often grow from lists that become too complicated to handle. A rule of thumb is to use lists for a maximum of 12 to 15 items, as it will be difficult to quickly comprehend and navigate longer lists. As can be seen in the list below, it is possible to introduce a form of hierarchy or order into a list, but only by using a specific order. It is not possible to describe any structure using a list.

- Field Marshal
- General
- Lieutenant-General
- Brigadier
- Colonel
- Lieutenant-Colonel
- Major
- Captain
- Lieutenant
- Second Lieutenant

Tree structures

Tree structures are a very common and useful taxonomy form as they provide a clear overview of categories and terms, and their relationships, from broad to narrow. This allows for much information to be added compared to a list. Moving upwards in a tree structure, we generalise based on similarities. Moving downwards in a tree structure, we specify by discriminating based on differences.

Trees are often not 'pure', i.e. perfectly consistent in their internal relationships. This means that in one part of a tree, a subdivision can mean 'A is a part of B', while in another part of the same tree it can mean 'C is a stage in process D'. An example of this is given in **Figure A.1**, where 'Chief Account Officer is working under Commissioner', while 'Korba project is executed by Deputy Commissioner 2'.

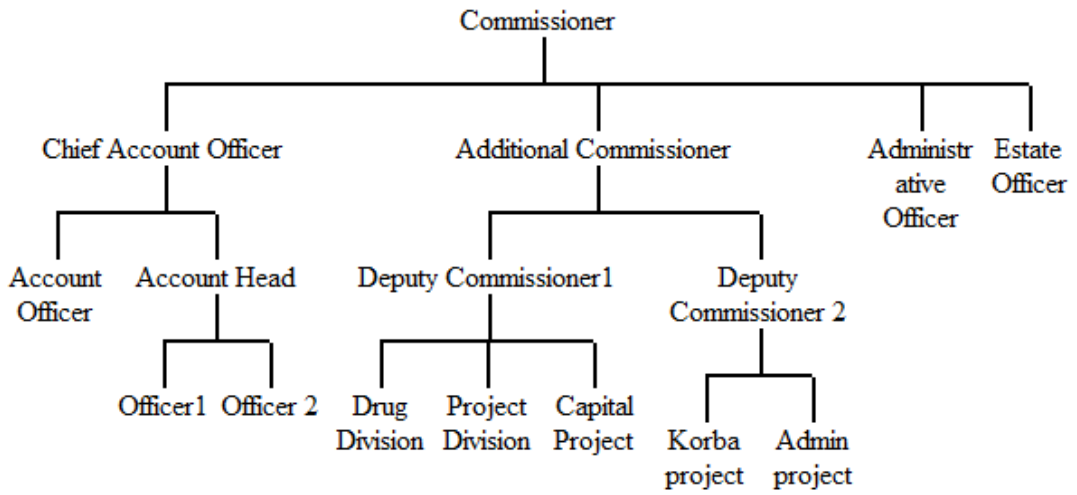


Figure A.1 – *Example of a typical tree structure* (Bhardwaj, 2020)

The lack of consistency is not necessarily a problem because a pragmatic approach can be more useful than striving for strict consistency. It does mean, however, that not all items on the same level in a tree reflect the same type of information ('Officer 1' vs. 'Admin project' in **Figure A.1**). Especially in larger structures, this can cause a loss of clarity.

Hierarchies

Hierarchies are a specific type of tree structure. Compared to a standard tree structure, a hierarchy is much more rigid in its structural consistency. The requirements for a 'scientific' hierarchy are:

- Inclusiveness: the top category includes all subordinate categories in the tree.
- Relational consistency: the kind of relationship between each level in the hierarchy is exactly the same
- Inheritance: subordinate categories in a hierarchy inherit all of the attributes of superordinate categories
- Mutual exclusivity: An entity can belong to only one class

An example of a scientific hierarchy is given in **Figure A.2**.

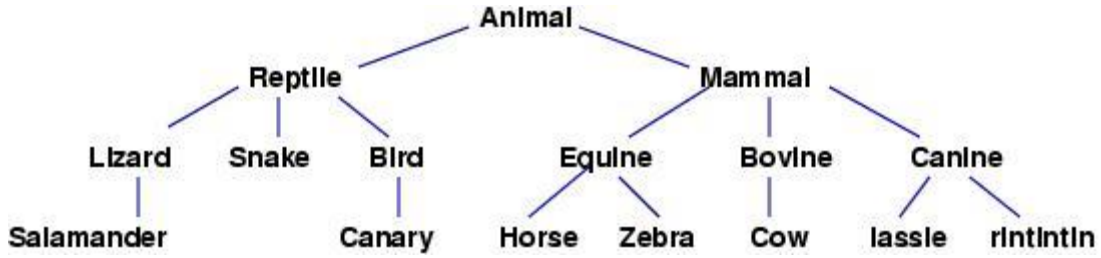


Figure A.2 – *Example of a scientific hierarchy* (Schmidt, 2020)

A scientific hierarchy can in principle be very attractive. Because it is so consistent and unambiguous, a hierarchy should make content easily findable. In practice, however, this is often not the case, since the real world is not so clearly defined and rigorously structured. Many of the definitions and categories used overlap with each other, or they can be ambiguous. Hierarchies are ideal for biology and hard sciences, but they are usually less useful for knowledge management purposes.

Polyhierarchies

Because hierarchies are often too rigid to work well for the purpose of knowledge management, polyhierarchies have been posed as an alternative. In a polyhierarchy, items that belong in more than one class are mapped into more than one superordinate class. This effectively breaks the inheritance and mutual exclusivity rules of the scientific hierarchy. Polyhierarchies can be problematic in knowledge management because cross-connecting hierarchies will erase its consistent structure, which is the main strength of a hierarchy. When too many cross-references start to occur, matrices and facet taxonomies work better (Pellini & Jones, 2011). An example of a polyhierarchy is given in **Figure A.3**.

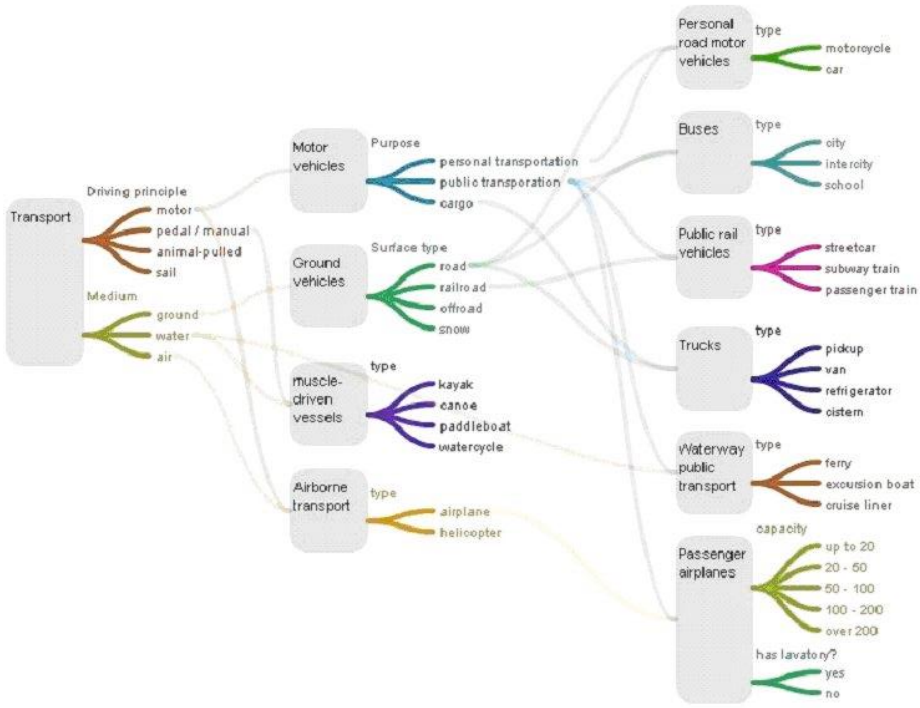


Figure A.3 – *Example of a polyhierarchy* (QNT, 2020)

Matrices

Matrices are useful when content can be categorised along the same dimensions. Due to the overview provided by the matrix structure, items can be compared, problems and opportunities can be identified, and gaps can be located. By comparison, tree structures can only subdivide groups based on one dimension at a time, e.g. 'is a part of'.

Matrices are most effective for categorising items along two dimensions, but they can also be useful for dealing with more than two dimensions. A good example is the periodic table of elements, displayed in **Figure A.4**. The horizontal axis is arranged into groups, while the vertical axis is arranged into periods. However, these are not the only ways that elements can be grouped. Therefore, multiple blocks (based on orbital valence), indicated using lay-out, and categories (metals, metalloids, non-metals, unknown), indicated using different colours, have been introduced. This enables us to categorise elements in a matrix based on four different dimensions.

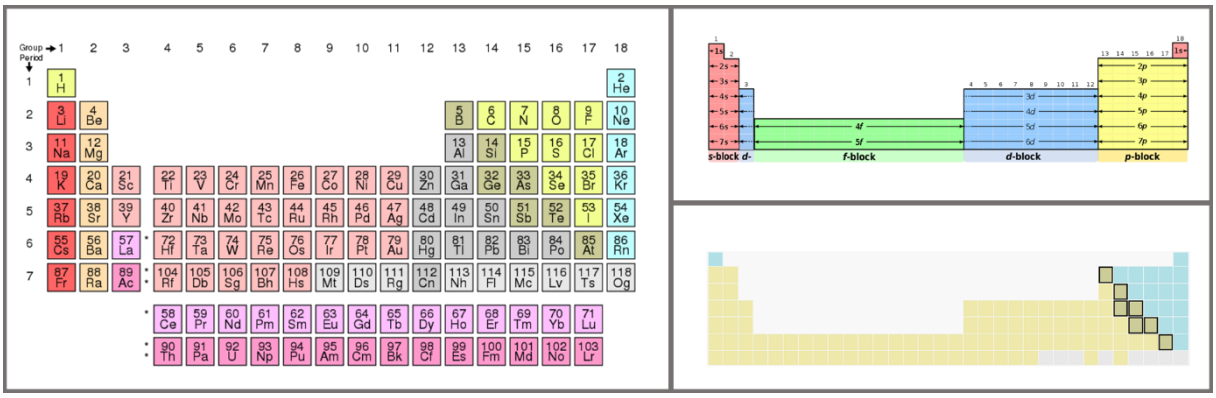


Figure A.4 *– Periodic table of elements as example of a matrix taxonomy. Top right: blocks, bottom right: categories (metals, metalloids, non-metals, unknown)* (Wikimedia Commons, 2020)

While it is possible to use up to four dimensions in a matrix, they are most effective for handling two or three dimensions. Adding more dimensions causes the matrix to lose clarity. Diverse collections of knowledge are not easily expressed using a matrix because not all entities can be described using the same few dimensions.

