

Research Document

Flooid



Students:

Nikita Gavrilov

Ilya Tsakunov

Alexander Vereshchagin

Victor Plesciuc

Gergana Agorasteva

Martens Yaniek

# Management summary

**Overview**

Datastic will assist a start-up Flooid with a project. The project will entail the development of a prototype for a knowledge network, where parties such as Fontys, Breda Universities of Applied Sciences and Delft University of Technology, will be able to discover information via a web interface powered by the Natural Language Programming idea.

The network will be created with the assistance of Flooid's partner EyeOnText, who offers Wowool SDK training. It is a cross-platform utility with Python and C++ language bindings. This tool assists in the ingestion of unstructured textual input and the delivery of structured semantic objects such as entities, sentiments, profiles, facts, and linkages.

**Problem**

Current problem

Many organizations of this community share a lot of information about themselves. This information is unstructured and people outside of the company but within the community cannot easily access it. This leads to time-wasting problems while two or more parties are working on similar projects apart.

Potential problems

* Getting out of project scope
* Misunderstanding withing a team
* Lack of technical knowledge
* Not meeting the deadlines

**Solution**

At the end of the project, the intended project result is a Proof of Concept of an information network that the community can and wants to use. In addition to the proof of concept, there will be a Business Plan created.

**HIGHLIGHTS**

* Datastic will deliver a working prototype as proof of concept (PoC) of application and advice about it.
* Target audience: Community of Organizations and universities.
* Responsibility of Datastic: Documentation, Data Aggregation, Natural Language Processing and Machine Learning, Obtaining end user feedback, Marketing.
* Team consists of 9 developers + process coach.
* Communication mostly via Zoom, weekly meetings on Friday.
* Business plan where business model will be described.
* Deadline: January 15th.

# Preface

*Company and project Information*

Flood is a start-up company which supports knowledge sharing activities between companies. It primarily serves as a networking industry consulting partner with platforms that use natural language processing and knowledge graphs. Flooid functions as a trigger in the system transition, influencing how employees perform, thrive, and care for the environment. They accomplish this by bringing together people, information, and artificial intelligence to boost personal qualities while fluidly connecting the online and offline worlds. Flooid has been a model of community-based learning and innovation supported by a platform using the Natural Language Processing concept and has a list of partner universities and mindlabs that are ready to be a part of the networking environment. The network will be designed with the help of Flooid’s partner EyeOnText, that provides workshops on the Wowool SDK. It is a cross-platform tool which helps to Ingest unstructured textual data and deliver structured semantic objects, such as entities, sentiments, profiles, facts and links.

Here comes the project with the DDBL (Data Driven Business Lab minor) project group - Datastic. The group consist of 6 international and dutch students from Fontys university who combine all their different study backgrounds to achieve the end result from this collaboration – to create a Proof of Concept of information network, core and frontend of prototypes, as well as advice for Flooid on project execution and a business plan.

*Relevance for Readers*

Within this document, the result of Datastic’s work over the creation and execution of the information network can be found. All content in this report is relevant for the contributing parties, which are the Flooid, all their partners and Fontys university.

*Word of Thanks*

A word of appreciation is sent to the Flooid company and its CEOs. The Datastic team wants to thank Stefan Groenendal and Rene Katerberg for their support and the chance to take on this exciting project. Sincere appreciation to Carolina Rubio and Philippe Forest for providing workshops and assisting with the technical aspects of the work. Big gratitude also to Simona Orzan for her work as a content coach of the group and her support during the whole project.

# Table of Contents

[Management summary 1](#_Toc93407336)

[Preface 2](#_Toc93407337)

[1 Introduction 4](#_Toc93407339)

[1.1 Purpose of the document 4](#_Toc93407340)

[1.2 Purpose of the project 4](#_Toc93407341)

[1.3 Parties involved 4](#_Toc93407342)

[1.4 Readers’s Guide 4](#_Toc93407343)

[2 Research results 5](#_Toc93407344)

[2.1 Web scraping 5](#_Toc93407345)

[2.2 NLP 6](#_Toc93407346)

[2.3 Neo4j 12](#_Toc93407347)

[3 Research questions 15](#_Toc93407348)

[3.1 What kind of data should be in network? 15](#_Toc93407349)

[3.2 What are the community's needs for the network? 15](#_Toc93407350)

[3.3 How can Datastic make a network using the tools (web scrapping, NLP, graphs) that are chosen? 15](#_Toc93407351)

[3.4 How can the data from network be stored? 15](#_Toc93407352)

[3.5 How can the network be presented to the end-user? 15](#_Toc93407353)

[3.6 How can this project be transformed to a business? 15](#_Toc93407354)

[4 Conclusion & recommendations 16](#_Toc93407355)

[4.1 Conclusion 16](#_Toc93407356)

[4.2 Recommendations 16](#_Toc93407357)

[5 Appendix 17](#_Toc93407358)

# Introduction

## Purpose of the document

In this document will the results of the project be discussed. The approach to the results and the conclusions that can be taken from the results. Further, there will be recommendation about how to proceed.

