Research report
Vitality platform

DATASTIC
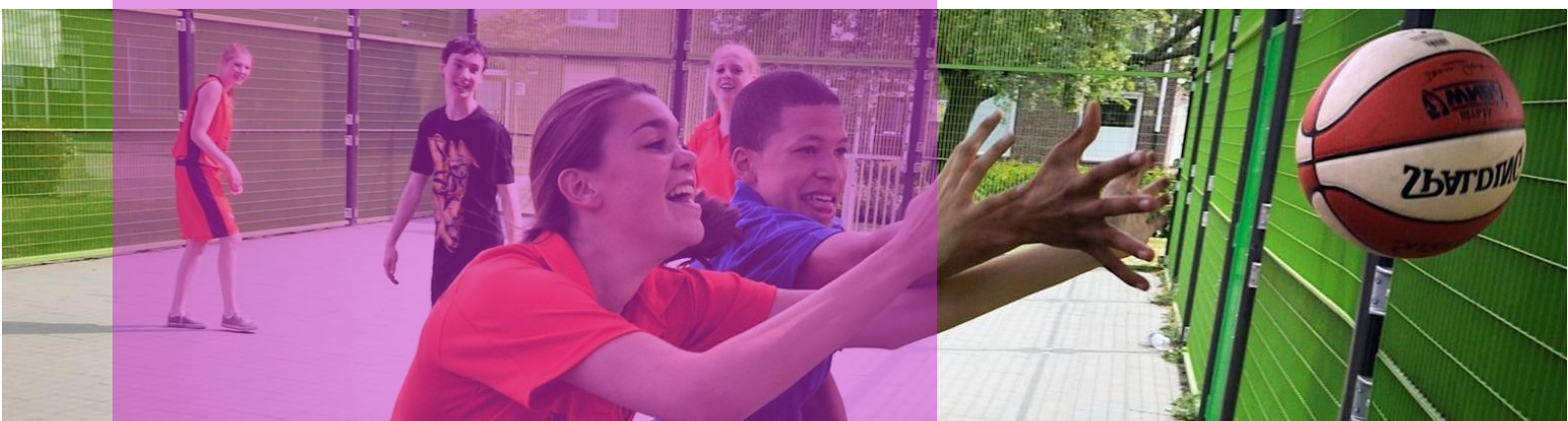
Research report Gemeente Eindhoven

Company	Gemeente Eindhoven
Company supervisor	Michiel Oomen
School	Fontys ICT Hogeschool
	Minor Data Driven Business Lab
Content coach	Jo-Ann Kamp
Process coach	Mr. Rob Verhoeven
Students	Gergana Agorastva
	Remco Bisschops
	Femke Boogerd
	Nikita Gavrilov
	Yaniek Martens
	Lieke Nijs
	Victor Plesciuc
	Ilya Tsakunov
	Aleksandr Vereshchagin

DATASTIC

DOCUMENT HISTORY

REVISIONS

Version	Status	Date	Changes
0.1	Concept	14-09-2021	Created Layout
0.2	Concept	22-09-2021	Added requirements
0.3	Concept	06-12-2021	Added Webscraping
0.4	Concept	15-12-2021	Added Matchmaking
0.5	Concept	17-12-2021	Added introduction
0.6	Concept	20-12-2021	Finished Personalized part
0.7	Concept	23-12-2021	Added Ethical Part
0.8	Concept	13-01-2022	Reviewed whole document and added advice
1.0	Finished	19-01-2022	Added Law Part

APPROVAL

This document has the following approvals:

Version	Date Approval	Name	Function	Paraaf

DISTRIBUTIE

This document is sent to:

Version	Name	Function
0.8	Jo-An Kamp	Lecture
0.8	Michiel Oomen	Client
1.0	Michiel Oomen	Client

DATASTIC

TABLE OF CONTENT

1 INTRODUCTION

Fontys Hogescholen HBO ICT & Business semester 7 consists of a Minor, which is Data Driven Business (DDBL). During the Minor an investigation is conducted at a company with an IT issue. This issue is being explored by a group of students called Datastic. The knowledge that the students have is combined and put into practice during the research. The project will take place from August 30 to January 21.

The client for the project will be the Municipality of Eindhoven. The municipality of Eindhoven wants to create a platform that makes it as easy as possible to find the sport & facility activities that match the citizens' specific preferences.

The municipality hired Bureau Moeilijke Dingen to research the user needs and in what way this can be met. They came up with a concept of an app where the citizens of Eindhoven can easily find information about sporting events. To make this app and integrate this they made a project plan where they work in phases. In phase 1 all the sport offers will be brought in scope and there will be looked at the matchmaking aspect of the app. Also, does phase 1 consist of law and ethic.

In phase 2 there will be looked at the central platform with integrated platforms while also maintaining control of privacy and with transparent AI. The last phase will consist of the development of the digital economy. This project will focus on the first phase.

Currently, the available sports information doesn't meet personal preferences. It takes quite some work to find the sport facilities that match the personal preferences. Research has shown that it is not always clear what the offer of sports in the environment looks like.

The municipality Eindhoven owns the websites eindhovensport.nl and Eindhoven365 owns thisiseindhoven.nl where information is offered related to sports. There are also external websites where information about sports can be found (Eventbrite.com and meetup.nl).
The information is spread over multiple websites which leads to the problem that there are no clear overviews of the events.

The main problem is that this information is not personalized, making the correct information more difficult to find. Personalization should provide even more overview for the best possible way to find sport information. This leads to a less user-friendly website than desired.

The goal of the project is to deliver an advice and improvements based on 3 aspects: Automate supply, Law and Ethics including GDPR and a personalized profile. This advice is supposed to help the Municipality to create an assignment and specifications for future development of the Eindhoven sport web platform to hand it over to third-party development team.

The project consists of a main question and sub questions to reach the project goal. The main question is: How can municipality Eindhoven create a platform with integrated intern and extern data for citizens of Eindhoven to find and fulfil their sport needs with matchmaking? Sub questions were made to answer the main question:

1. What are the end users' requirements for a sport platform?
2. How can sport activities that exist on different websites be combined?
3. What is the most suitable matchmaking technique to show user's sport preferences?

4. How can the chosen matchmaking technique be implemented to display the user's sport preferences?
5. How can the matchmaking be ethically responsible and transparent?

This document has been prepared to capture all relevant basic information and principles of the project so that it can be properly managed. It aims to define the project on the one hand and, on the other hand, to serve as a basis for management and enable assessment of the project's success.

Chapters 2 to 6 describe the process and implementation of the research for each sub-question. Finally, chapter 5 follows in which the conclusion and recommendations are formulated.

2 REQUIREMENTS END USERS (Q1)

Sub question one will be answered in this chapter, which is: What are the end users' requirements for a sport platform? To answer this question Datastic used an approach based on the ICT research methods. The process of the methods can be seen in figure 1.
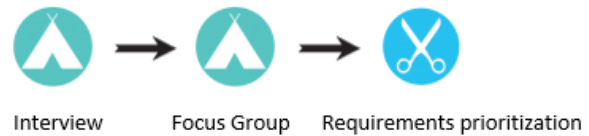
Figure 1 Research methods Question 1

In this chapter, the requirements of the end-users will be discussed. To find out what the requirements are, interviews were held with people in the focus group. With the interviews, a list is made with the requirements which is important for the end-users to have in the app.

2.1 INTERVIEW RESULTS

First, there will be looked into the interviews that were held. In appendix 8.1, the interviews can be found. From these interviews, multiple conclusions can be drawn.