Facets

Facets were first introduced in 1932 as a way to improve the taxonomy methods used by librarians at the time. It was presupposed that books belong in one specific location or category, while they often fit into multiple different categories. Facets allow items to be classified into multiple categories simultaneously, each facet functioning as a mini taxonomy, often in the form of a list, so they can be identified from different searching points.

Initially, this method was not very feasible because books are physically bound to one specific location. However, with the rise of the digital age, this method became much more effective. After all, digital knowledge is not bound to a physical location, so documents can end up in different locations without a problem. A digital structure also makes it easy to reach documents through different routes (Pellini & Jones, 2011).

Facets work well when dealing with large amounts of content, especially digital files which use tags and metadata (e.g. author(s), title, keywords, date of publication, place of publication). This is why the use of facets is very common in online databases, where users can narrow their search by adding filters. Only results that fit into every category or contain the right metadata will be presented. This principle is graphically displayed in **Figure A.5**.

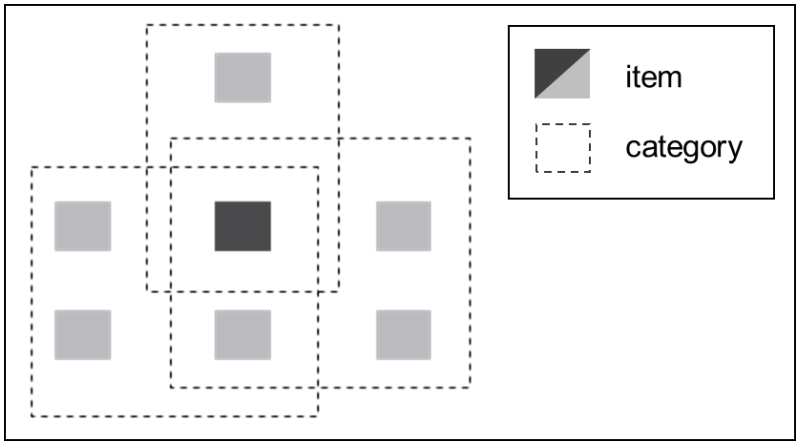


Figure A.5 *– Graphical depiction of a faceted knowledge taxonomy* (Cloudset Solutions, 2020)

One downside of a faceted taxonomy is that its use requires a certain level of subject knowledge. Facets are less effective when the user does not easily understand the structure of the taxonomy, or if the user is not familiar with the categories used. This is not a problem for online stores, but it could be a problem when dealing with specialist knowledge presented to general users. This means that it is important to understand the knowledge base of the target audience.

System maps

A system map is a visual representation of a knowledge domain, where connections between entities are used to express their relationships implicitly or explicitly. It can be descriptive, representing a real-world domain such as a metro line map, illustrated in **Figure A.6**. System maps can also be conceptual, representing non-physical constructs, such as business processes. An example of a conceptual system map, or concept map, is given in **Figure A.7**. System maps have a very strong representational power due to the visual format, but they are not ideal when situations become too complex. The more complex a representation becomes, the more difficult it is to convey information visually. It is also difficult to represent hierarchy using system maps.

Figure A.6 *– Descriptive system map representing the Moscow Metro* (Cambooth.net, 2020)

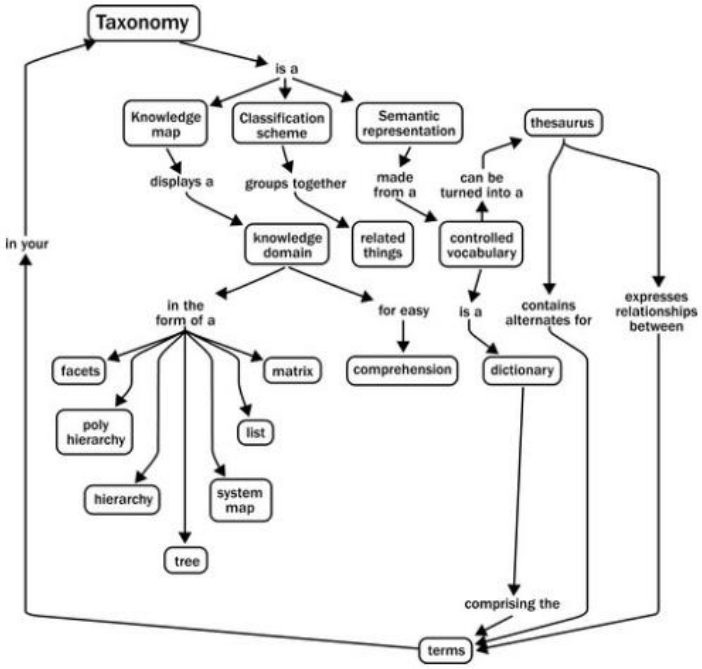


Figure A.7 *– Concept map describing taxonomies* (Lambe, 2007)

Appendix B – Interviews

In total, eight interviews have been conducted. Each interview is covered separately in the subsections of this appendix.

B.1 Interview David Kroekenstoel

This interview was conducted om the 11th of May 2020, via Skype. The interview was in Dutch. An English summary with the most important information is provided below.

English summary
David Kroekenstoel has been working at RWS for almost 20 years. His work mainly concerns production of knowledge, but he also uses existing river knowledge. He is used to sharing river knowledge on Kennisplein and usually does so, except when dealing with classified information, which is often the case in international cooperations. He is usually able to find the required knowledge on Kennisplein, but older reports can be difficult to find because they have been badly labelled or have not been shared at all. The proposed structure seems applicable but is dependent on the discipline of the users. It will only function well if users make an effort to apply the correct labels. It can be difficult to serve all different types of users simultaneously. It should be kept in mind that the website will only be used if it has an added value to the users.

1.	***Kunt u kort iets vertellen over uw functie, achtergrond en werkzaamheden bij RWS?***
Ik werk op dit moment bijna 20 jaar bij RWS. Ik heb civiele techniek gestudeerd aan de Universiteit Twente en ben in 1999 afgestudeerd. Bij RWS ben ik begonnen bij de voorloper van WVL, op de afdeling rivieren, als riviermorfoloog. De afgelopen jaren is het accent van mijn werk verschoven van hydraulica naar hoogwaterveiligheid. Op dit moment ben ik vooral bezig met projecten die vallen onder BOA (DGWB). Verder ben ik actief in internationale samenwerkingen. Zo ben ik onderdeel van de internationale werkgroep Rijn en Maas en ben ik bezig met een studie naar overstromingsrisico's op de grens tussen Duitsland en Nederland. Een derde van mijn tijd besteed ik aan internationale projecten, een derde aan projecten onder BOA en een derde aan andere projecten.	
2.	***Op wat voor manier bent u bezig met het produceren en/of het gebruiken/verwerken van rivierkennis?***
Ik ben voornamelijk werkzaam in de ontwikkeling van rivierkennis. Binnen het BOA werk ik aan ontwikkelingsprojecten over de systeemwerking van hydraulica. Verder doe ik onderzoek naar negatieve systeemwerking van rivieren. Daarnaast pas ik ook bestaande kennis toe, bijvoorbeeld in het gebruik van bestaande hydraulische modellen.	
3.	***Deelt u altijd nieuwe rivierkennis via Kennisplein of andere ontsluitingsmethodes? Waarom wel of niet?***
De afgelopen jaren is het gebruikelijk geworden om kennis standaard te delen op Kennisplein, maar dit verschilt sterk per project. Bij sommige projecten is het vanzelfsprekend om de kennis te delen. Bij andere projecten niet, vaak vanwege gevoeligheid van informatie. Bij internationale projecten is het afhankelijk wat de internationale partners ervan vinden. Kennis kan ook gedeeld worden via de Helpdesk Water, waar een aparte projectpagina kan worden aangemaakt.	
4.	***Kunt u de documenten die u nodig hebt altijd vinden via Kennisplein e.d.?***
Meestal kan ik de kennis die ik nodig heb vinden via Kennisplein. De afgelopen jaren zijn rapporten beter actief gedeeld op Kennisplein. Oudere rapporten uit het papieren archief zijn vaak slecht vindbaar omdat ze niet goed zijn gelabeld of simpelweg niet gedeeld. Ook zijn ze vaak slecht ingescand.	
5.	***Bent u tevreden met de huidige manier van kennisontsluiting of vindt u dat het beter kan? Zo ja, hoe?***

Het kan natuurlijk beter, maar realistisch gezien is het erg lastig om iets te maken en bij te houden wat aan ieders wensen voldoet. Voor mijn eigen doeleinden voldoet de huidige manier van kennisontsluiting.	
6.	***Hoe zou u zelf documenten met rivierkennis indelen (welke categorieën/wat voor structuur)? Waarom op deze manier?***
In ieder geval de vijf verschillende thema's en een indeling tussen Rijn en Maas. Verder kan er projectsgewijs worden ingedeeld en een map met wetenschappelijk onderzoek worden gebruikt. Documenten in mijn persoonlijke opslag gebruik ik uitgangspunten gekoppeld aan modellen: modelschematisaties, gegevens over bodemligging, ruwheid, afvoer, golfvorm, etc. Maar dit is voor algemene kennisontsluiting niet een logische indeling. Het rapport over hydraulische belastingen verschijnt iedere 6 jaar voor elke riviertak. Die zou ik graag gebundeld beschikbaar zien.	

[Uitleg van resultaten literatuur]

7.	***Denkt u dat deze structuur bruikbaar is? Waarom wel of niet?***
De structuur lijkt bruikbaar, maar het staat of valt met welke labels gebruikt worden. Zorg dat de juiste labels uniform gebruikt worden. Ook is er van de gebruikers een zekere discipline nodig om bij het delen van bestanden altijd de juiste labels aan documenten toe te kennen. Je zou een label vrij kunnen laten voor iemand om zelf te bepalen, zodat het niet te strikt wordt. Vergeet niet dat als iets wordt gebouwd, het ook moet worden onderhouden en voldoende gebruikt moet worden.	
8.	***Denkt u dat het haalbaar is om op deze of een andere manier rekening te houden met de verschillende types gebruikers (RWS-werknemers, wetenschappelijke gemeenschap, marktpartijen)?***
Dit is lastig, omdat verschillende doelgroepen verschillende eisen stellen aan de website. Een gebruiker die al prima op andere manieren bediend wordt, zal er niet veel moeite in steken. De gebruiker die je bedient, moet het meeste voordeel hebben bij zo'n website. Houd goed in je achterhoofd welke partijen er het meeste baat bij hebben.	