## Purpose of the project

The main idea of this project is to have a network is to share partner’s knowledge and connect people to work together based on needed skills with a help of neo4j graph data platform. This network is also supposed to prevent time-wasting problems while two or more parties are working on similar projects apart.

## Parties involved

|  |  |
| --- | --- |
| Name | Function |
| Datastic | Project group |
| Rob Verhoeven | Process Coach DDBL |
| Simona Orzan | Content Coach DDBL |
| Flooid | Client/Partner |
| Carolina Rubi | EyeOnText, NLP-expert |

## Readers’s Guide

The first chapter that will be discussed in this document is the research results. In this chapter will the approach and the results be discussed from each method that is used in this project. Furthermore, will each method have a conclusion. Chapter 3 draws conclusions and gives advice to the project.

# Research results

In this chapter the result of the research will be discussed.

## Web scraping

To create a final deliverable, the data is needed. In scope of this project, the data should contain information about community's partners, including basic information about the company, projects it is involved in, founders’ information etc. The way to get this information is to extract it from the official partners’ websites. It could be done manually, however as the number of partners is about 20 and this number will further increase, the manual approach is not suitable. Moreover, the information on websites is updated regularly, so the new information should be in the database as well. Here the web scraping comes into. Generally, web scraping is a process of extracting data from websites. There are different tools for web scraping, such as SaS solution and Python libraries. For this project Datastic decided to use BeatifulSoup4 (BS4), which is a Python library. The principle of this tool as based on html structure of web page. BS4 allows to extract information between html tags, or just the whole page content.

### Approach

Datastic tried 2 main approaches. The first approach was based on extracting only necessary information (founders, stack of technologies etc.). To do that, Datastic created a list of partners with websites. Then for each website the content between html tags was extracted. Datastic was mainly searching for div-tags with specific classes.

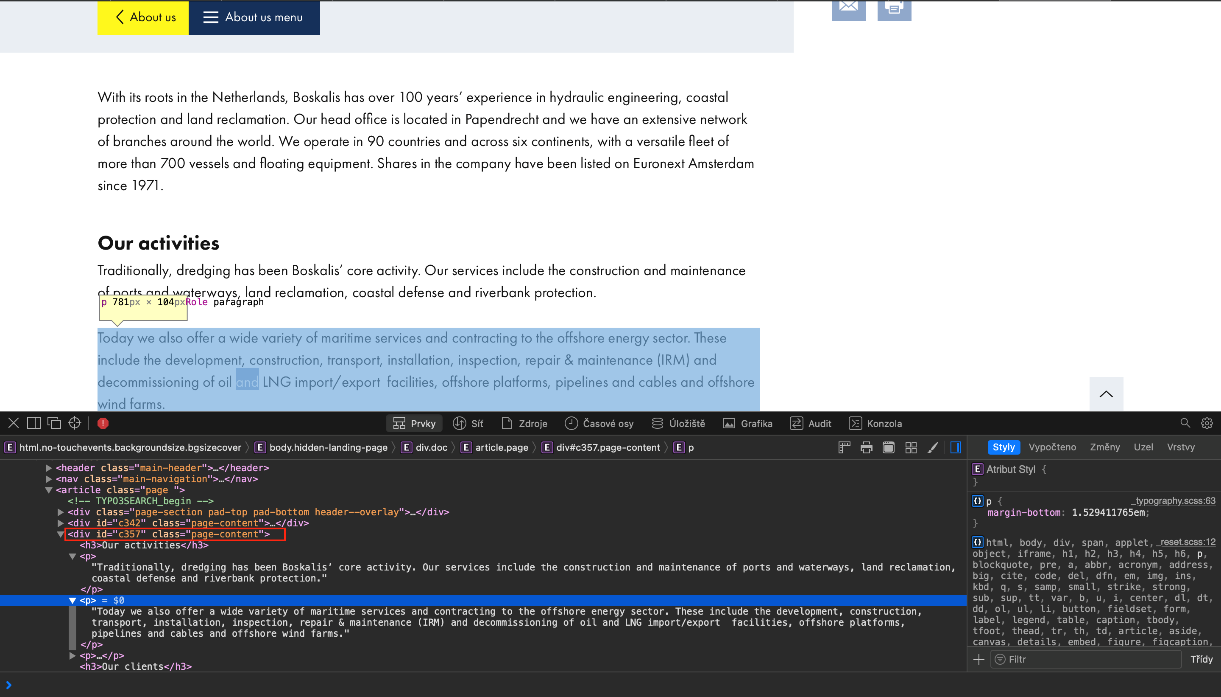


Figure 1 Div-tags with specific classes

The second approach was not expecting extracting information between specific html tags, but the whole page content. That was done because of fact that further during the project NLP be used, so extracting the necessary information would be the task of NLP.

The next problem was that partners have different structure of pages with projects. So, in addition to general script Datastic created separate scripts for several partners to extract information about projects.