All the interviews and surveys that were held with inhabitants of the municipality Eindhoven. There are both Dutch and internationals that filled in the survey, which gives a view on different nationalities user requirements. All the respondents are currently practicing a kind of sport. From this, it can be concluded that it is possible to find different sports easily enough to participate.

However, most of the respondents stated that they use the www.ssceindhoven.tue.nl website. On this website, all the different kind of sports and activities are shown. This indicates that it is desirable to have one website with an overview of all the possibilities. Next to that, it is mentionable that this website is only for students, so it cannot be used for other inhabitants of Eindhoven.

All respondents said that their main source of information is online, either just on the internet or via Facebook. This affirms the need of an online tool since this is the main source of information.
When looking into the motivation, most respondents stated that they might be interested in a new sport if Eindhoven sports suggested one for them. Therefore, a personalized sports profile is desirable. However, for this profile personal data is necessary and not every respondent is willing to share their personal data.

There has also been looked into the information respondents wanted to see about the sport activities. A short summary of the most mentioned parts is listed:
- Description
- Timetable
- Place
- What do you need to bring?
- Location
- Recent activity
- Age and experience level
- Price

2.2 FOCUS GROUP

The results from the interviews were discussed with a focus group. This was done by asking them what their thoughts were on the conclusions that could be made. The questions and answers can be found in appendix 8.6.

oOo DATASTIC

An important thing both the respondents and the focus group agreed on, was that the main source of information is the internet. Sometimes the first source is different, for example friends or a poster, but the group stated that even then they would look more into it via the internet.

Next to that, everybody agreed that for students it is easier to find information about different sports than non-students because of the www.ssceindhoven.tue.nl website.

When it came to sharing data, there were more different opinions. Not everybody is willing to share the same amount of personal data. This will have an impact on how personalized a personal sport account can be. Besides that, the focus group stated that if you are willing to share more data, that the recommended sport should be more accurate. On top of that, they indicated that seeing recent activities and progress would be interesting. For this, it is however necessary to share more data.

With these options, the focus group stated that they would be motivated to sport more often and to try out new things. The last thing presented to the focus group was the information that the respondents would like to see. They agreed to earlier stated parts. Next to that, it would also be appreciated to see if there is a limit of participants and how many spaces are left.

2.3 REQUIREMENT PRIORITIZATION

In this chapter will be described how Datasctic team applied MoSCow prioritization method tool to sort out the most important aspects for the project from final user perspective. (MoSCoW Prioritization, n.d.)

Before running a MoSCoW analysis, a few things need to happen. First, key stakeholders and the team need to get aligned on objectives and prioritization factors. Then, all participants must agree on which initiatives to prioritize. In this case Datastic had interviews with focus group, personal interviews, and survey.

At this point, also discussed how any disagreements will be settled in prioritization. If it is possible to establish how to resolve disputes before they come up, it can help prevent those disagreements from holding up progress.

DATASTIC

Pros of using this prioritization framework:
- It's good for involving stakeholders without a technical background in the product prioritization process.
- Quick, easy, and intuitive way of communicating priorities to the team and the customers.
- It allows you to think about resource allocation when you classify your features and requirements into each bucket.

Cons of using this prioritization framework:
- It's tempting for teams and stakeholders to overestimate the number of Must-Have features.
- It's an exercise in formulating release criteria more than a prioritization method.
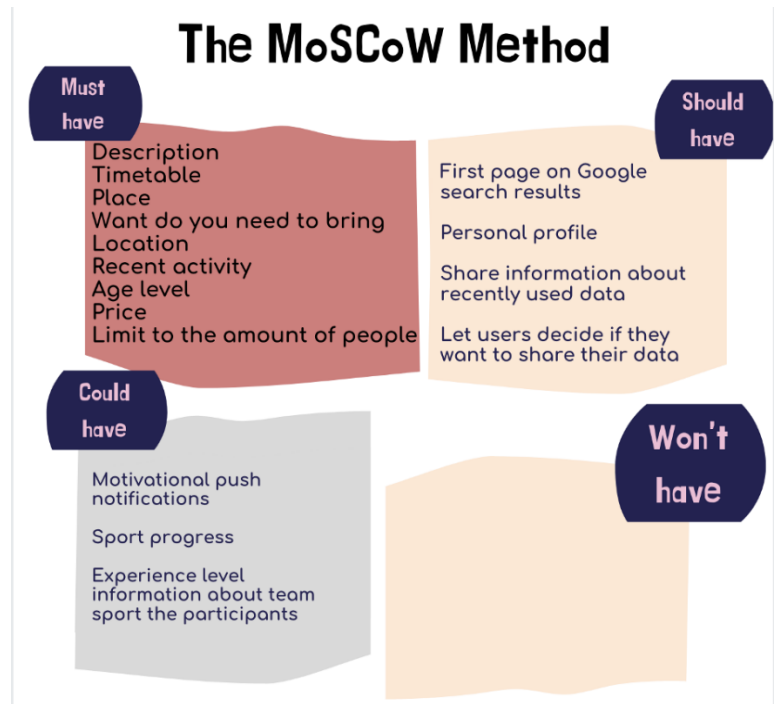
Figure 2 the moscow method

Must-haves are the features that must be present for the product to be functional at all. They're non-negotiable and essential. If one of these requirements or features isn't present, the product cannot be launched, thus making it the most time-sensitive of all the buckets. Should have requirements are important to deliver, but they're not time sensitive. Could-have is a feature that's neither essential nor important to deliver within a timeframe. They're bonuses that would greatly improve customer satisfaction, but don't have a great impact if they're left out. Last of all, won't-have features are the least critical features, tasks, or requirements (and the first to go when there are resource constraints). These are features that will be considered for future releases.

The MoSCoW model is dynamic and allows room for evolving priorities. At the same time, a feature that was considered a "Won't-Have" can one day become a must-have depending on the type of product and user's upcoming needs.

From current research it is possible to identify the most needful information about sport activities that must be available for users from the beginning:
- Description
- Timetable
- Place
- What do you need to bring?
- Location
- Recent activity
- Age level
- Price
- Limit to the amount of people

With this list of requirements, the platform will give users basic information about sport activities that can be extended in future with creating personal profile, where people could control what kind of data they share and if they want to share it. In addition, focus group suggested extra features like

oUO DATASTIC

motivational push notifications, sport progress information and opportunity to check a level of players for team sport activities.

2.4 CONCLUSION

From the interviews, it can be concluded that having one website with an overview of the different sport activities is desirable. The focus group also confirms this. For the use of personal data, it is important to make it clear where the data is used for. Users should be able to use the platform without sharing their personal data if they are not willing to do this. The list of requirements that came from both the interviews and focus group that needs to be available is:

- Description
- Timetable
- Place
- What do you need to bring?
- Location
- Recent activity
- Age level
- Price
- Limit to the amount of people

3 WEB SCRAPING (Q2)

Sub question two will be answered in this chapter, which is: How can sport activities that exist on Eindhoven Sport website be combined? The process of the methods Datastic used to answer the question can be seen in figure 3.

To extract and collect any public information from websites the web scraping was used. To successfully implement the web scraping technique in practice, investigation and theoretical exploration were performed. This chapter describes five DOT framework methods that were used and the result of the investigation.