B.2 Interview Hendrik Buiteveld

This interview was conducted om the 14th of May 2020, via Skype. The interview was in Dutch. An English summary with the most important information is provided below.

English summary
Hendrik Buiteveld studied applied physics at the UT and eventually started working for RWS. In recent years, he has been dealing with that enters the Netherlands from abroad. He mainly works on knowledge production, sometimes in cooperation with KNMI or Deltares. He usually shares reports on Kennisplein and sometimes on Helpdesk Water, but not memos that may contain important information as well. documents on Kennisplein are findable, but this is not always easy. It usually only works if you search very precisely and have extensive knowledge on the subject. The proposed structure sounds flexible in searching, which is good, but it requires discipline from the uploader to label every document correctly. People are usually more inclined to start something new than to correctly wrap up previous work, so archiving does not always happen. People sometimes use different names for the same terms, so standardising will be necessary. Implementation into an existing structure is preferable over creating a new website, to make the process easier for users. A universally applicable website for different audiences would be best. A label with the type of publication (research, policy article, background report) could help with.

1.	***Kunt u kort iets vertellen over uw functie, achtergrond en werkzaamheden bij RWS?***
Ik heb technische natuurkunde gestudeerd aan de UT, afgestudeerd op supergeleiding. Daarna ben ik bezig geweest met biomedische optische meettechniek. Van daaruit gewerkt aan het meten van stoffen d.m.v. remote sensing. Daarna onderzoek gedaan voor RWS in Delft. Toen bij RWS begonnen aan de afdeling hydrologie en veiligheid, de voorloper van hoogwaterveiligheid. De afgelopen jaren ben ik voornamelijk bezig geweest met water dat Nederland binnenkomt en o.a. aan afvoerverwachtingen en de effecten van maatregelen.	
2.	***Op wat voor manier bent u bezig met het produceren en/of het gebruiken/verwerken van rivierkennis?***
Veel wat wij doen is kennisproductie, ook samen met het KNMI of Deltares. In het verleden hebben we veel samengewerkt met een instituut in Duitsland. Wij kijken hoe maatregelen in het stroomgebied uitpakken voor klimaatscenario's. Dit is in dienst van DGWB, dus uiteindelijk heeft dit beleidsimplicaties.	
3.	***Deelt u altijd nieuwe rivierkennis via Kennisplein of andere ontsluitingsmethodes? Waarom wel of niet?***
Geproduceerde rapporten worden gedeeld op Kennisplein. Er worden ook veel memo's geproduceerd die niet op Kennisplein terechtkomen. Soms staat in deze memo's wel essentiële informatie die vervolgens niet wordt gerapporteerd. In de regel deel ik nieuwe kennis wel op Kennisplein en sommige dingen op Helpdesk Water. Kennis die ontwikkeld is voor Deltares en KNMI wordt apart op hun eigen websites gedeeld.	
4.	***Kunt u de documenten die u nodig hebt altijd vinden via Kennisplein e.d.?***
Over het algemeen zijn documenten wel te vinden, maar Kennisplein is niet heel gebruiksvriendelijk. Uit een zoekopdracht komt vaak een grote waslijst, maar niet altijd waar je naar zoekt. Als je heel precies weet wat je zoekt, dan lukt het wel. Algemeen zoeken werkt niet. Als je nog niet bekend bent met een onderwerp, zul je hulp nodig hebben met zoeken. Met een simpele zoekopdracht zal het dan niet lukken.	
5.	***Bent u tevreden met de huidige manier van kennisontsluiting of vindt u dat het beter kan? Zo ja, hoe?***
Het zou wel beter moeten kunnen. Alle kennis zou op Kennisplein terecht moeten komen. Als overheid is het belangrijk dat dit netjes gebeurt, want het wordt met belastinggeld betaald. In theorie zou dit via Kennisplein moeten kunnen, maar ik weet niet zo snel hoe.	
6.	***Hoe zou u zelf documenten met rivierkennis indelen (welke categorieën/wat voor structuur)? Waarom op deze manier?***

Je kunt onderscheid maken tussen hoogwater en laagwater en indelen op chemische of biologische thema's. Ook kun je onderscheid maken tussen beleidsmatige en wetenschappelijke documenten en op geografisch gebied/locatie indelen. De indeling hoeft niet zo gedetailleerd als in bibliotheken, dat kan het proces van uploaden moeilijk maken.

[Uitleg van resultaten literatuur]

7.	***Denkt u dat deze structuur bruikbaar is? Waarom wel of niet?***
Het klinkt flexibel in het zoeken en dat is goed, maar het vergt wel discipline van de uploader om alle documenten goed te labelen. Mensen vinden het altijd leuker om met iets nieuws te beginnen dan het oude netjes af te ronden, daarom gebeurt het archiveren niet altijd goed. Zelfbedachte namen voor categorieën, mensen gebruiken soms hun eigen verschillende termen voor bestaande categorieën, werken niet met deze structuur. Het is handig om te standaardiseren, maar dat zal sowieso nodig zijn voor een nieuwe structuur. Ik zou niet gelijk een nieuwe website hiervoor maken, maar het binnen een bestaande structuur proberen te verwerken. Dit is vooral belangrijk om het proces makkelijk te maken, zodat mensen het in de praktijk gebruiken.	
8.	***Denkt u dat het haalbaar is om op deze of een andere manier rekening te houden met de verschillende types gebruikers (RWS-werknemers, wetenschappelijke gemeenschap, marktpartijen)?***
Een universeel bruikbare website zou makkelijk zijn. Een label met het type publicatie (onderzoek, beleidsartikel, achtergrondrapport) kan hierbij helpen. De hoeveelheid labels is wel een afweging die je moet maken. Als er te veel categorieën/labels worden gebruikt, wordt het veel werk om te delen en zien mensen er het voordeel misschien niet van in.	

B.3 Interview Emiel Kater

This interview was conducted om the 15th of May 2020, via Skype. The interview was in Dutch. An English summary with the most important information is provided below.

English summary
Emiel Kater is a river consultant (rivierkundig adviseur) at ON, working mainly on flood safety. He has also worked on improving information supply and knowledge disclosure within ON. He does not share river knowledge centrally, as it is not clear whether documents should be shared. He is usually able to find relevant knowledge on Kennisplein, but Helpdesk Water is considered more of an authority. The facet structure seems appropriate, as it is useful to represent overlapping categories. The main functions, geographic areas and metadata are straightforward and useful categories. Different information products could be named (soil position, morphology, vegetation, water levels, etc.) and a distinction between scientific publications, practical research and policy documents could be made. One should be cautious in adding too many categories, as they might change over time, making already labelled documents problematic. A solution for this can be to label using keywords, as opposed to a large number of specified categories. People will be more inclined to use this structure if it can be implemented into an existing website like Kennisplein or Helpdesk Water, but pragmatically it will be easier to create something new. It is possible to serve the different types of users in this manner, perhaps by providing a different entrance to the website for each target audience.

1.	***Kunt u kort iets vertellen over uw functie, achtergrond en werkzaamheden bij RWS?***
Ik ben rivierkundig adviseur bij ON. Hier werken we aan het beheren van de Rijntakken, zelf ben ik vooral bezig met hoogwaterveiligheid. Vanuit ON ben ik bezig geweest de informatievoorziening en kennisontsluiting voor onszelf op een rijtje te zetten. In Nijmegen heb ik biologie/milieukunde gestudeerd en daarna ben ik vrij snel bij RWS terechtgekomen. Ik ben een tijdje weggeweest, maar sinds 5 jaar weer terug. Ik heb altijd iets met rivieren en waterveiligheid gedaan, ook buiten RWS. Werk ook met GIS.	
2.	***Op wat voor manier bent u bezig met het produceren en/of het gebruiken/verwerken van rivierkennis?***
De afgelopen jaren ben ik het meest concreet bezig geweest met informatie en kennis, bijvoorbeeld het werken aan de rivierkundige informatiemap. De ambitie daarvan was om informatieproducten en documenten die rivierkundigen nodig hebben beschikbaar te maken op een centrale plek. Dit heb ik geprobeerd te maken naast mijn gewone taken, om in mijn eigen team van 10 mensen makkelijk te kunnen ontsluiten.	
3.	***Deelt u altijd nieuwe rivierkennis via Kennisplein of andere ontsluitingsmethodes? Waarom wel of niet?***
Dat heb ik nooit bewust gedaan. Het is mij niet duidelijk wanneer iets daar wel of niet terecht moet komen. Ik heb niet aan kennisontwikkeling gewerkt, maar wel praktische onderzoeken gedaan die gedeeld zouden kunnen worden. Dat zou ik graag wel doen, maar het komt er niet van. Misschien heb ik me er zelf te weinig in verdiept en mogelijk is het een cultuurprobleem. Als de omgeving eraan gewend raakt en het gaat wat meer leven, misschien dat het dan beter gaat. Ook moet het duidelijk worden of iets wel geschikt is om gedeeld te worden.	
4.	***Kunt u de documenten die u nodig hebt altijd vinden via Kennisplein e.d.?***
Meestal wel, de Helpdesk Water is helder. Zoeken doe ik op thema en dat lukt bij Kennisplein minder. De Helpdesk Water is meer een autoriteit. Op de interne schijf waar we zelf dingen bewaren, kan ik dingen beter vinden.	
5.	***Bent u tevreden met de huidige manier van kennisontsluiting of vindt u dat het beter kan? Zo ja, hoe?***
Helpdesk Water werkt goed, vooral voor het opzoeken van sleuteldocumenten die bij een bepaald beleid of project horen. Als het gaat om onderzoeksrapporten kan het beter. Niet alleen via Kennisplein, maar ook via Intranet is het niet altijd duidelijk.	

6.	***Hoe zou u zelf documenten met rivierkennis indelen (welke categorieën/wat voor structuur)? Waarom op deze manier?***
Dat is voor iedereen verschillend, ON en WVL zouden heel verschillende manieren toepassen. De indeling in de vier hoofdfuncties is zinnig, maar waterveiligheid is een breed begrip. Afvoerverdeling is iets waarop ik ook zou willen zoeken. De vraag is of je eerst op gebied of op thema wilt zoeken, diezelfde discussie hebben we bij de riviermap. Uiteindelijk hebben we voor gebied gekozen en een soort boomstructuur gemaakt.	