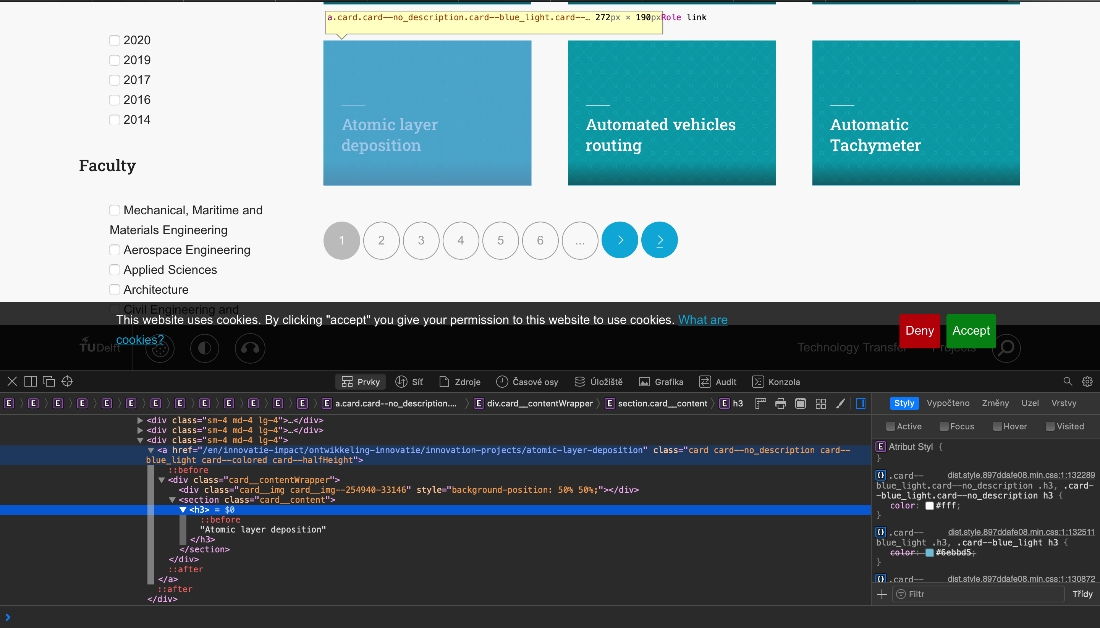


Figure 2 Div-tags with specific classes

For instance, some partners’ websites contain pagination, the structure of which differs from website to website. To iterate through each page Datastic mostly used an approach of changing the structure of url. In the example below the first rectangle is a year of projects and the second shows the page number.

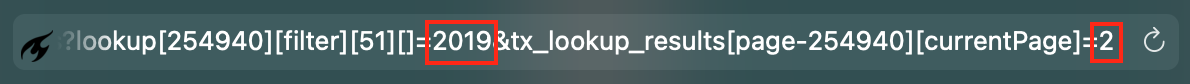


Figure 3 Pagination

### Result

The result of web scraping part of this project is a bunch of Python scripts with documentation. The general script is responsible for scraping the whole pages’ content. Another script is made to extract the information about projects from 3 partners universities (TU Delft, Fontys, BUAS). The output of scripts is a folder which contains folders for each company with .txt files inside.

### Conclusion

To sum up, Datastic prepared Python scripts that can extract the necessary information from partners’ websites. Regarding some topics (like projects partner is involved in) another script is needed. Unfortunately, as was said above, different websites have different html structure, so the script often needs to be adjusted.

## NLP

Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.

NLP: interaction between computer - human

* Speech recognition
* Language generation
* Machine translation
* Natural language understanding

### Approach

**Wowool language**

This language already has pre-installed libraries with different languages (Add list of languages) and domains (add list and explain)

Wowool is a text-oriented pattern matcher engine. It has two main components:

**Lexicons (Like vocabulary)**

Easiest way to annotate, just matching literal or stems; add knowledge of the world, simplify rule writing.

Example: In our case we had a problem that NLP was detecting words (Engels,  
manage, As, ways, Sophia Rotterdam) as city. Datastic had to filter out those words and let the program know that they are not cities ( = NoCity)