Figure 3 Research methods question 2

The first investigation started with literature study in web scraping techniques. This is done to give a global view of the existing ways of web scraping. Web scraping techniques can be used for different purposes and in different scenarios. The techniques were investigated based on the requirements of the tool and the knowledge of the team. The following techniques were researched: HTML Parsing, HTTP Programming, Text Pattern Matching, DOM Parser and WebScraper.io. Datastic concluded this research is good to understand the ways of web scraping but is too broad to choose a specific way of scraping. So, new research was started to zoom more into the best tool for web scraping. First, it started with literature study to look for the best languages and platforms.

Datastic did a research on the following languages and platforms: Python, Node.js, Ruby, C++ and PHP. Based of (Prowebscraper, 2021) and (JobsPikr, 18) Python is the best option for web scraping. In python you can use dozens of libraries for web scraping. 5 libraries were investigated to look for the best option. This is based on the research of (elitedatascience, 2020). In this report they decided to feature the 5 Python libraries for web scraping that they found the most interesting. There were several requirements that the library should need:

- Easy to learn
- Able to scrape from multiple websites at once
- Free accessible to everyone

Based on these requirements the library's Scrapy, lxml and request didn't work for the project. They weren't beginner-friendly, supported by difficult documentation, or were not able to do HTML parsing. The most common used library was beautiful soup and based on the advantages, Datastic is going to use this library. For the full research of the techniques and tool see Appendix chapter 9.3 and 9.4.

3.1 BEST GOOD AND BAD PRACTICES

To get more familiar with web scraping before the prototype will be made, some of the best, good and bad practices will be searched. The goal of this part is to get familiar with common ways of working during web scraping. Also, some of the common mistakes are tackled out before Datastic is starting on the prototype. Underneath, a table is created which contains best, good and bad practices which Datastic will use when they start web scraping:

Best practices	Good Practices	Bad practices
1. API: Web scraping would be useless if there is an API available.	1. Don't breach the GDPR	1. Don't hard code session cookies
2. Don't hit the servers too often otherwise you will get blocked.	2. Use captcha solving services	2. Don't DOS Websites
3. Website administrators create sometimes the robot.txt file for scraping information	3. Use headless browsers	3. Don't copy and paste reusable code
4. Don't follow the same scraping pattern	4. Watch out for honeypot traps (bot detection system)	4. Don't write single threaded scrapers
5. Don't violate copyright, always consider the web data if its copyrighted data.		5. Don't use the same pattern of scraping.

Table 1 best good and bad pratices

3.2 SUMMARY EXPERT INTERVIEWS

To get a better understanding of webscraping, Datastic conducted an interview with two experts. These experts are Pepijn Verburg from Bureau Moelijke Dingen and Marco Langhuizen from DDBL.

The outcomes of the interviews are summarised based on different subjects. The interviews can be found in appendix 8.5.

Libraries

The interviews revealed that beautiful soup is a good library to use for webscraping. It is a library with a good comprehensive documentation. Both interviewees also recommend looking into scrapy which is more of a crawling tool than a scraping tool. The whole idea is to scrape HTLM pages that contains elements. Also searching for tags, once one has been identified, search for the content in the tag. There are two ways of searching, CSS matching and Xpath matching. While scrapy was less-functional version of beautiful soup it is easier to use according to mister Langenhuizen.

When asked if Webscraper.io is a good tool to use, both interviewees expressed that it is a good idea to use it. It is useful to get to know what is available on the website. Recommended is to start with a small number of websites and expand it if it is working.

Webscraping

The easiest way to webscrape a lot of websites would be if most of them have the same html structure. Unfortunate, this is not the case for most websites. This way it is hard to use a basic template to use for all the websites. Furthermore, if the HTML structure changes the whole script won't work anymore. This is less of problem if you're using a crawler like Scrapy, where the crawler looks for links instead of specific classes or ids.

The best way to prevent harming website with queries is to slow down the process and checking the cache headers. To store data, it is recommended to use some kind of database. This depends on the project wherefore it is used

Legal

Every website is different. For this project there are mostly websites that are hosted by IT-providers. They might ban the webscraper if it creates a huge load on the server. Furthermore, before webscraping you can also look at the Robots.txt files and html header to figure out if the owners want certain information to be webscraped. In most cases it is legal to webscrape information, it can become an issue if you want to use the information. E.g., for your own website. But for more information about the legal parts of webscraping it is recommended to have an expert legal person look at this.

3.3 PITFALL REPORT

During the investigation on websites, the "Eindhoven sport" portal was used. The website contains 165 sports providers that operate on the territory of Eindhoven. To obtain more information than basic description, contact information, and address that was stored on "Eindhoven sport" each website of the sport providers was thoroughly examined.

During the examination of websites, Datastic defined several features that websites should have. To begin with, the organization needs to be commercial that is why all associations were filtered out. Moreover, the schedule and open activities need to be present and up to date on the website. Finally, the HTML structure needs to be clear and understandable in order to extract data from the website using web scraping.

3.3.1 Website analysis

To have a clear overview on the websites that were examined an excel sheet was created. The excel file has several columns such as Name of the website, Page which was extracted from "Eindhoven sport", URL link of the website. These topics were put in the columns such as Association (yes/no), Activities, Activities up to date, Training schedule, Everything on the same page, Correct HTML structure, Commercial (yes/no), Any open activities, Overall Usability of website.

To mark the websites the binary variable was used such as simple answers "Yes" and "No".
More in depth overview of the excel file will be described in the following chapter.

Name Website	Page	URL	Association (Yes/No)	Activities?	Activities up to date? (Yes/No)
afromoon	1	https://www.afromoonworld.com/	No	No	No
aikido ando	1	https://aikido.school/	No	No	No
airtricks	1	https://www.airtricks.nl/	Yes	No	No
Almonte Basketball	1	https://www.almonte.nl/	Yes	No	No
ana danst	1	https://www.anadanst.nl	No	Yes	Yes
area 51	1	https://www.area51eindhoven.nl/	No	Yes	No
artisport	1	https://www.artisport.nl/	No	No	No
badmintonclub eindhoven	1	https://bceindhoven.nl/	Yes	Yes	No
badmintonclub Smashing Bruang	1	https://www.smashingbruang.com/	Yes	Yes	No
basketbalvereniging achilles '71	1	https://achilles71.nl/en/welkombijachilles71basketk	Yes	No	No
BC Hanevoet	1	https://www.bchanevoet.nl/	Yes	Yes	Yes
BCAB Eindhoven	1	https://www.bcabeindhoven.nl/	Yes	Yes	No
Bewegen Blijft	1	https://www.bewegenblijft.nl/	Yes	No	Yes
Boksclub the golden gloves	1	https://goldengloves.nl/	Yes	No	No
Brabant Diving	1	https://brabantdiving.nl/	No	Yes	No

Figure 4 list of websites from eindhovensport.nl

To get more insights in the websites that were provided by the municipality, Datastic added some python statics to it. From that all the websites that were provided, only 47% contained activities (Figure 3). Since Datastic saw that almost all these activities were for members of an organization, they also looked to which activities are open for everyone. In Figure 3 the number of open activities is shown. This is only 24.7% of the available websites with activities, containing 19 websites. Only 10 of these websites are up to date (Figure 2).
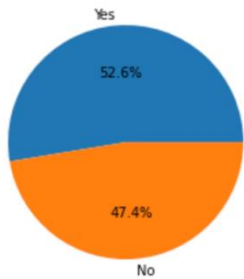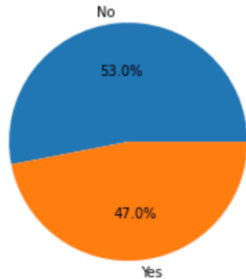
Figure 5 - Activities

Once Datastic went one step further and look if the HTML is scrapable, it had the following outcome, which will mean that only 7 websites out of the 165 are currently useable: When there is a look at the deviation between commercial companies and associations, most of the websites are associations. In Figure 6 is shown that 65.9% of the websites is from associations.