[Uitleg van resultaten literatuur]

7.	***Denkt u dat deze structuur bruikbaar is? Waarom wel of niet?***
Zoiets als dit lijkt mij een goede structuur. Het is nuttig dat je documenten niet op meerdere plekken kunt/hoeft op te slaan omdat het mogelijk is om overlap mee te nemen.	
8.	***Denkt u dat de genoemde categorieën bruikbaar zijn? Waarom wel of niet?***
De vier hoofdfuncties zijn belangrijk, dat is de kern van RWS. Het ligt voor de hand om die te gebruiken, net als de gebieden. De tags en metadata zijn ook recht toe recht aan. De types kennis (systeemkennis, modellen en instrumenten, monitoring) zijn nog een beetje vaag. Systeemkennis is erg breed. Je zou informatieproducten kunnen noemen: bodemligging, morfologie, vegetatie, waterstanden, etc. Verder kun je onderscheid maken tussen wetenschappelijke publicaties, praktijkonderzoek, beleid. Als categorieën veranderen of worden toegevoegd, wordt het een probleem. Je moet dan alle gelabelde documenten opnieuw langsgaan. Misschien moet je het nu niet te specifiek maken, maar meer met kernwoorden werken. Als het nu te specifiek wordt, dan is het risico te groot dat er iets aan verandert. I.p.v. meer facetten zou je meer kernwoorden kunnen gebruiken (bodemligging, morfologie, vegetatie, waterstanden).	
9.	***Denkt u dat de nieuwe structuur kan worden toegepast op een bestaande website zoals Kennisplein of Helpdesk Water***
Het liefst integreren in iets wat al bestaat, dan zullen mensen eerder geneigd zijn om het te gebruiken. Als er iets nieuws is, gaan we het dan niet meer op Kennisplein zetten? Als er verschillende systemen naast elkaar worden gebruikt, dan moeten er goede afspraken over worden gemaakt. Ik zit niet te wachten op verschillende systemen voor iedereen. Vanuit het gebruik zou ik één geheel dus mooi vinden, maar pragmatisch gezien zou ik iets nieuws maken.	
10.	***Denkt u dat het haalbaar is om op deze of een andere manier rekening te houden met de verschillende types gebruikers (RWS-werknemers, wetenschappelijke gemeenschap, marktpartijen)?***
Ik denk dat dat wel kan, als je het voor elkaar krijgt een flexibele website te maken, met per doelgroep een verschillende ingang naar de website. Groep A zoekt zo en groep B zoekt zo. Dat is technisch gezien geen belemmering, maar vanuit het beheer is het niet mogelijk om iets te specifiek of te organisch te maken. Verder moet er bewust gekozen worden welke documenten wel en niet gedeeld worden.	

B.4 Interview Margriet Schoor

This interview was conducted om the 18th of May 2020, via Skype. The interview was in Dutch. Due to time constraints, some questions have been skipped. An English summary with the most important information is provided below.

English summary
Margriet Schoor works at the department of Water quality and Ecology, mainly on ecology. She is involved in monitoring to generate advice and plans ecologically tinted studies. She always shares new knowledge on Kennisplein and then shares a link with people whom she thinks are interested. She finds it difficult to find the right documents on Kennisplein. The search function yields many irrelevant results, even when searching for the right title or author. This is because old documents have not been shared or have been labelled incorrectly. The proposed structure appears applicable because overlapping categories can be represented. For an online database/website, this seems like a logical structure. Some products cannot be classified using these categories. Studies on sustainability, for example, cannot be coupled to one of the main functions or a specific area. The main functions seem fixed but until a few years ago, water quantity/availability was not used as a separate category. This means that even the main functions are subject to change.

1.	***Kunt u kort iets vertellen over uw functie, achtergrond en werkzaamheden bij RWS?***
Ik werk bij de afdeling Waterkwaliteit en Ecologie, met name aan ecologie. Ik houd me bezig met monitoring om adviezen te geven. Ik werk ongeveer 30 jaar bij RWS, waarvan de laatste 13 jaar bij waterkwaliteit en ecologie, daarvoor morfologie. Op dit moment laat ik veel studies uitvoeren die ecologisch getint zijn.	
2.	***Op wat voor manier bent u bezig met het produceren en/of het gebruiken/verwerken van rivierkennis?***
Ik laat specifiek onderzoek uitvoeren, dit is dus het produceren van kennis. Toen het nog kon, ging ik naar bijeenkomsten en symposia om kennis op te doen.	
3.	***Deelt u altijd nieuwe rivierkennis via Kennisplein of andere ontsluitingsmethodes? Waarom wel of niet?***
Ja, eigenlijk altijd. Wat uit onderzoeken komt, zet ik op Kennisplein en dat komt dan openbaar beschikbaar. De link hiervan deel ik met mensen van wie ik denk dat ze er belangstelling bij kunnen hebben.	
4.	***Kunt u de documenten die u nodig hebt altijd vinden via Kennisplein e.d.?***
Je moet heel veel ervan weten, wil je documenten goed kunnen vinden. Er komen veel irrelevante documenten uit de zoekfunctie, zelfs als je zoekt op de juiste auteur. Dit kan ook komen doordat oude documenten niet goed zijn gedeeld of slecht zijn gelabeld op Kennisplein. Kennisplein zelf vind ik tegenwoordig lastig te vinden op intranet. Vroeger was Kennisplein direct vindbaar, nu moet je veel doorklikken. Als je wilt dat mensen het meer gebruiken, dan moet het beter toegankelijk zijn.	
5.	***Bent u tevreden met de huidige manier van kennisontsluiting of vindt u dat het beter kan? Zo ja, hoe?***
De vindbaarheid van documenten moet beter kunnen. Als je op de titel van een document zoekt, komt dit er bijvoorbeeld vaak niet als eerste resultaat uit. Ook zijn veel oude fysieke documenten slecht ingescand, wat enorme bestanden met een slechte kwaliteit oplevert. Voor het bestaan van de online database, moesten bestanden al aan de bibliotheek gestuurd worden. Dan zou je die moeten kunnen vinden i.p.v. een scan. Zelfs mijn eigen oude werk is niet meer vindbaar op Kennisplein, 20 jaar geleden kon ik dat uit de bibliotheek lenen. Documenten die nu worden gedeeld moeten over 20 jaar nog wel vindbaar zijn.	
6.	***Hoe zou u zelf documenten met rivierkennis indelen (welke categorieën/wat voor structuur)? Waarom op deze manier?***
Hier heb ik geen duidelijk idee over. Mijn hoofd is niet op die manier gestructureerd. Dat werkt meer op een andere manier van zoeken.	

[Uitleg van resultaten literatuur]

7.	***Denkt u dat deze structuur bruikbaar is? Waarom wel of niet?***
De facettenstructuur lijkt een nuttige toepassing, het is goed dat overlap tussen categorieën kan worden meegenomen op deze manier. Lijkt een logische structuur voor een online database/website.	
8.	***Denkt u dat de genoemde categorieën bruikbaar zijn? Waarom wel of niet?***
Sommige documenten zijn niet in te delen in deze categorieën. Denk bijvoorbeeld aan modelinstrumentarium, dit kan niet direct gekoppeld worden aan de hoofdfuncties of een specifiek gebied. Ook studies over duurzaamheid of gebiedsinrichting zijn moeilijk in te delen. De gebieden moet je eerst indelen in stroomgebieden, dan riviertakken. Je zou een onderscheid kunnen maken tussen inhoud en proces. Proces: hoe is iets georganiseerd. Inhoud: bepaalde kennis in een vakgebied. De vier hoofdfuncties lijken vast te staan, maar tot een paar jaar terug werd zoetwaterbeschikbaarheid/ waterkwantiteit niet als aparte categorie/onderwerp gebruikt. Zelfs dit kan dus veranderen. De vraag is: als de wereld verandert of de kijk van mensen op de wereld verandert, kun je daarmee omgaan?	

B.5 Interview Daniël van Putten

This interview was conducted om the 20th of May 2020, via Skype. The interview was in Dutch. An English summary with the most important information is provided below.

English summary
Daniël van Putten works as a river consultant at ON. A large part of his work is judging projects within flood plains of rivers. An ancillary task is to produce, collect and distribute information about the Rhine branch system. He does not share river knowledge on Kennisplein because it is not clear what should and should not be shared. When he has a question about a certain subject, he will search for a theme but not a specific document. Many documents are already available at the department, so it is not necessary to search centrally. It is important that the connection between departments is improved, so people know whom to ask when looking for something. The proposed structure makes sense and it is good that overlapping categories can be taken into account, enabling to search for themes and not just specific documents. Distinctions can be made by geographic area, hydraulics/morphology, and scale level. A category like 'models' is too broad and vague. It is preferable to have all knowledge in a central location, rather than introducing a new website. If a new website is introduced, make sure that it is linked to Kennisplein. For serving members of the scientific community, research data could be added in addition to the main report file. Other than that, this structure could serve each target audience well.

1.	***Kunt u kort iets vertellen over uw functie, achtergrond en werkzaamheden bij RWS?***
Ik ben net als Emiel rivierkundig adviseur bij ON. Een groot deel van mijn werk is het beoordelen van projecten in uiterwaarden van de rivier. Hierbij zorg ik dat dingen netjes worden aangelegd, er geen overlast ontstaat voor hoogwaterveiligheid, ecologie, etc. Een neventaak is om rivierkundige informatie van het Rijntakkensysteem te ontwikkelen, verzamelen en verspreiden. Ik heb civiele techniek gestudeerd aan de UT, met een master Water Engineering & Management.	
2.	***Op wat voor manier bent u bezig met het produceren en/of het gebruiken/verwerken van rivierkennis?***
Waar ik mee bezig ben zie ik meer als informatie dan als kennis, maar het verschil is niet altijd duidelijk. Een voorbeeld is de Betrekkingslijnen Rijntakken 2018, wat ik net heb gestuurd.	
3.	***Deelt u altijd nieuwe rivierkennis via Kennisplein of andere ontsluitingsmethodes? Waarom wel of niet?***
Nee, dat heb ik nooit gedaan. Het is voor mij niet duidelijk wat wel of niet gedeeld moet worden. Veel documenten en memo's lijken niet 'gelikt' genoeg om als eindproduct gedeeld te worden.	
4.	***Kunt u de documenten die u nodig hebt altijd vinden via Kennisplein e.d.?***
Ik zoek nooit echt gericht op Kennisplein. Soms heb ik wel een vraag over een onderwerp, maar dan zoek ik niet op een specifiek document, meer op een thema. Veel documenten zijn op de afdeling zelf al beschikbaar, dan ga ik niet op Kennisplein zoeken. In mijn ervaring zijn documenten vaak nog wel vindbaar op Kennisplein.	
5.	***Bent u tevreden met de huidige manier van kennisontsluiting of vindt u dat het beter kan? Zo ja, hoe?***
Ik denk dat het beter kan. RWS is een grote organisatie, als enkeling ben je niet altijd op de hoogte van alles wat relevant kan zijn. Dan helpt het als relevante kennis makkelijk vindbaar is. Soms wordt er dubbel onderzoek gedaan omdat men niet op de hoogte is wat er al bekend is. Het is belangrijk dat de verbinding tussen afdelingen beter gaat. Als je weet wie je waarvoor nodig hebt, komt je vraag op de goede plek terecht. Als je elkaar opzoekt, weet je nog niet alles, maar het helpt om bij de juiste persoon aan te kloppen.	
6.	***Hoe zou u zelf documenten met rivierkennis indelen (welke categorieën/wat voor structuur)? Waarom op deze manier?***
Daar heb ik nog nooit zo over nagedacht. Je kunt het geografisch indelen (Rijn/Maas). Hydraulica en morfologie zijn belangrijke thema's. Indelen kan op schaalniveau (specifieke plek vs. groot gebied). Verder	

zijn er nog belangrijke onderwerpen als vegetatie, waterveiligheid, kribben, ijs, morfologie, ecologie. In plaats van het gebruik van mappen, zou je ook alleen labels kunnen gebruiken voor een zoekfunctie.