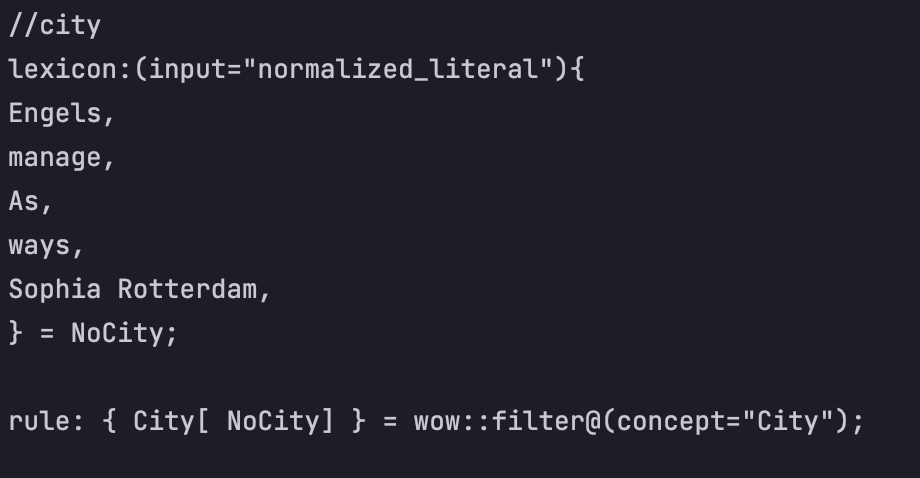


Figure 4 City lexicon

(input="normalized\_literal") - Matches the literal token regardless of capitalization or any accents or special characters.

Lexicon for Identifying different phone formats from web sites. The problem is that every website uses its own format, but they are limited.

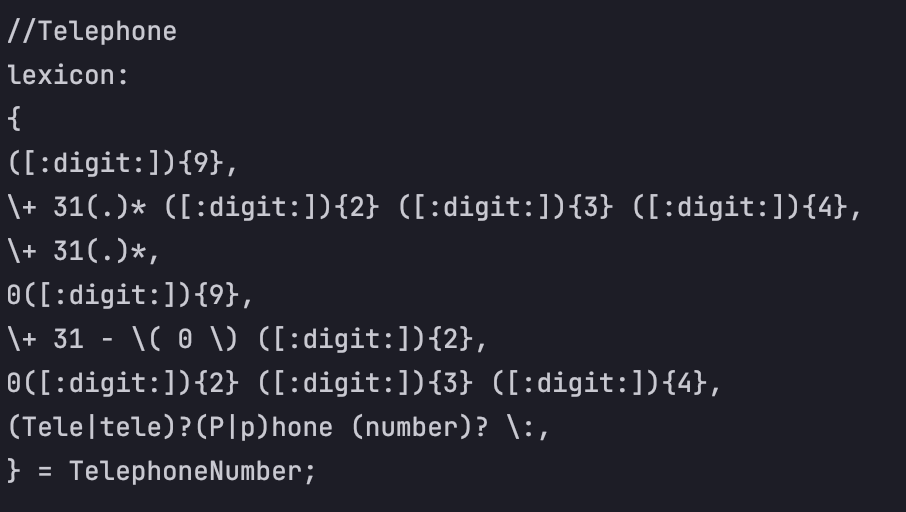


Figure 5 Telehone lexicon

* ([:digit:]){9} - if there is 9 digits - it is a Telephone number (Random example: 061234567).
* (.)\* - space between numbers.

**Rules**

In comparison to lexicons, rules provide a more expressive tool for capturing linguistic patterns. The use of lexicons in the creation of dictionaries, gazetters, and rules in the creation of larger and more complex patterns:

With rules you can:

* address stems, literals and part of speech, as done in lexicons
* use other annotations (from lexicons or other rules)
* Address the context of the expression
* Make sub-annotations
* Use the relative position of the sentence
* Make coreference rules
* Use filters
* call python

**Example:**

To identify what kind of technologies the given text has, to extract this data and mark it as “Technology”. The rule says that if there are more than 0 words and less or equal 3 before word “Device” and those words are Noun or Adjective or Adverb – this combination of words is Technology.

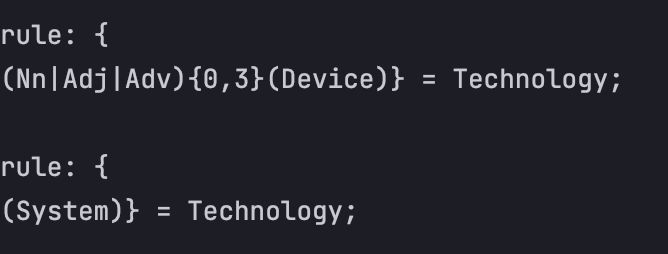


Figure 6 Rules for technology

**Link file**

Link file used to set a connection between needed topics and create visual relationship between them.

Example: Datastic team would like to see what kind of projects every university has and how do they cross with each other based on topics.



Figure 7 Example on topic connections

This code set the connection between uploaded file to scan (“Doc”) and names on university projects with “Projects” label. But to see what those projects are about there must be created one more connection between project names and topics. This way will help to see if different universities have same project topics that they could potentially work on together.