If Datastic dives deeper in the information that is shown per organization type, we can conclude that the percentage that is shown for activities in Figure 9, won't really differs that much.
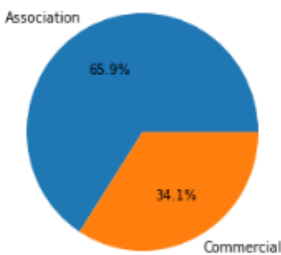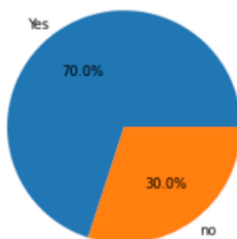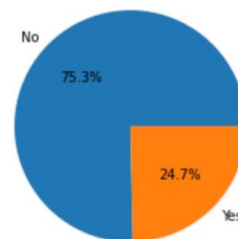
Figure 7 activities per organization type   Figure 8 html structure   Figure 9 open activities

More in-depth information about the open activities per organizational structure is telling that commercial companies overall offer more open activities then the associations. This is visualized in Figure 10.
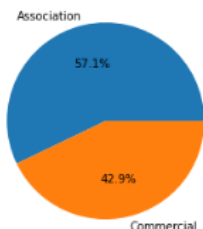
Figure 10 - Open activities per organization type

3.3.2 Webscraping prototype

In this section of the document, Datastic is going to describe web scraping prototype which was based on the Python library BeautifulSoup. Datastic is going to describe the code, explain the outcome and difficulties that company faced.

During the prototype phase of the project, Datastic created several scripts for the web scraping. An example of such script can be found in the Appendix section of the document and was used to web scrape data from "E.S.K.B.V. Impact".

First, Datastic imported libraries into Python to enable BeautifulSoup, connect to websites and create spreadsheets.

```python
from bs4 import BeautifulSoup, Comment
import requests
import pandas as pd
import re
import openpyxl
```
Figure 11 - Libraries

Then Datastic connected to a website that wanted to web scrape and specified the class from the HTML code which is called "classes_slide_wrap". After that Datastic specified the file where the results will be stored, in this case, it is an Excel file called "results".

```python
url = 'https://www.eskbvimpact.nl/'            # URL of main page
r = requests.get(url + '#classes')        # Link to page with events
soup = BeautifulSoup(r.content, "html.parser")
event_info = soup.find_all('div', attrs={'class': 'classes_slide_wrap'})     # Parsing all <a> tags with links from page
file = "C:\\Users\\nikit\\Desktop\\Minor\\results.xlsx"            # File for results
```
*Figure 12 - Excel definition*

In the following script, Datastic excludes all the activities that are intended for members of the club. After that, titles and descriptions are extracted from the website using classes "classes_title" and "classes_text".

After that, Datastic extract information about the day, time, and location. In order to do so we extract everything that stands after the words Day, Time, and Location and symbol ":".

```python
forbidden_level = '(Members only)'
for element in event_info:
    #if forbidden_level not in element.text:
    titles = [re.sub(r"[\n][\W]+[^\w]", "\n", e.text.strip()) for e in soup.find_all('div', attrs={'class': 'classes_title'})]      # Extracting titles
    descriptions = [re.sub(r"[\n][\W]+[^\w]", "\n", e.text.strip()) for e in soup.find_all('div', attrs={'class': 'classes_text'})]       # Extracting descriptions
    days = []
    times = []
    locations = []
    for description in descriptions:
        description_parts = description.split("\n")
        for part in description_parts:
            if 'Day' in part:
                days += [part[part.index(':') + 2:]]
            elif 'Time' in part:
                times += [part[part.index(':') + 2:]]
            elif 'Location' in part:
                locations += [part[part.index(':') + 2:]]
```
Figure 13- Extraction

Finally, Datastic extracts the outcome of the web scraping into an Excel file. In order to do so the pandas' library comes to use. Datastic specify the columns that are going to be present in the spreadsheet and populate the Excel file at the end.

```
df = pd.DataFrame(list(zip(titles, descriptions, days, times, locations)), columns=['Title', 'Description', 'Day', 'Time', 'Locations'])
df = df[~df['Description'].str.contains(forbidden_level)]        # Filtering for open events only
df = df.drop('Description', 1)
df.to_excel(file)
```

Figure 14 - Write to excel

Down below you can see the outcome of the script. The outcome is stored in the Excel file in a form of a spreadsheet. There result is split between four columns: Title, Day, Time, and Locations which contain relative information.

resultsboxing

	Title	Day	Time	Locations
2	Wednesday Kickboxing &	Wednesday		Sporthal 3A
4	Friday Introduction t	Friday	21:30-23:00	Dojo

Figure 15 - Result

3.4 CONCLUSION AND ADVICE

While Datastic performed some library research they came out to the conclusion that beautiful soup was the most common library to use while web scraping. The reason why Datastic also had chosen to use this was the following:
- It's easy to learn
- You are able to scrape from multiple websites at once
- Its free accessible for everyone.

The best, good & bad practices had some of the same outcomes. Mainly the most common outcome was preventing to getting blocked by a website. This could be handled by using a timer. Also, the experts confirmed that using beautiful soup was a good starting point. They said that web scraping is pretty hard to learn and because of all the different HTML structures, it's hard to manage. The legal part of web scraping is still unknown, also for them. That's why it is important to look at the outcome of the 5th sub question.

During the research of every website Datastic concluded that it is hard to find websites that are containing open activities to scrape. Even when 157 of the websites weren't usable, Datastic started prototyping a script that was able to scrape from some websites.

The above prototype shows that it is possible to web scrape data from the website. If the content is present on the website, it is always possible to extract it using the classes of HTML where such objects are stored. However, there are some disadvantages of such a method.

First, the classes where the data is stored need to be identified manually, meaning that you first need to go to the website and then inspect the webpage and identify the class which leads to another major issue. Currently, it is not possible to fully automate web scraping techniques to extract data from websites. Even if Datastic assume that all websites contain required information the HTML structure is different for each website meaning that the script needs to be adjusted.

Finally, maintenance of such a script is hard and time-consuming. If we assume that organization would like to update the website and add another open activity the script will not track it down because it

will not store information about newly added classes. Thus, the script needs to be updated manually and regularly.

Taking everything into consideration, it is possible to web scrape data from the website, but it is hard to make it automated for many websites and a lot of manual work is still required.

Datastic advises the municipality to think about the consideration of using another method then web scraping. Creating an automated web scraping script that can scrape from different websites containing different HTML elements is a very tough project. This means it will cost the municipality a lot of money. Even when someone can create this kind of script, there is no content available to scrape, worth the costs and time.

Datastic is advising one of the following two options:
- Create your own websites where associations or commercial companies can put their activity into a form. This means every activity will be putted into the Municipality's site in the same structure, so it is easy usable for their platform (this will reduce the costs).
- Contact the organizations about your plans and perform research if they are willing to update their websites with the content the municipality likes to scrape.

4 MATCHMAKING (Q3)

Sub question three will be answered in this chapter, which is: What is the most suitable matchmaking technique to show user's sport preferences? The process of the methods Datastic used to answer the question can be seen in the figure 13.