[Uitleg van resultaten literatuur]

7.	***Denkt u dat deze structuur bruikbaar is? Waarom wel of niet?***
De structuur is logisch, het is goed dat overlap tussen categorieën meegenomen kan worden. Het is mooi dat er thematisch gezocht kan worden om een algemeen beeld te krijgen van een onderwerp en je niet alleen kunt zoeken op een specifiek document.	
8.	***Denkt u dat de genoemde categorieën bruikbaar zijn? Waarom wel of niet?***
Documenten die we nu hebben zijn nog niet zo ingedeeld. Een categorie als 'modellen' is te groot en algemeen. Hydraulica en morfologie kunnen toegevoegd worden, maar die zitten al wel verstopt achter de hoofdfuncties, dus het is de vraag of dat helpt.	
9.	***Ziet u deze structuur liever op een nieuwe website of een bestaand platform? Waarom?***
Het heeft mijn voorkeur om kennisdocumenten op één plek te hebben. Als elk thema zijn eigen website krijgt, word ik daar niet blij van. Als er een nieuwe website komt, zorg dan dat er goed wordt gelinkt van de rivierkenniswebsite naar Kennisplein en vice versa. Het is belangrijk om te weten of je documenten dan alleen op de nieuwe website moet delen of ook via Kennisplein.	
10.	***Denkt u dat het haalbaar is om op deze of een andere manier rekening te houden met de verschillende types gebruikers (RWS-werknemers, wetenschappelijke gemeenschap, marktpartijen)?***
Ik heb zelf de ervaring dat documenten op Kennisplein redelijk vindbaar zijn. Misschien dat je onderzoeksdata kunt toevoegen voor wetenschappelijke gebruikers, dus niet alleen een pdf van het hoofdrapport. In principe is deze structuur prima voor verschillende gebruikers.	

B.6 Interview Rien van Zetten

This interview was conducted om the 25th of May 2020, via Skype. The interview was in Dutch. An English summary with the most important information is provided below.

English summary
Rien van Zetten works as project manager at GPO in Utrecht and has worked in the water domain for around 40 years. Until 6 years ago, he worked at WVL. He has been involved in the different 'Stories of the River' and is part of the specialist pool of POR. He is not used to sharing river knowledge on Kennisplein, and usually asks someone to do that for him. He rarely uses Kennisplein, as his work usually requires knowledge to be produced instead of used. When he did use Kennisplein, he was able to find what he needed. Using the right keywords is important in making knowledge findable. New terms could be coupled to existing keywords in the future to ensure that documents remain findable when new terminology is introduced. A database structure that enables to search on different themes is preferred. Distinctions could be made by upper/lower river region, and the main functions. The use of the right keywords is essential. Changing terminology over time can cause problems. Linking new keywords to old ones can be a solution to this. If a new structure is introduced, someone must be appointed to managing the system on a daily basis. A new structure will require a new website, but people will be more inclined to share knowledge on an existing platform. Serving different user types should not be a problem because the same terms will be topical in different fields.

1.	***Kunt u kort iets vertellen over uw functie, achtergrond en werkzaamheden bij RWS?***
Ik werk als projectmanager bij GPO in Utrecht. Bij GPO worden alle weg- en waterprojecten groter dan €50 miljoen uitgevoerd. Ik heb civiele techniek gestudeerd aan de HTS in Dordrecht (nu Hogeschool Rotterdam), toegepaste natuurwetenschappen en filosofie gestudeerd aan verschillende hogescholen. Ik werk nu zo'n 40 jaar en heb altijd in het waterdomein gewerkt. Binnen die sector heb ik alles gedaan, de hele keten van beleidsvoorbereiding tot aan realisatie. Ik heb zes jaar bij WVL gewerkt, tot ongeveer zes jaar geleden omdat ik graag weer aan grote projecten wilde werken.	
2.	***Op wat voor manier bent u bezig met het produceren en/of het gebruiken/verwerken van rivierkennis?***
Op verschillende manieren. Ik ben betrokken bij het Verhaal van de Rijn-Maasmonding, Verhaal van de Rivier 2.0 en het Verhaal van de Rijntakken. In het POR ben ik aanwezig als onderdeel van de deskundigenpool. De verhalen van de Rivier is het genereren van kennis, de deskundigenpool levert ook advies.	
3.	***Deelt u altijd nieuwe rivierkennis via Kennisplein of andere ontsluitingsmethodes? Waarom wel of niet?***
Nee, meestal vraag ik iemand om dat voor me te doen. Ik ben zelf niet goed met het delen van dingen en word er wel eens op gewezen dat terugkoppeling beter moet. Als er een beter systeem was, zou ik het eerder doen, dat nodigt meer uit om actief te delen.	
4.	***Kunt u de documenten die u nodig hebt altijd vinden via Kennisplein e.d.?***
Ik zoek niet vaak op Kennisplein, waar ik werk moet kennis vaak juist nog gegenereerd worden. Voor mijn werk hoef ik dus niet vaak op Kennisplein te kijken. Wanneer ik Kennisplein heb gebruikt, kon ik wel vinden wat ik nodig had.	
5.	***Bent u tevreden met de huidige manier van kennisontsluiting of vindt u dat het beter kan? Zo ja, hoe?***
Daar is maar één antwoord op, het kan beter. Het is belangrijk om de juiste trefwoorden te gebruiken. Die moeten zo breed zijn dat ze in de toekomst toepasbaar blijven, maar tegelijk specifiek genoeg om documenten direct te kunnen vinden. Je zou nieuwe begrippen in de toekomst kunnen linken aan bestaande begrippen die verwant zijn, zodat documenten vindbaar blijven als er nieuwe terminologie wordt geïntroduceerd.	
6.	***Hoe zou u zelf documenten met rivierkennis indelen (welke categorieën/wat voor structuur)? Waarom op deze manier?***

Een database structuur waarin je op verschillende thema's kunt filteren. Je kunt onderscheid maken tussen het boven- en benedenrivierengebied en de hoofdfuncties. De juiste trefwoorden zijn nodig om goed te kunnen zoeken.

[Uitleg van resultaten literatuur]

7.	***Denkt u dat deze structuur bruikbaar is? Waarom wel of niet?***
De structuur op zichzelf is naar verhouding minder belangrijk. De manier waarop je kunt zoeken is naar verhouding belangrijker. Probeer goed te kopiëren van systemen die werken, iets is beter goed gepikt dan slecht bedacht. Het grootste probleem is dat termen over de tijd kunnen veranderen. Probeer het mogelijk te maken om nieuwe trefwoorden te koppelen aan oude. Dit zou je kunnen doen door een trefwoordenboom toe te voegen aan het begin van de pagina, waarin de termen en hun onderlinge verhoudingen worden weergegeven. Vraag aan een wetenschappelijke bibliotheek hoe zij omgaan met nieuwe termen. Zorg ook dat je breder kijkt dan alleen mensen van WVL, omdat die niet altijd buiten hun eigen manier denken. Als deze structuur er komt, is het noodzakelijk dat er iemand komt die het onderhoudt. Iemand voor dagelijkste handelingen is dus belangrijk. Kijk voor de Rijntakken hoe geografen dit indelen, bijvoorbeeld in een atlas. Buiten het eigen vakgebied kun je slimme suggesties meekrijgen.	
8.	***Ziet u deze structuur liever op een nieuwe website of een bestaand platform? Waarom?***
Als er een nieuwe structuur komt, dan een nieuwe website. Eigenlijk moet gebruik gemaakt worden van bestaande platforms, dat zijn mensen eerder geneigd te gebruiken. Zo mogelijk Kennisplein, hier staat kennis breder dan alleen rivieren, wat ook relevant kan zijn.	
9.	***Denkt u dat het haalbaar is om op deze of een andere manier rekening te houden met de verschillende types gebruikers (RWS-werknemers, wetenschappelijke gemeenschap, marktpartijen)?***
In principe zou er geen verschil in moeten zitten, kennis blijft hetzelfde. Trefwoorden die voor ons nu actueel zijn, zullen voor marktpartijen op een identieke manier actueel zijn.	

B.7 Interview Ralph Schielen

This interview was conducted om the 25th of May 2020, via Skype. The interview was in Dutch. An English summary with the most important information is provided below.

English summary
Ralph Schielen has a studied mathematics and has a background in hydraulics and morphology. He has worked for RWS since 2000. In recent years, the focus of his work has shifted towards the knowledge component. Finding knowledge gaps and then producing new system knowledge is the main part of his work. Until last year, he did not regularly share knowledge on Kennisplein because it seemed complicated. It turned out to be easy, so it is now routine to share knowledge on Kennisplein. His experience with findability is positive, but new reports have often not been shared by colleagues. Not all search terms are recognised, so broader search possibilities would be nice. The proposed structure seems applicable. The main functions are a logical choice, but water quality and ecology should be taken separately. Morphology and sediment management should be added as a separate category. The proposed knowledge types could be useful. Labels for data, governance and historical publications could be added. Geographical distinctions can be made by the upper/lower river regions, the Meuse and Rhine branches. Implementing this new structure into Kennisplein would be preferable.