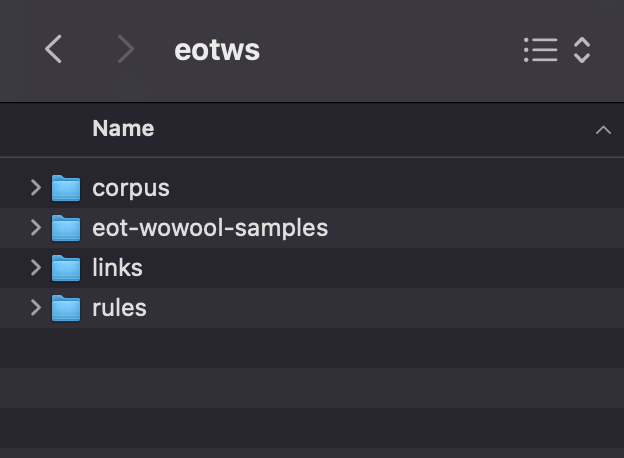
**Domain**

A domain is a collection of wow files (rules and lexicons) that are grouped for a single purpose (healthcare, finance, recruitment). Those files work as a library to extract needed data from txt files based on topic required to identify.

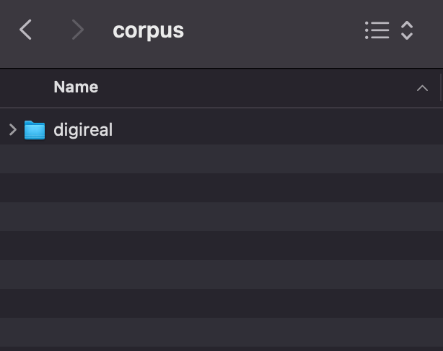
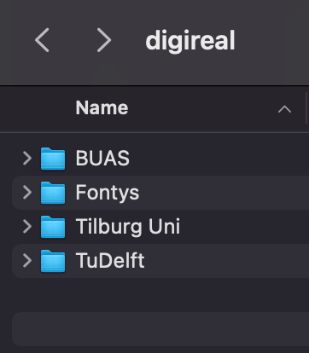
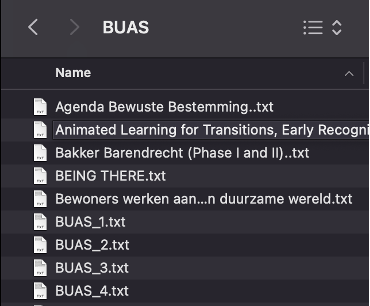
**Environment**

File storing solution

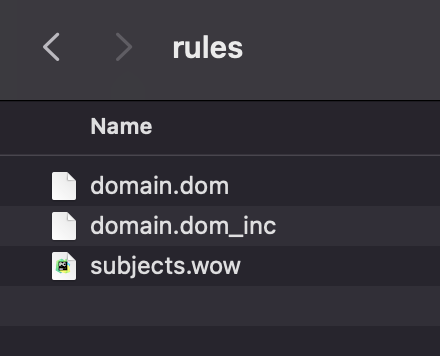
All files and folders are stored in “eotws” folder:



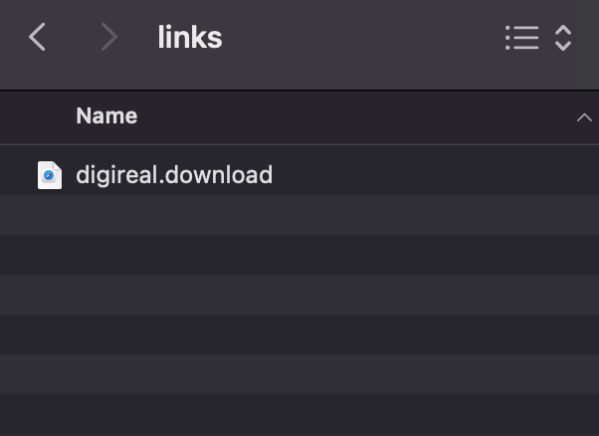
Files to scan:

Rules for NLP are adjusted and stored in subjects.wow file.



Links that create connection between different topics.



### Result

The result is that Datasctic has created a list of lexicon and rules to clean and adjust the accuracy of correct result. To do so, following lexicons, rules and connections were created:

* NoCity
* NoPerson
* University
* Topics
* Position
* TelephoneNumber
* PObox – separation of physical addresses and post box addresses
* Univercity\_Projects
* Partner
* HR
* Organization

### Conclusion

The Wowool NLP comes with a lot of knowledge encoded, for your benefit, it is possible to choose from several pre-set domains to get started and that you can modify or build upon. You can use the Wowool language to completely customize your solution and connections.

## Neo4j

Neo4j is a native graph database that stores and manages data in a more natural, connected state, maintaining data relationships that deliver efficient queries, deep context for analytics, and an easily modifiable data model.

In a graph database, relationships are stored at the individual record level, while a relational database uses predefined structures (table definitions). Relational databases are faster when handling huge numbers of records because the structure of the data is known ahead of time. This also leads to a smaller memory footprint. Graph databases don’t have a predefined structure for the data which is why each record must be examined individually during a query to determine the structure of the data. The fundamental components of a graph database are Nodes, Relationships, Labels and Properties.

Nodes: main entities in a graph

Relationships: connection between nodes

Labels: attributes that group similar nodes together.

Properties: key/value pairs stored within nodes or relationships

### Approach

To create a knowledge graph that can allow us to view the relationships between different companies/universities/employees etc. we need to use a graph database.