Figure 16 Research methods question 3

4.1 LITERATURE STUDY MACHINE LEARNING

The reason to dive into Machine Learning is that client would like to see sport recommendations for user based on personal preferences and previous user's sport activities. To get a better understanding of this topic, Datastic conducted an interview with machine learning expert/Fontys teacher form DDBL minor Simona Orzan. The main idea of this interview was to check what type of library the Datastic team could use for Municipality project to create personal matchmaking process for vitality platform. The outcome from this meeting was an expert advice to use clustering approach: K means clustering.

4.1.1 Clustering approach overview

Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. In our case, sport activities can be divided by groups based on type of activity (Fighting and Defence sport or team sports with balls) (Clustering, n.d.)

4.1.2 What Is K Means Algorithm

K Means Algorithm is an Iterative algorithm that divides a group of n datasets into k subgroups /clusters based on the similarity and their mean distance from the centroid of that subgroup/ formed. K, here is the pre-defined number of clusters to be formed by the Algorithm. If K=3, It means the number of clusters to be formed from the dataset is 3 (K-Means Clustering in Python: A Practical Guide, n.d.).

4.1.3 Algorithm steps Of K Means

The working of the K-Means algorithm is explained in the below steps (Step by Step KMeans Explained in Detail, n.d.):

Step-1: Select the value of K, to decide the number of clusters to be formed.
Step-2: Select random K points which will act as centroids.
Step-3: Assign each data point, based on their distance from the randomly selected points (Centroid) to the nearest/closest centroid which will form the predefined clusters.
Step-4: Place a new centroid of each cluster.
Step-5: Repeat step no.3, which reassign each datapoint to the new closest centroid of each cluster.
Step-6: If any reassignment occurs, then go to step-4 else go to Step 7.
Step-7: Finish
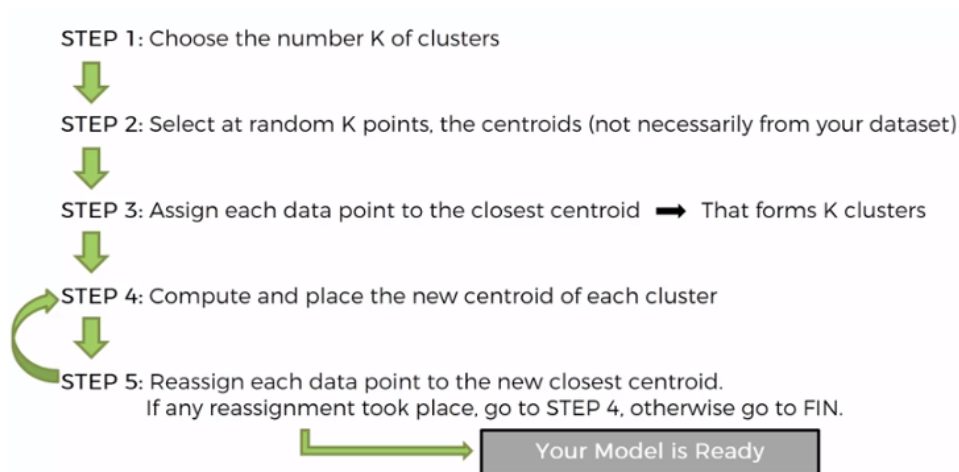
Figure 17 steps K-Means Algorithm
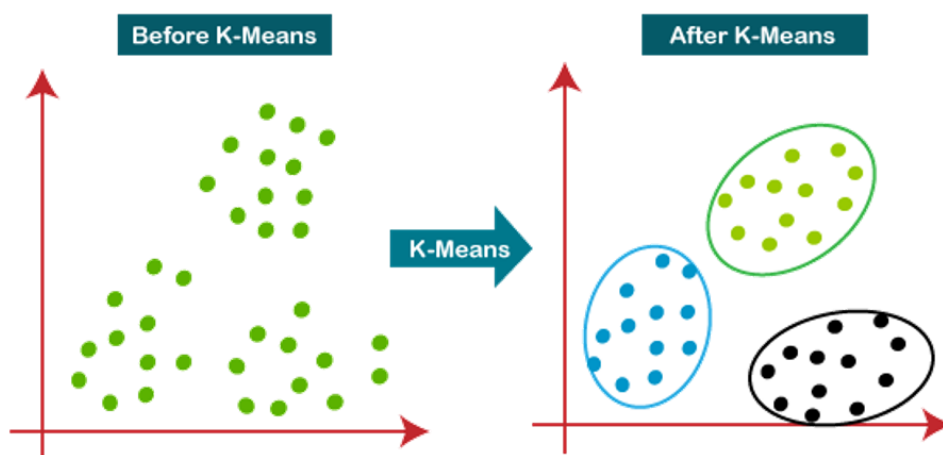
Figure 18 before and after K-Means

4.1.4   Decision tree

The decision tree is the most powerful and popular tool for classification and prediction (Geeks for Geeks, 2021). A decision tree is a flowchart like a tree structure, where each internal node designates a test on an attribute, each branch represents a result of the test, and each leaf node contains a class label.

Figure 19 Example decision tree

The construction of the decision tree classifier does not require any domain knowledge or parameterization and is therefore suitable for exploratory knowledge discovery (Geeks for Geeks, 2021). Decision trees can handle large data. In general, the decision tree classifier has good precision. Decision tree induction is a typical inductive approach to learning classification knowledge.

Strengths and Weakness of Decision Tree approach

The strengths of decision tree methods are:

- Decision trees can generate understandable rules.
- Decision trees perform classification without requiring much computation.

- Decision trees can handle both continuous and categorical variables.
- Decision trees provide a clear indication of which fields are most important for prediction or classification.

The weaknesses of decision tree methods:
- Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Decision trees are prone to errors in classification problems with many classes and relatively small number of training examples.
- Decision tree can be computationally expensive to train.

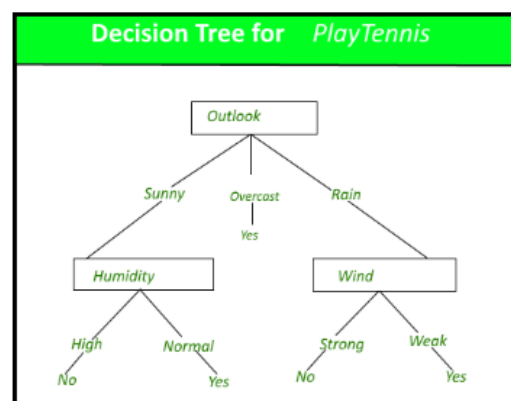## 4.2 LITERATURE STUDY RULE ENGINE
In this chapter literature study will be done to get more information about rule engine. This method can be used in the future as it comes to matchmaking.

4.2.1 What is a rule engine?
A rule engine is a piece of software, that runs a set of rules on data and if any condition matches then it executes the corresponding actions (F.Marchioni, 2012). These rules are described in a declarative manner and are often expressed as, "When some conditions are evaluated to true, then do some tasks ".

In a system there can be multiple rules that can lead to multiple rules being true at the same time. Then are the rules in conflict. To prevent this can a rule engine use different conflict resolver strategies. For this reason, are there two execution methods that can be used.

Forward chaining: a "data-driven" method.
When a fact is inserted or updated will the rule engine use available facts and exciting rules to extract more facts until a goal is reached. The rule engine will start with facts and end with a conclusion.

Backward chaining: a "goal-driven" method.
The reversed method from forward chaining. This method starts with a conclusion or a list of goals. When these goals cannot be reached than it will search for sub-goals that can satisfy some part of the main goal. This process will go on until the conclusion is proven or there are no more sub-goals.