1.	***Kunt u kort iets vertellen over uw functie, achtergrond en werkzaamheden bij RWS?***
Ik werk sinds 2000 bij RWS. Ik werk aan grote beleidsmatige projecten zoals ruimte voor de rivier, het deltaprogramma en Rivers2morrows. Ik heb een hydraulische en morfologische achtergrond, het beoordelen van projecten op hydraulische en morfologische effecten. De laatste jaren meer de kenniscomponent: wat weten we al over het rivierensysteem en het opvullen van gaten waar kennis nog mist of nodig is. Ik heb eerst aan de UT gewerkt als intermediair tussen RWS en de universiteit, nu heb ik die functie in Delft.	
2.	***Op wat voor manier bent u bezig met het produceren en/of het gebruiken/verwerken van rivierkennis?***
Rivierkennis is de kern van mijn werk. Ik ben bezig met het ontwikkelen van systeemkennis en het vinden van gaten. Mijn interesse ligt in het bovenrivierengebied en het bouwen met natuur. Ik ben minder betrokken bij het maken van modellen, maar die worden wel constant verbeterd met nieuwe kennis.	
3.	***Deelt u altijd nieuwe rivierkennis via Kennisplein of andere ontsluitingsmethodes? Waarom wel of niet?***
Dat deed ik voorheen nauwelijks. Een jaar of 3-4 geleden ben ik kennis gaan verspreiden naar collega's via e-mails. Sinds afgelopen jaar deel ik standaard kennis via Kennisplein. Ik deed het eerst niet omdat het ingewikkeld leek, maar het blijkt heel simpel te zijn. Een groot voordeel van het delen via Kennisplein is dat het ook direct openbaar wordt via PUC.	
4.	***Kunt u de documenten die u nodig hebt altijd vinden via Kennisplein e.d.?***
Mijn ervaringen zijn heel goed. Ik heb veel gezocht op bodemsamenstelling en ik heb het idee dat ik alles wel vind. Als ik iets niet kon vinden, bleek het later ook niet te bestaan. Historische data is goed vindbaar, dat is allemaal van papieren archieven overgezet naar Kennisplein. Nieuwe rapporten zijn veel minder online gezet door collega's, een grove schatting is 30-40%.	
5.	***Bent u tevreden met de huidige manier van kennisontsluiting of vindt u dat het beter kan? Zo ja, hoe?***
De manier van kennisontsluiting is wel oké. Soms heb ik het idee dat niet alle zoektermen herkend worden, een iets brede zoekmogelijkheid zou handig zijn. Typfouten worden bijvoorbeeld niet herkend, er wordt niet zoals op Google een alternatieve suggestie gegeven. Verder hebben documenten vaak rare namen (codes, projectnummers) waardoor ze minder goed vindbaar zijn.	
6.	***Hoe zou u zelf documenten met rivierkennis indelen (welke categorieën/wat voor structuur)? Waarom op deze manier?***

Wat zou kunnen is een soort verdeling tussen hydraulica, morfologie en ecologie en een aparte categorie voor data (gekoppeld aan hydraulica, morfologie en ecologie). Vanuit mijn technische hoek zijn labels als scheepvaart of bouwen met natuur van toepassing.

[Uitleg van resultaten literatuur]

7.	***Denkt u dat deze structuur bruikbaar is? Waarom wel of niet?***
Op zich lijkt mij dit een goede structuur.	
8.	***Denkt u dat de genoemde categorieën bruikbaar zijn? Waarom wel of niet?***
Gezien de recente discussies bij IRM is het gebruik van de hoofdfuncties een logische keus. Waterkwaliteit en ecologie zou ik niet samennemen, maar apart. Ook zou je een label governance kunnen toevoegen. Morfologie en sedimentmanagement zou ik wel als losse categorie toevoegen, al is het heel erg gekoppeld aan andere thema's. Er wordt namelijk veel onderzoek uitgevoerd met morfologie/sediment als hoofddoel. De types kennis zoals die nu genoemd zijn kunnen nuttig zijn, ik zou data/data-analyse toevoegen. Ook kun je historische publicaties (voor 1950) toevoegen. Geografisch kun je onderscheid maken tussen het boven- en benedenrivierengebied. De Rijntakken zou ik in een boom toevoegen, zodat je ook bijvoorbeeld alleen op de Waal kunt zoeken. Voor de kernwoorden heb ik liever een lijst woorden waar ik uit kan kiezen dan dat ik zelf woorden moet opgeven.	
9.	***Ziet u deze structuur liever op een nieuwe website of een bestaand platform? Waarom?***
Dat maakt mij niet heel veel uit, maar ik zou het op Kennisplein zetten.	

B.8 Interview Arjan Sieben

This interview was conducted om the 27th of May 2020, via Skype. The interview was in Dutch. An English summary with the most important information is provided below.

English summary
Arjan Sieben is river consultant in the field of hydraulics and morphology and he supervises research done by Deltares in these areas. He has a background in civil engineering, studied at TU Delft. He does not use Kennisplein or Helpdesk Water to share or find river knowledge, but this has no specific reason. It has never been a problem for his work that some knowledge is not available. It is very useful to have a structure that can represent overlapping categories. It could be useful to have a different entrance to the website per institute. Other than that, preferably categorise using commonly used themes. Themes like morphology or navigability and subsoil could be used. It is difficult to choose the 'right' categories, since the best way to showcase documents is to apply the same format that is used in the report itself. Searching for author might not work well for RWS documents, as these are usually written by teams rather than individuals. Make sure that a new website is well connected to existing systems.

1.	***Kunt u kort iets vertellen over uw functie, achtergrond en werkzaamheden bij RWS?***
Ik ben rivierkundig adviseur op het gebied van hydraulica en morfologie. Aspecten hiervan zijn aanleg- en het onderhoudsprojecten zoals baggercontracten, herstel van oevers en morfologische effecten van maatregelen. Ik begeleid onderzoek naar die aspecten door Deltares. Mijn achtergrond is civiele techniek, gestudeerd in Delft.	
2.	***Op wat voor manier bent u bezig met het produceren en/of het gebruiken/verwerken van rivierkennis?***
Kennis wordt gebruikt bij het maken van adviezen. Deze advisering is vaak weer inspiratie voor nieuwe kennisvragen en onderzoek, waaruit weer nieuwe kennisproducten ontstaan.	
3.	***Deelt u altijd nieuwe rivierkennis via Kennisplein of andere ontsluitingsmethodes? Waarom wel of niet?***
Ik zet zelf niet actief bestanden op Kennisplein. Ik heb Kennisplein eigenlijk nooit gebruikt, maar dat heeft geen speciale reden. Bij onderzoek dat ik begeleid proberen we een eigenaar in de regio te zoeken. Dat is het eerst verspreidingskanaal. Ook op de gemeenschappelijke schijf van RWS wordt kennis gedeeld. Er is een nieuwsbrief die vrij breed wordt rondgestuurd.	
4.	***Kunt u de documenten die u nodig hebt altijd vinden via Kennisplein e.d.?***
Ik gebruik Kennisplein en Helpdesk Water nooit, ook niet om dingen te vinden. Wel gebruik ik de service desk data, het dataloket van RWS.	
5.	***Bent u tevreden met de huidige manier van kennisontsluiting of vindt u dat het beter kan? Zo ja, hoe?***
Kennisontsluiting is erg afhankelijk van het netwerk van collega's. Een belangrijk onderdeel is dat je via collega's. Als kennis niet te vinden is, dan zoek je een oplossing. Als het je werk niet blokkeert, dan is het geen probleem dat kennis niet toegankelijk is. Dat is bij mij nooit het geval geweest.	
6.	***Hoe zou u zelf documenten met rivierkennis indelen (welke categorieën/wat voor structuur)? Waarom op deze manier?***
Wat makkelijk is, is een indeling per instituut. Als je kennis zoekt, kun je het vinden door te bedenken wie het heeft gemaakt. Ook indelen op thema's, het liefst de thema's zoals je ze in de meeste bibliotheken tegenkomt. Een beheerder heeft meestal ingedeeld op objecten (bruggen, etc.). Je kunt ook op meer algemene dingen indelen, zoals het thema morfologie of scheepvaart en ondergrond.	

[Uitleg van resultaten literatuur]

7.	***Denkt u dat deze structuur bruikbaar is? Waarom wel of niet?***

Jawel, hoe meer overlap je kunt, hoe beter. Anders word je bij het zoeken op een spoor gezet waar je het misschien niet vindt.	
8.	***Denkt u dat de genoemde categorieën bruikbaar zijn? Waarom wel of niet?***
Ik zou het zelf handig vinden om per kennisinstituut een ingang te hebben (RWS-diensten, Universiteiten, KNMI, Deltares). Voor mezelf heb ik andere categorieën in gebruik. Het is een beetje dubbel om te vragen om de 'juiste' categorieën, voor een goede etalage van documenten moet je je zo veel mogelijk houden aan het format van het onderzoek zelf. De hoofdfuncties zijn duidelijk, maar als je al langer dan 10 jaar in het vak zit, dank je niet altijd in deze hoofdfuncties. Zoeken op auteur kan binnen een overheid lastig zijn, de cultuur is dat het niet snel wordt gekoppeld aan auteurs omdat er vaak in groepen wordt gewerkt. In de wetenschappelijke wereld is dat juist andersom.	
9.	***Denkt u dat de nieuwe structuur kan worden toegepast op een bestaande website zoals Kennisplein of Helpdesk Water***
Je kunt proberen een dummy rapport op Kennisplein te delen om te testen of het goed vindbaar is. Dan zou je als gedachtenexperiment kunnen proberen aan te tonen dat het beter werkt dan Kennisplein.	
10.	***Denkt u dat het haalbaar is om op deze of een andere manier rekening te houden met de verschillende types gebruikers (RWS-werknemers, wetenschappelijke gemeenschap, marktpartijen)?***
Dat hangt ervan af. Als het vooral RWS'ers zijn die het gebruiken, dan kan het wel. RWS, Deltares, ingenieursbureaus hebben allemaal hun eigen systeem. In die zin is het een beetje dubbel om een nieuwe website te maken. Zorg dat het zo veel mogelijk aansluit bij wat er al loopt.	

Appendix C – Naïve Bayes model

This appendix contains the training data used to set up and train the Naïve Bayes model, the frequency of each keyword in every document and the MATLAB script used to extract keywords from the documents.

C.1 Training data

The documents and keywords used to train and test the model have been provided by Ralph Schielen, David Kroekenstoel, Hendrik Buiteveld and Emiel Kater. A list of the documents can be found in **Table C.1**.