Neo4j is ideal for storing and showcasing data and insight generated by using Natural Language Processing (NLP) (insert reference). EyeOnText(reference) analyses the text data that we have and allows us to understand the meaning without having to read the text itself.

Once a text is analysed, we can then create a structure that allows us to save the data in a graph format and showcase it in Neo4j.

Visualizing data in Neo4j is done using the Cypher Query Language, which is easy to learn because of its similarity to SQL and other languages, and its intuitiveness.

### Result

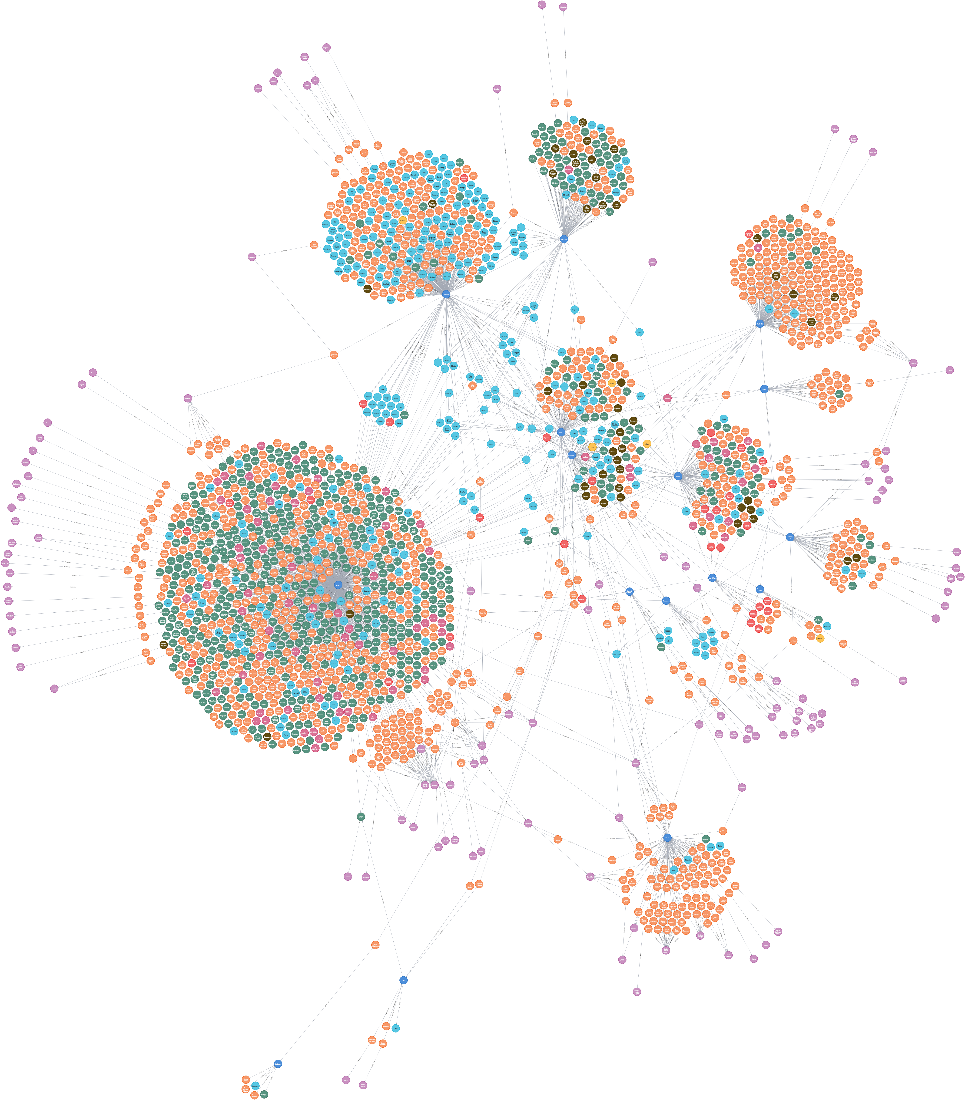


Figure 8 Example of network

As you can see in Figure the resulting graph database created based on processed web scraped content analysed with NLP.

Each of the individual circle represents a type of information and is connected to other nodes, thus allowing us to see connections that were not there before. In perspective, the algorithm that helps us analyse the text has its flaws and so the graph isn’t as accurate as we would imagine, but this gives us the possibility to showcase what would be possible, if the algorithm were to be improved.