Advantages of rule-engine (Katiyar, 2019)
- Rules are easy to make, also for non-technical persons.
- Rules are stored at a centre storage.
- Logic is managed separately from core application logic so it can be managed and reused.
- In rule-engine, we use different pattern matching and conflict resolving algorithms, which give high performance.
- Rules can easily be updated. No code changes are required.
- The different applications can use the same rule-engine for the same logic. It increases reusability.

When not to use rule engine: (Study tonight, sd)
- Rule engine is an overkill for small project with less than 20 rules.
- If the business logic is static.
- If the rules are simple or self-contained, when it is only for a single object.

- If the project where it is used for is a one-time thing.

4.2.2   Conclusion

This research has shown that rule engine a good method is to use for this project. It is easy to use and understand for all project members. For now, the rule engine can be a little overkill but with the future in mind where more kinds of events will be added is a rule engine a good choice to use for this project. Furthermore, the decision tree is a good match with the rule engine.

4.3   BEST, GOOD AND BAD PRACTICES

The purpose of this research is to outline the best, good, and bad practices of Rules Engine which helps in creating optimized, less error prone and easy to maintain rules. This research is based on the following sources: (Gainsight, 2020), (Berlioz, 2019), (Fuchs, 2014).

Table 2 Best Good Bad Practices Rule Engine

Best practices	Good practices	Bad practices
Use filters: for example, wherever possible try to process only incremental data based on last modified date and add other appropriate filters.	**Update in one go:** When multiple rules are configured to update/Load data to an object, try to point those rules to an object and use a single rule to load data from the object to another object.	**Avoid Multiple Datasets transformations and merges:** for example, if multiple objects can be joined using lookups in MDA/SFDC then use lookups to fetch data using single data set instead of having multiple datasets and merges.
Use Built-in functions: Use in built function and make sure to add appropriate date filters to the source dataset.	**Include and exclude operators:** If multiple filter condition on the same field must be added to the rule, then use Includes/Excludes option instead of multiple OR conditions.	**Do not use OR and ELSE:** OR makes it difficult to track the business performance of each scenario. the best design will use default actions if they exist in your system. A default action is typically defined at the rule set level and applies when none of the rule's fire.
Duplicates in the target data: Make sure Identifier fields will never receive NULL values from the source and Source should not have duplicates based on identifier fields. Insert should be avoided in the rules. Make sure to use same set of identifiers while inserting/updating data across different rules/Actions.	**Numbering rules:** Use a numbering scheme to make it easy to remember the order when constructing the rule chain.	**Parallel Execution:** If there is no dependency between rules then go with individual schedules instead of rule chain. Scheduling rules in rule chain will prevent rules from running parallel.
Brainstorm: Brainstorm extensively about the variables	**Color Code:** Color code the variables and solutions so that	**Rules with same name:** avoid creating multiple rules with same name.

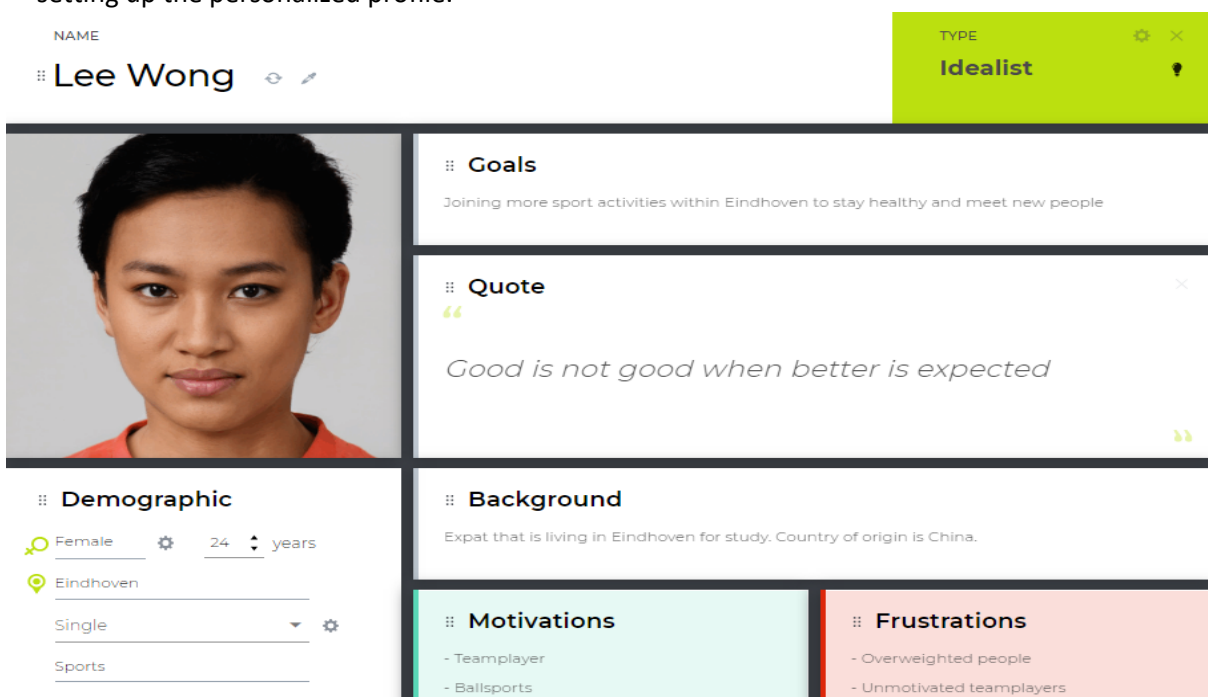|		
and common problems customers may encounter. Consider all solutions that customers may attempt while troubleshooting.	the decision tree is easily understandable.	

4.4 DATA ANALYSIS

In order to create a prototype to test Machine Learning or Rule engine outcome the Mockup data was created. The data set consists of six columns: Company Name, sport category, Title, Date and Time of open activity and last but not least Location of sport facilitator. The data is completely random. However, Sport Category was written based on sport categories from eindhovensport.nl. Title column rows consist of name of sport provider and sport category plus "Open activates".