Table C.1 – List of documents used as training data

Nr.	Title	Function	Geographic area
1	Duurzame Vaardiepte Rijntakken (DVR2)	NW	Rhine
2	Zoetwatervoorziening in Nederland	FA	All
3	Handboek Overstromingsrisico's op de kaart	WS	All
4	Kalibratie en aanpassingen HBV model voor de Rijn voor laagwater	FA	Rhine
5	Afvoerverdeling Rijntakken	WS	Rhine
6	Invloed van de kribverlagingsmaatregel bij het Pannerdensch Kanaal op de ontwikkeling van de ijsdammen	WS	Rhine
7	Consequentieanalyse GRADE	WS	All
8	Hotspotanalyse voor het Deltaprogramma Zoetwater	FA	All
9	Normering kanaaldijk Kreekrakpolder als regionale kering	WS, NW	Other
10	Overstromingsrisico Dijkring 42 Ooij en Millingen	WS	All
11	Eindevaluatie Ruimte voor de Rivier	WS	All
12	Eindrapport overstromingsrichtlijn	WS	Other
13	Hydraulische randvoorwaarden 2006 voor het toetsen van primaire waterkeringen	WS	All
14	Rijn en Maas	All	All
15	Nationale klimaatadaptatiestrategie 2016	All	All
16	Maximale Rijnafvoer bij Lobith	WS	Rhine
17	Internationaal gecoördineerd overstromingsrisicobeheerplan van het internationaal Rijndistrict, deel A	WS	Rhine
18	Overstromingsrisico's in Nederland	WS	All
19	Afvoer(beperkingen) van de Overijsselse Vecht in extreme omstandigheden	WS	Other
20	Knikpunten in het waterbeheer van het Maasstroomgebied als gevolg van klimaatverandering	All	Meuse
21	Grensoverschrijdende effecten van extreem hoogwater op de Niederrhein	WS	Rhine
22	Driejaarlijks rapport waterkwaliteit Scheldestroomgebiedsdistrict	WN	Other
23	Onderzoek Oplevering & Overdracht Hondbroeksche Pleij	WS	Rhine
24	Jaarrapport Maas 2018	FA	Meuse
25	Evaluatie van de reductie van het overstromingsrisico rekening houdend met de types van maatregelen en beschermingsdoelen conform richtlijn	WS	Rhine
26	Hoogwaterevaluatie Rijn 2011	WS	Rhine
27	Hoogwaterevaluatie Maas Winterseizoen 2010-2011	WS	Meuse
28	Samenvatting van onderzoek met GRADE naar implicaties van nieuwe klimaatprojecties voor rivierafvoeren	WS, NW, FA	All

29	Stapsgewijs naar autonoom varen	NW	All
30	Synthesedocument Rivieren	WS	Rhine
31	Rivierkundig beoordelingskader voor ingrepen in de grote rivieren	WS	Rhine
32	Leidraad rivieren	WS	Rhine
33	Hydraulische randvoorwaarden voor primaire waterkeringen	WS	All
34	IJsvorming op de Nederlandse Rivieren	WS	All
35	Beheerplan Natura 2000 Rijntakken	WS, WN	Rhine
36	Kerndocument netwerkbeheervisie RWS	All	All
37	Druk op de dijken 1995	WS	Rhine

Table C.2 contains the list of keywords used to identify the geographic areas. More keywords have been indicated, but not all of them occur in the training data. Those have been excluded from the model.

Table C.2 – List of keywords used to identify geographic areas

$\boldsymbol{w_1}$	Maas	$\boldsymbol{w_6}$	Regelwerk	$\boldsymbol{w_{11}}$	Kribverlaging
$\boldsymbol{w_2}$	Beleidslijn	$\boldsymbol{w_7}$	Splitsingspunt	$\boldsymbol{w_{12}}$	Lent
$\boldsymbol{w_3}$	Rijn	$\boldsymbol{w_8}$	Bodemkribben	$\boldsymbol{w_{13}}$	Vaste laag
$\boldsymbol{w_4}$	Ruimte voor de rivier	$\boldsymbol{w_9}$	Waal	$\boldsymbol{w_{14}}$	IJssel
$\boldsymbol{w_5}$	Afvoerverdeling	$\boldsymbol{w_{10}}$	Langsdam	$\boldsymbol{w_{15}}$	Dieren

Table C.3 and **Table C.4** contain the model parameters used to set up the Naïve Bayes model for the geographic areas.

Table C.3 – Model parameters $P(C_j)$ based on training data

$\boldsymbol{C_j}$	$\boldsymbol{P(C_j)}$
$Meuse$	$P(Meuse) = 18/37 = 0.486$
$Rhine$	$P(Rhine) = 30/37 = 0.811$

Table C.4 – Model parameters $P(w_i \mid C_j)$ based on training data

	$\boldsymbol{w_1}$	$\boldsymbol{w_2}$	$\boldsymbol{w_3}$	$\boldsymbol{w_4}$	$\boldsymbol{w_5}$	$\boldsymbol{w_6}$	$\boldsymbol{w_7}$	$\boldsymbol{w_8}$
$\boldsymbol{P(w_i \mid Meuse)}$	0.071	0.087	0.120	0.224	0.101	0.006	0.009	0.034
$\boldsymbol{P(w_i \mid Rhine)}$	0.022	0.024	0.017	0.038	0.079	0.003	0.001	0.029
	$\boldsymbol{w_9}$	$\boldsymbol{w_{10}}$	$\boldsymbol{w_{11}}$	$\boldsymbol{w_{12}}$	$\boldsymbol{w_{13}}$	$\boldsymbol{w_{14}}$	$\boldsymbol{w_{15}}$	
$\boldsymbol{P(w_i \mid Meuse)}$	0.009	0.003	0.011	0.004	0.001	0.002	0.001	
$\boldsymbol{P(w_i \mid Rhine)}$	0.000	0.033	0.002	0.069	0.007	0.038	0.011	

C.2 Keyword matrices

Figure C.1 contains the frequency of each keyword identifying the main functions in the training data. This is the output of the MATLAB script and used to set up the Naïve Bayes model in Excel.

Nr.	hoog	risico	overs	dijk	wate	herh	slach	scha	overs	overs	frequ	afvoe	faalk	laagv	afvoe	drink	verdi	zout	wate	klass	vaste	ladin	moda	door	vaarc	droo	uiter	wate	flora	faun	vege	chlor	temp	klima	stuw
1	26	8	2	3	58	0	0	1	0	0	0	30	0	27	19	0	0	0	0	0	45	0	0	0	7	0	9	0	1	1	0	0	0	10	10
2	6	24	4	50	88	6	0	52	0	2	9	90	0	1	2	79	10	113	10	13	0	2	2	3	31	285	1	13	1	3	23	34	23	202	15
3	8	191	264	87	51	1	0	15	0	8	12	10	0	0	0	8	0	1	0	21	0	0	0	0	0	5	0	0	0	0	0	0	0	0	1
4	6	0	0	0	15	0	0	0	0	0	0	147	0	131	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	32	1	1
5	10	0	1	4	24	0	0	0	1	0	0	200	0	0	54	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	1	0
6	11	5	9	4	11	2	0	1	1	4	1	40	0	0	3	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	5	1	5
7	0	0	7	8	31	43	0	0	0	0	0	87	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0
8	0	5	0	15	92	3	0	4	0	0	10	196	0	5	1	27	5	61	31	3	0	3	0	0	34	33	0	29	0	0	0	19	3	90	37
9	1	9	33	26	47	1	1	78	0	0	3	1	0	0	0	1	0	0	0	23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
10	23	227	405	834	75	0	98	125	134	20	6	12	166	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	1	11
11	7	185	5	26	13	0	0	4	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	3	0	0	2	3
12	1	202	459	7	3	0	1	20	2	0	9	4	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4	0	0	0	2	0	7	78	2
13	38	0	3	495	150	0	0	2	0	8	191	121	1	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
14	66	10	45	109	58	0	2	23	0	0	0	100	0	17	4	15	1	20	4	0	1	6	0	0	3	16	35	6	0	4	6	0	6	18	60
15	0	52	25	8	1	0	2	55	2	0	5	8	0	0	0	6	0	0	0	0	0	0	1	0	0	18	0	0	3	3	0	0	17	293	0
16	8	0	38	52	19	1	0	0	0	0	0	135	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	6	1
17	91	199	202	18	25	14	0	11	0	6	1	44	0	1	0	1	0	0	0	1	0	0	0	0	0	0	4	0	1	1	0	0	1	21	5
18	8	219	419	54	23	3	44	89	15	5	9	24	0	0	0	5	0	1	0	16	0	0	0	0	0	0	1	0	0	0	0	0	0	41	0
19	3	1	27	16	25	8	0	0	0	0	0	133	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4
20	12	17	4	11	192	6	0	20	0	0	17	257	1	15	0	32	16	13	8	12	0	0	0	0	0	141	1	18	3	8	3	2	72	433	30
21	33	0	45	326	90	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	3	0	1	0	1
22	1	6	5	0	0	0	0	2	0	0	5	4	0	9	0	1	0	5	0	14	0	0	0	0	0	1	0	0	2	4	0	3	11	5	1
23	0	11	0	1	3	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	18	1	1	4	0	0	13	0	0	6	146	0	6	1	234	8	7	11	2	0	0	0	0	0	108	0	7	0	0	0	4	21	35	13
25	32	127	127	4	20	8	0	105	27	0	7	1	0	0	0	2	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
26	271	0	0	14	201	1	0	0	0	0	5	414	0	1	17	0	0	0	0	2	0	0	0	0	0	3	11	0	0	0	0	0	0	0	59
27	250	0	0	1	152	1	0	0	0	0	0	302	0	0	13	0	0	0	0	3	0	0	0	0	0	8	1	0	0	0	0	0	1	0	76
28	33	4	18	23	2	4	0	0	1	3	5	155	0	15	0	0	0	2	0	0	0	0	0	0	0	5	0	0	0	0	0	0	7	54	1
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	0	0	0	0	0	0	0	0	0	0	0	5
30	137	181	151	1169	160	1	22	56	19	7	0	272	3	4	51	5	0	1	3	3	0	0	0	0	3	1	21	4	2	2	2	0	2	190	16
31	60	9	4	71	227	14	0	14	3	0	17	327	0	8	62	1	0	2	2	48	1	0	0	0	0	0	9	0	0	0	44	0	0	0	26
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	88	1	7	522	136	0	0	1	0	6	52	49	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4
34	1	0	1	0	64	0	0	0	0	0	0	29	0	0	9	0	0	2	0	0	0	0	0	0	0	0	6	0	0	0	0	0	92	19	4
35	28	16	11	146	10	0	3	14	0	0	10	30	0	0	0	4	0	0	0	0	0	0	0	0	0	22	108	0	4	20	49	0	0	5	38
36	1	8	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	7	0
37	156	1	10	305	138	3	0	31	0	4	11	88	1	1	0	0	0	0	0	0	0	0	0	0	0	4	11	0	0	0	0	0	5	1	5

Figure C.1 *– Frequencies of keywords identifying main functions in training data*

Figure C.2 contains the frequency of each keyword identifying the geographic areas in the training data. This is the output of the MATLAB script and used to set up the Naïve Bayes model in Excel.