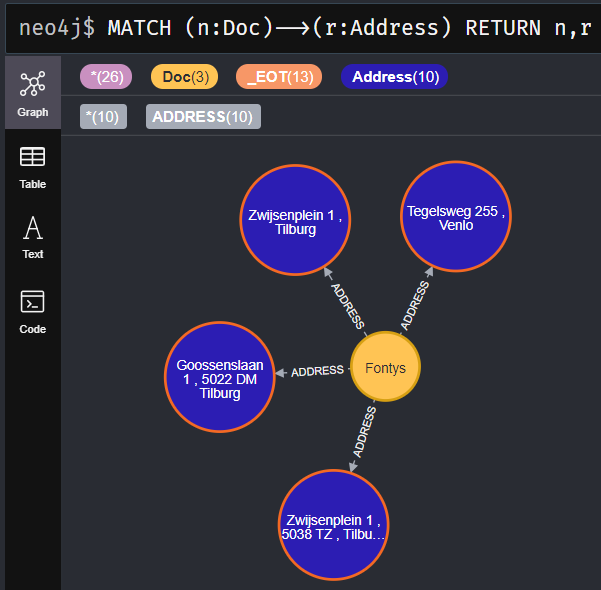


Figure 9 The graph

Cypher command for showing the Addresses connected to Fontys

### Conclusion

In conclusion, using Neo4j can allow us to see insights into data that aren’t usually possible in a very intuitive way. This can allow partners to create a knowledge sharing network where data transparency is key.

# Research questions

## What kind of data should be in network?

In scope of this project, the data should contain information about community's partners, including basic information about the company/university, projects it is involved in, founders’ information, positions, contact information etc. The way to get this information is to extract it from the official partners’ websites via web scraping approach. This list of network data can be extended based on partner’s request and needs. In this case, Datasctic included the following data:

* City names
* Persons
* University names
* Physical addresses
* Topics
* Job positions
* Telephone numbers
* PObox addresses
* University Projects names
* Emails

## What are the community's needs for the network?

Datastic had a meeting with Petra van Dijk from Mindlabs about needs the following is mentioned:

* Show where the company connections are, to which sectors they belong, places and the connection to other companies. In other words, the visualisation of the network.
* Get an inside of the ecosystem that is already in place and where the network can be expanded (new connections outside the network)
* Have an inside in methodologies that are used in the projects from the companies inside the network.

## How can Datastic make a network using the tools (web scrapping, NLP, graphs) that are chosen?

Datastic performed research on techniques which allow us to retrieve content from websites that are connected to the community. Based on the research and Flooid’s advice the technique chosen is called ‘webscraping’. With the help of BeautifulSoup library the team retrieved the content from webpages. Datastic gathered this content and structured it into a folder containing textfiles named after the entities that provided the data.

Secondly, Wowool language was used to process the webscraped data and make it ready to be Nodes, Relationships and of future properties of Neo4j database. In order to identify the wods by machine the set of so called “rules” and “”lexicons” were used. Rules were used to automatically identify the words by certain regularities such as “+31” entity for Dutch phone numbers. Lexicons were used to hardcode certain entities that does not have any regularities such as names of university projects.

Lastly all data was transferred to Neo4j environment. During this step Datastic launched Neo4j web application. The Neo4j graph database used as validation of previous step where the quality of outcome of Wowool rules and lexicons were evaluated. Datastic also provides code for the visualisations in Neo4j using the CYPHER query language.

## How can the data from network be stored?

Created database contained more that 1000 nodes and this number will increase with new partners connecting to network. As final application would be distributed as SaaS, the best solution would be to use cloud storage. Companies who provide such storage include Google, Amazon, IBM and so on. Besides that, Neo4j also has its own service – AuraDB, which is apparently well integrated with Neo4j graphs.

## How can the network be presented to the end-user?

Graph database is a way of displaying relations between different entities. However, for end-users it’s essential to have an ability to search the information in graph in intuitive way. In other words, the frontend part of application is needed. Datastic came to conclusion that the best way to distribute network would be creating web-based application. Neo4j provides some solutions to create it. Features of these solutions may not meet all user requirements though. Another possibility would be to create own application, so that could be more flexible solution.

## How can this project be transformed to a business?

The business of Flooid will be a collaborative learning environment in the form of a graph database powered by Neo4j and Natural Language Processing (NLP) technology. A digital part of them can be seen as a digital twin of the real-life network where people interact and exchange information and collaborate with each other.

The product aims primarily on B2B market and targets the following industries:

1. Knowledge institutions

2. Governments

3. Private business partners

4. Social institutions

5. Startup participants

6. Partnerships that unite groups mentioned above

To promote the product among the parties mentioned above several promotional strategies will be used such as Word of Mouth (Marketing), Earned Media and Paid Media, Industry events, and direct selling.

The software will be distributed using the software-As-A-Service (SaaS) model. Using SaaS the application runs on a Neo4j server and users access them using any internet-connected device. The payment model in this case is based on subscription. The subscription model offers flexibility in being able to subscribe per month, quarter or year.