Sport Provider	Sport Category	Title	Date	Start Time	Location
Auer-Schulist	Custom Sports	Auer-Schulist Custom Sports Open Activity	01.10.2021	20:00:00	58674 Pawling Junction
Murphy-Treutel	Sport and Game	Murphy-Treutel Sport and Game Open Activity	25.09.2021	15:00:00	788 Dapin Circle
Gerlach Group	Agility Sport	Gerlach Group Agility Sport Open Activity	30.11.2021	18:00:00	260 Onsgard Road
Cole, O'Connell and Hauck	Gymnastics	Cole, O'Connell and Hauck Gymnastics Open Activity	27.09.2021	21:00:00	35208 Northport Lane
Leuschke-Ledner	Dancing	Leuschke-Ledner Dancing Open Activity	21.10.2021	13:00:00	4 Corben Junction
Moore Inc	Gymnastics	Moore Inc Gymnastics Open Activity	21.11.2021	16:00:00	040 Graedel Trail
Farrell, Hegmann and Stoltenberg	Cardio and Strength training	Farrell, Hegmann and Stoltenberg Cardio and Strength training Open Activity	10.09.2021	15:00:00	5 East Terrace
Douglas-Lindgren	Winter Sports	Douglas-Lindgren Winter Sports Open Activity	28.09.2021	16:00:00	381 Glendale Center
Gorczany-Boehm	Gymnastics	Gorczany-Boehm Gymnastics Open Activity	25.12.2021	17:00:00	2158 Surrey Alley
Rippin Inc	Dancing	Rippin Inc Dancing Open Activity	27.10.2021	21:00:00	2 Springview Trail
Lemke, Towne and Simonis	Mind Sport	Lemke, Towne and Simonis Mind Sport Open Activity	24.09.2021	18:00:00	1 Talisman Parkway
D'Amore-Halvorson	Urban Sports	D'Amore-Halvorson Urban Sports Open Activity	01.09.2021	22:00:00	44 Schmedeman Road
Tromp, Nader and Upton	Agility Sport	Tromp, Nader and Upton Agility Sport Open Activity	27.09.2021	13:00:00	561 Evergreen Park
Welch-Zieme	Mind Sport	Welch-Zieme Mind Sport Open Activity	29.10.2021	14:00:00	22835 Riverside Point
Gaylord, Nitzsche and Harris	Climbing and Mountaineering	Gaylord, Nitzsche and Harris Climbing and Mountaineering Open Activity	12.12.2021	17:00:00	37 Kim Court
Hilpert Inc	Agility Sport	Hilpert Inc Agility Sport Open Activity	24.09.2021	14:00:00	207 Comanche Plaza
McDermott-Kirlin	Mind Sport	McDermott-Kirlin Mind Sport Open Activity	20.11.2021	14:00:00	50 Mosinee Way
Keebler LLC	Ball sport	Keebler LLC Ball sport Open Activity	19.01.2022	19:00:00	891 Del Sol Parkway
Raynor-Stehr	Sport and Game	Raynor-Stehr Sport and Game Open Activity	09.12.2021	20:00:00	6269 Sauthoff Avenue
Considine-Feil	Climbing and Mountaineering	Considine-Feil Climbing and Mountaineering Open Activity	21.12.2021	21:00:00	5 Bobwhite Junction
Kub, Larkin and Mante	Recreational sports	Kub, Larkin and Mante Recreational sports Open Activity	18.12.2021	17:00:00	53873 Burrows Parkway

Figure 20 information different websites

4.5 PERSONA

Datastic designed a persona to create reliable and realistic representations of the key audience segments for reference. Underneath the picture of the persona that will be used to test the matchmaking process, is shown. Based on this persona the questions will be answered regarding setting up the personalized profile.

NAME

Lee Wong

TYPE

Idealist

Goals

Joining more sport activities within Eindhoven to stay healthy and meet new people

Quote

Good is not good when better is expected

Demographic

Female   24 years

Eindhoven

Single

Sports

Background

Expat that is living in Eindhoven for study. Country of origin is China.

Motivations

- Teamplayer
- Ballsports

Frustrations

- Overweighted people
- Unmotivated teamplayers

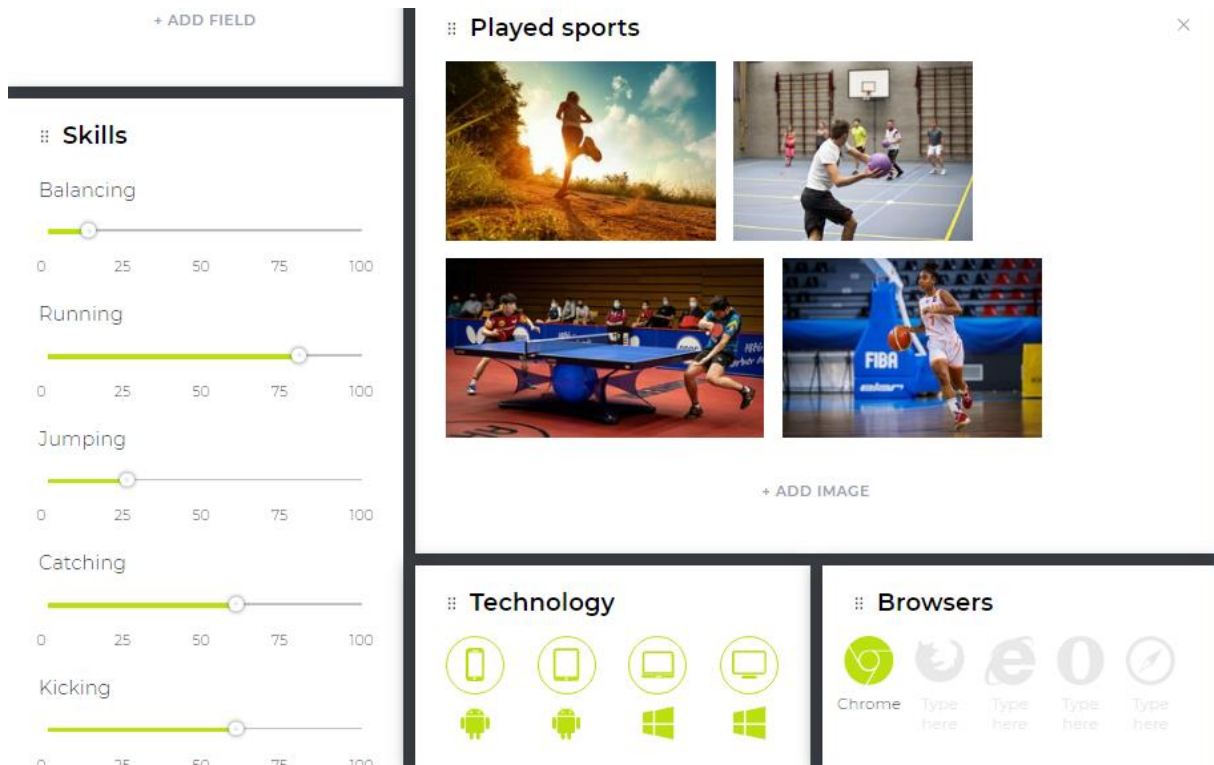*Figure 21 example persona*

DATASTIC

Figure 22 example persona

4.6 CONCLUSION

During the different research approaches in this subchapter, Datastic found what kind of matchmaking tool they will use for the platform of vitality. Firstly, machine learning was further investigated. During the investigation the machine learning technique "Clustering" came as best option. One of those models that is using clustering was K-means. After a bit more in-depth research and some conversations with an expert, Datastic concluded that the data output from research question 2 (webscraping) was too less to use with the clustering model.

Secondly, the rule engine was investigated. Rule engine is an easy way to get to an end-goal. This can be done by creating a decision tree. Here you define the steps that should be followed and that are reaching the result. Datastic has chosen for this method because it's clear and easy to use for every team member.

To start using the rule engine, there are a couple of things to avoid. By researching the bad practices Datastic will not do the following things while creating it:

- Use multiple datasets

Some things Datastic will do (good practices) are:

- Using filters
- Using Built-in functions

At the end of this research, Datastic created a persona that will be used in the future prototype. This persona is created based on the scope of the project (expats).

oᑐ DATASTIC

5   PERSONALIZED DATA (Q4)

Sub question four will be answered in this chapter, which is: How can the chosen matchmaking technique be implemented to display the user's sport preferences?

In this chapter, the prototype for matchmaking will be discussed. First, the python script that is made will be explained. Then future steps that need to be made will be discussed.

Figure 23 research methods personalized data

5.1   PROTOTYPE

In this section of the research document, Datastic is going to describe Rule engine prototype which was based on the Python code and Pandas library. Datastic is going to describe the code, explain the outcome and difficulties that company faced.