Nr.	maas	beleidslijn	rijn	ruimte vo	afvoerver	regelwerk	splitsings	bodemkri	waal	langsdam	kribverlag	lent	vaste laag	ijssel	dieren
1	0	0	5	0	19	0	19	10	0	99	3	0	45	0	0
2	2	0	16	0	2	0	0	0	0	0	0	13	0	7	0
3	0	1	1	0	0	0	0	0	0	0	0	0	0	8	0
4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
5	0	0	17	0	54	133	22	0	0	0	0	0	0	0	0
6	0	0	2	0	3	0	2	0	0	0	98	1	0	0	0
7	0	0	12	0	2	0	1	0	0	0	0	0	0	0	0
8	2	0	15	0	1	1	0	0	1	0	0	1	0	2	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	6	0	17	0	0	0	0	0	0	3	5	0	0	3	2
12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
13	0	0	19	0	3	0	3	0	1	0	0	0	0	16	0
14	31	2	27	6	4	0	8	2	1	0	2	0	1	1	23
15	0	0	0	0	0	0	0	0	2	0	0	0	0	2	8
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	48	3	0	0	0	0	3	0	1	2	0	0	0
18	1	0	6	1	0	0	0	0	0	0	0	0	0	2	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0
20	25	5	1	0	0	0	0	0	0	0	0	0	0	0	6
21	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	4	0	0	6
23	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0
24	6	0	5	0	1	0	0	0	0	0	0	3	0	0	3
25	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0
26	0	0	76	0	17	0	4	0	1	0	0	0	0	0	0
27	40	0	0	0	13	0	0	0	0	0	0	0	0	0	0
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
30	15	5	6	8	51	0	21	0	4	14	2	0	0	5	1
31	11	2	3	1	62	5	36	0	0	1	0	1	1	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	1	0	19	0	0	0	0	0	1	0	0	3	0	12	0
34	0	0	7	0	9	0	1	0	3	0	0	50	0	0	0
35	0	2	7	0	0	0	0	0	2	3	0	2	0	35	4
36	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
37	0	0	3	3	0	0	1	0	0	0	0	0	0	0	3

Figure C.2 *– Frequencies of keywords identifying geographic areas in training data*

Figure C.3 contains the frequency of each keyword identifying the main functions in the test data. This is the output of the MATLAB script and input for the Naïve Bayes model in Excel.

Nr.	hoog	risico	overs	dijk	wate	herha	slach	schac	overs	overs	frequ	afvoe	faalk	laagw	afvoe	drink	verdr	zout	wate	klass	vaste	ladin	moda	doorv	vaarc	droog	uiter	wate	flora	fauna	veget	chlor	temp	klima	stuw
1	0	19	3	14	16	0	0	107	0	0	0	28	1	2	0	55	182	48	7	0	0	1	0	0	5	177	0	56	3	4	4	17	6	10	5
2	0	0	2	1	97	0	0	0	0	0	0	29	0	1	0	0	0	0	6	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	19
3	68	0	49	381	141	35	0	0	3	0	5	96	0	1	0	2	0	0	0	2	0	0	0	0	0	6	0	0	0	0	8	0	0	0	17
4	34	2	12	154	23	0	0	0	0	0	4	12	0	3	0	0	0	0	0	0	0	0	0	0	0	4	32	0	0	1	12	0	0	5	3
5	23	7	0	3	52	0	0	17	0	0	0	39	0	2	3	0	0	0	2	3	24	6	1	0	6	0	12	0	1	1	1	0	0	5	4
6	29	3	0	1	21	0	0	2	0	0	20	33	0	0	3	0	0	0	0	0	0	0	0	0	0	0	4	0	0	1	10	0	0	0	3
7	5	0	4	50	3	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	2	0	0	0	0	0
8	45	1	0	11	96	0	0	6	0	0	14	125	0	7	42	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	7	0	0	3	17

Figure C.3 *– Frequencies of keywords identifying main functions in test data*

Figure C.4 contains the frequency of each keyword identifying the geographic areas in the test data. This is the output of the MATLAB script and input for the Naïve Bayes model in Excel.

Nr.	maas	beleidslijn	rijn	ruimte vo	afvoerver	regelwerk	splitsingsp	bodemkri	waal	langsdam	kribverlag	lent	vaste laag	ijssel	dieren
1	1	1	3	0	0	0	0	0	1	1	0	1	0	10	3
2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	209	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	5	0	0	0	0	1	0	0	1	0	0	8
5	0	1	0	0	3	0	4	0	0	13	3	0	24	0	0
6	0	0	4	0	3	11	1	0	1	0	0	3	0	0	0
7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
8	0	0	69	0	42	0	12	0	1	0	0	0	0	0	0

Figure C.4 *– Frequencies of keywords identifying geographic areas in test data*

C.3 MATLAB script

```
%% General info

% This MATLAB script is able to load documents, extract the file
text and
% then search for predefined keywords indicating main functions and
% geographic areas in each document. This is done for a set of
training
% documents and a set of test documents.

% The MATLAB script produces four matrices containing the frequency
of
% every main function keyword and every geographic area keyword in
all
% training documents and all test documents. These four matrices
containing
% keyword frequencies are exported to Excel to build and then test a
Naïve
% Bayes model.

% Made by T.J.A. Luyten, June 2020

%% Load documents and extract file text
clear, clc

%Documents are loaded into MATLAB for every .pdf file in the
training
%directory and the name of each document is extracted
trainDir = "C:\Users\Tim\Documents\MATLAB\NBTrain";
trainFiles = dir(fullfile(trainDir,'*.pdf'));
FilesTrain = extractfield(trainFiles,'name');

%The file text of every document in the training directory is
extracted
for i = 1:length(FilesTrain)
    file = trainDir+"\"+FilesTrain{i};
    Train(i) = {extractFileText(file)};
end

%Documents are loaded into MATLAB for every .pdf file in the test
%directory and the name of each document is extracted
testDir = "C:\Users\Tim\Documents\MATLAB\NBTest";
testFiles = dir(fullfile(testDir,'*.pdf')); %LL
FilesTest = extractfield(testFiles,'name');

%The file text of every document in the test directory is extracted
for i = 1:length(FilesTest)
    file = testDir+"\"+FilesTest{i};
    Test(i) = {extractFileText(file)};
end

%% Define keywords

%Keywords identifying main functions are defined
KeywordsF =
{'hoogwater','risico','overstroming','dijk','waterstand',...
```

```
    'herhalingstijd','slachtoffers','schade','overstromingskans',...
'overschrijdingskans','frequentie','afvoer','faalkans','laagwater',...
    'afvoerverdeling','drinkwater','verdringingsreeks','zout',...
    'waterverdeling','klasse','vaste laag','lading','modaliteit',...
'doorvaarthoogte','vaardiepte','droog','uiterwaarde','watertekort',...
'flora','fauna','vegetatie','chloride','temperatuur','klimaat','stuw
'};

%Keywords identifying geographic areas are defined
KeywordsG = {'maas','beleidslijn','rijn','ruimte voor de rivier',...
'afvoerverdeling','regelwerk','splitsingspunt','bodemkribben',...
    'waal','langsdam','kribverlaging','lent','vaste
laag','ijssel','dieren'};

%% Create matrices with keyword frequencies

%Empty matrices for training and testing are created, to be filled
with the
%frequency of each keyword per document. This is done for keywords
%identifying main functions and geographic areas.
TrainMatrixF = zeros(length(FilesTrain),length(KeywordsF));
TestMatrixF = zeros(length(FilesTest),length(KeywordsF));
TrainMatrixG = zeros(length(FilesTrain),length(KeywordsG));
TestMatrixG = zeros(length(FilesTest),length(KeywordsG));

%Training matrix is filled with the frequency of main function
keywords in
%a document
for i=1:length(FilesTrain)
    m = Train{i};
    for j=1:length(KeywordsF);
    word = strfind(m,KeywordsF(j));
    TrainMatrixF(i,j)=length(word);
    end
end
TrainMatrixF;

%Test matrix is filled with the frequency of main function keywords
in
%a document
for i=1:length(FilesTest)
    m = Test{1,i};
    for j=1:length(KeywordsF);
    word = strfind(m,KeywordsF(j));
    TestMatrixF(i,j)=length(word);
    end
end
TestMatrixF;
```

```
%Training matrix is filled with the frequency of geographic area
keywords in
%a document
for i=1:length(FilesTrain)
    m = Train{i};
    for j=1:length(KeywordsG);
    word = strfind(m,KeywordsG(j));
    TrainMatrixG(i,j)=length(word);
    end
end
TrainMatrixG;

%Test matrix is filled with the frequency of geographic area
keywords in
%a document
for i=1:length(FilesTest)
    m = Test{i};
    for j=1:length(KeywordsG);
    word = strfind(m,KeywordsG(j));
    TestMatrixG(i,j)=length(word);
    end
end
TestMatrixG;

%% Export keyword frequency matrices

%Frequencies of main function keywords in training data are exported
to
%Excel along with column and row titles.
xlswrite('NaïveBayes',TrainMatrixF,'Training data_F','B2')
xlswrite('NaïveBayes',FilesTrain','Training data_F','A2')
xlswrite('NaïveBayes',KeywordsF,'Training data_F','B1')

%Frequencies of main function keywords in test data are exported to
%Excel along with column and row titles.
xlswrite('NaïveBayes',TestMatrixF,'Test data_F','B2')
xlswrite('NaïveBayes',FilesTest','Test data_F','A2')
xlswrite('NaïveBayes',KeywordsF,'Test data_F','B1')

%Frequencies of geographic area keywords in training data are
exported to
%Excel along with column and row titles.
xlswrite('NaïveBayes',TrainMatrixG,'Training data_G','B2')
xlswrite('NaïveBayes',FilesTrain','Training data_G','A2')
xlswrite('NaïveBayes',KeywordsG,'Training data_G','B1')

%Frequencies of geographic area keywords in test data are exported
to
%Excel along with column and row titles.
xlswrite('NaïveBayes',TestMatrixG,'Test data_G','B2')
xlswrite('NaïveBayes',FilesTest','Test data_G','A2')
xlswrite('NaïveBayes',KeywordsG,'Test data_G','B1')
```