The market research used as a validation of the main idea of the product. The Datastic looked through different aspects of Micro, Meso and Macro environments. First of all, the Microenvironment shows the factors that could influence the performance result of the company. The target setting of the product Flooid wants to establish on the market is a digital twin of the real-life network. Thus, the main market strategy aims at the B2B sector with the focus on the business with the use of sales agents. Secondly, the Meso environment indicates that there are a couple of platforms and applications that offer similar functions. However, the platforms that Flooid wants to establish is new and it contains a lot of functions that are new, thus, it cannot be placed in the same category as other platforms. Taking this into consideration, there is a certain demand on the main idea of ​​collaborative learning among the companies, without any direct competitors to the Flooid’s solution, meaning that Flooid can fit in the market quite easily. Last but not least, the Macro environment consists of factors that are in the play where the business is active. The socio-cultural research has shown that nowadays what regular COVID-19 outbreaks, people are working from home. All meetings, conferences, and communication, in general, take place online. With an online way of working, it is hard to figure out who is involved in a certain project. This situation creates an opportunity for the solution that Flooid has.

More in-depth information about marketing strategy can be found in Business plan document located in the Appendix section of this document.

# Conclusion & recommendations

In this chapter, a conclusion of the research will be made. Furthermore, recommendations will be given on what should be done by the next project group.

## Conclusion

Taking everything into consideration, the project was quite challenging but at the same time quite interesting. The group consisted mostly of non-IT students and because of that a lot of effort has gone into understanding Python and Wowool languages. The group decided to use the Agile approach and divided the project into sprits each of them lasting for one week. With this approach, Drastic researched, implemented, and test new ideas every week and then received feedback on them during the weekly meeting with Flood partners. Thus, the learning went smoothly.

At the beginning of the project, the group created a Project plan where the most important guidelines of the project were described, including the main research question, sub-questions, the scope of the project, methodologies that were used throughout the project, and settled up initial deadlines. The project plan was reviewed several times by the client and by Datastic’s coach in order to provide quality feedback and agree on the terms of the project. When both parties were agreed on provided way of working that was documented in the project plan, it was validated.

Then, Datastic successfully used Web scraping in order to gain the required data. The group quickly got a hang of Cypher queries that neo4j uses because it is similar to SQL queries that Datastic used before. In the end, the “Business plan” document is used as a validation of the main idea of the project. However, the group had some difficulties during the project as well. First of all, is that different websites use different HTML structures, so the script has to be adjusted for some websites individually which requires a lot of additional work. Secondly, NLP was a new technology for the team and had to be studied from scratch. Last but not least, we wanted to make the graph database bigger with all data that we have obtained, however, the larger graph gets the poorly it runs on the device with constant lagging and delay.

In the conclusion, we can say that it is important to listen to the end-user and their needs. The meeting with MindLabs indicates that it is essential to provide an overview of where the company connects with each other and knows what methodologies are in use. Taking this into account drastic found, extracted and included: City, Persons, Universities, Address, Topics, job positions, phone numbers, PO Boxes addresses, university projects, and emails. Using Porto typing methodology, literature study, and workshops the group successfully created a few fully working prototypes in order to test and get more insights into the technology. The created database stores about 1000 nodes and this number increase with new parties connecting o the network. The initial text files from each of the data that was extracted were located locally on the devices of Datastic and updated via GitLab.

## Recommendations

**Web application**

The result of this project is a PoC of graph database. This PoC assumes to be used mostly by developers or as a demo version. For the end users a web application needs to be created. Web application should allow to interact with the database. For instance, a particular company or project could be found in database including the corresponding relations. That could be done either by using Neo4j platforms, or by creating an own application. Another possibility is to use combine both.

Neo4j has different platforms to create visualizations, including Neo4j Bloom and Neo4j Graph Apps. Graph Apps (https://install.graphapp.io) are simple web applications with HTML and JavaScript which interact with Neo4j databases through Neo4j Desktop. Such applications are not complex enough and may not meet the requirements of this project. Bloom is a web interface, which is mostly used to find patterns in data or to investigate it from different perspectives.

Taking in consider the above, our recommendation would be to create own web interface. The possible stack may include Python, Flask, AJAX, React, JavaScript. Examples and tutorials could be found on Neo4j website.

To store the data, we recommend using Neo4j AuraDB cloud storage as it was developed especially for Neo4j graphs. It has free subscription with some limitations as well as paid, starting from 65$ per month. Other cloud storages may include Amazon EC2, Microsoft Azure, Google Cloud Platform.

**Updates on documentation**

During the beginning of project we created “Manual” document with explanation of how to set up the webscraping, docker, GitHub and Neo4j environments. We advise to constantly work on updating “Manual” document so every further R&D team will have a smooth start.

**Data from end user**

Datastic’s researched showed that it is possible to work with Webscraping in order to obtain required data about organizations. However, if organizations contribute to providing data, it will significantly improve the quality of the data set.

**Law and Ethics**

Some of the organizations’ websites prohibit the use of webscraping algorithms. It is also a grey area when the data from open sources used as backup for database. Thus, data is very sensitive matter when it comes to ethical and law considerations. In order to ensure the legitimacy of the project the ethics and law must be researched and discussed.

# Appendix

**List of files:**

* Partners\_Final
* Statistics Flooid
* Webscraped content
* Business plan Flooid