During the prototype phase of the project, Datastic created several scripts for the Rule engine. First, Datastic imported Panda library into Python. Pandas - pandas is a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language.

```
1    #import libraries
2    import pandas as pd
```

Figure 24 importing Pandas

In our case, Pandas are used to scan Excel files to import sport data base with information about sport provider, sport category, title, date, start time and location. (MOCK_DATA)

Figure 25 list of all available websites

To make user's preferences (input) matching with the information from database, an additional Excel file with sport characteristics was created. Those characteristics can be found in the table below. (RULES_MATCHMAKING)

Figure 26 list sport characteristics

To import those files, the following script is used:

```
5    #import file
6    data = pd.read_excel(r'/Users/alexandrvereshchagin/Desktop/Minor/Eindhoven/rule/MOCK_DATA (1).xlsx')
7    match = pd.read_excel(r'/Users/alexandrvereshchagin/Desktop/Minor/Eindhoven/rule/RULES_MATCHMAKING.xlsx')
```

Figure 27 Import of sport provider's database and sport activity characteristics

In the following script, Datastic implemented questions like "*Do you like outdoor sport?*" and separation of sport activity into different categories: Outdoor, Indoor, Team, Individual, Ball, Swim, Bike.

Also, in this part of the code if-else statement is used to math and show information based on user's input. If the answer to the given question is "Yes", the script matches sport activity that are located outdoor and give this category one point score, that will be used to filter the best match. If the answer is "No", no point to this category is given.

```
12   #Questions
13       #Outdoor
14   def OutdoorCategory():
15       a = input("Do you like outdoor sports?")
16       if 'yes' in a.lower():
17           match['Location'] = match['Location'].replace(['Outdoor'], 1)
18           print("Outdoor")
19       else:
20           match['Location'] = match['Location'].replace(['Outdoor'], 0)
```

Figure 28 Example of questions and scoring system

The same script structure is provided for Outdoor, Indoor, Team, Individual, Ball, Swim, Bike categories with following questions:

- Do you like outdoor sports?
- Do you like indoor sports?
- Do you like team sports?
- Do you like individual sport?
- Do you like ball sports?
- Do you like swimming?
- Do you like cycling?

oUO DATASTIC

```
83        #input
84        OutdoorCategory()
85        IndoorCategory()
86        TeamCategory()
87        IndiCategory()
88        BallCategory()
89        SwimCategory()
90        BikeCategory()
```

Figure 30 input script for question categories

```
Do you like outdoor sports?yes
Outdoor
Do you like indoor sports?no
Do you like teamsports?yes
Teamsport
Do you like individual sport?no
Do you like ball sports?yes
Ball
Do you like swimming?no
Do you like cycling?no
```

Figure 29 example user's input

To test if the script identifies "Yes/No" input correctly, positive answers are shown after the question. It can be cleaned by deleting the last print() command in "Questions" section of the code.

Table 3 presents the input score result, where "1" shows a match of the category or sport type with user's input and calculate a sum of scores for a specific sport in a purpose of filtering from the highest matching score to the lowest.

Sport	Sport_type	Location	Ball	Swim	Bike	Score
Agility Sport	0	0	0	0	0	0
Athletics	0	0	0	0	0	0
Ball sport	1	1	1	0	0	3
Cardio and Strength training	0	0	0	0	0	0
Climbing and Mountaineering	0	0	0	0	0	0
Combat and Defense Sports	0	0	0	0	0	0
Custom Sports	0	0	0	0	0	0
Cycling	1	1	0	0	0	2
Dancing	0	0	0	0	0	0
Fresbee	1	1	0	0	0	2
Gymnastics	0	0	0	0	0	0
Mind Sport	1	0	0	0	0	1
Racket and net sports	0	0	1	0	0	1
Recreational sports	0	0	0	0	0	0
Scouting	1	1	0	0	0	2
Swimming and Water sports	1	0	0	0	0	1
Urban Sports	0	1	0	0	0	1
Winter Sports	0	1	0	0	0	1

Table 3 input score results

This script matching sport type and categories with "Yes" input to create and display Table 3.

```
92        match["Score"] = match.sum(axis=1)
93        print(match)
```

Figure 31 script matching sport and category based on score

To filter and display the final input result the following script is applied. First it identifies the highest sum of scores and match it with type of sport.

```
95        #Filter data
96        score = match["Score"]
97        a = score.max()
98        print("Score:")
99        print(a)
100
101       result_match = match[match['Score']==a]
102       print("Sports:")
103       print(result_match)
```

Figure 32 script filtering data

oOo DATASTIC

Second, type of sport with highest score matching results are matched with the main list of sport activities to display a list of sport providers with their main information.

```
105    #filter database to input
106    s = result_match['Sport']
107    fsport = s.values.tolist()
108    print("Sports:")
109    print(*fsport, sep=", ")
110    data_new = data[data['SportCategory'].isin(fsport)]
111    print("Sport Providers:", data_new)
```

Figure 33 script matching user's preferences with sport provider database

Example of preferred sport activities based on rule engine:

```
Sport Providers:                    SportProvider  ...                Location
31                  Keebler LLC  ...       891 Del Sol Parkway
32                 Jacobi-Borer  ...       42 Mockingbird Pass
33             McKenzie-Rosenbaum  ...       63932 Lake View Hill
34                Veum and Sons  ...         78 Moulton Park
35              Leannon-Kautzer  ...         87 Reinke Trail
36           Metz, Price and Swift  ...          7 Duke Avenue
37                 Will-Krajcik  ...   06545 Charing Cross Pass
38                  Morar Group  ...       85 Sachtjen Parkway
39    Gislason, Bernier and Hudson  ...    77219 Morningstar Park
40                  Labadie LLC  ...      06 Northport Parkway
41     Legros, Glover and Johnston  ...        8573 Oriole Pass
42              Hagenes and Sons  ...         393 Fulton Plaza
43                    Terry Inc  ...      1502 Elmside Terrace
```

Figure 34 Example of final list of preferred sport activities

5.2 FUTURE STEPS

The results from prototype showed positive result towards the actual working solution. Thus, we can explore what are the next steps in terms of technical background of the matchmaking can be implemented in future.

5.2.1 Data storage

Storing the data about the users' preferences and interests in the sports would be a good idea in this particular case. First of all, user do not have to answer all questions once again since the application is already storing the data about user's interests in the data warehouse. Secondly, the stored data can be used later on to compare the user answers with similar answers from the other users', this technique is called collaborative filtering and it will be described in the following chapter.

5.2.2 Collaborative filtering

Collaborative Filtering is the most common technique used when it comes to building intelligent recommendation system that can learn to give better recommendations as more information about users is collected. Most websites like Amazon, YouTube, and Netflix use collaborative filtering as a part of their sophisticated recommendation systems.

Collaborative filtering uses similarities between users and items simultaneously to provide recommendations. Thus, collaborative filtering models can recommend an item to user A based on the interests of a similar user B. Furthermore, the embeddings can be learned automatically, without relying on hand-engineering of features. Last but not least, the good thing about collaborative filtering in this particular case is since it looks at the interests of other users it matches people with the same interest and experience together.

Example:
Considering a sport recommendation system in which the training data consists of a feedback matrix in which:

- Each row represents a user

- Each column represents a sport

The feedback about sports falls into one of two categories:

- Explicit – users specify how much they liked a particular sport by providing an answer to question in matchmaking.
- Implicit – if a user joins the sport, the system infers that the user is interested

Since the results is based on the answers of the users, we will use binary matrix, where value 1 indicate positive answer to a question and thus interest in sport.

When a user receives a result of matchmaking, the system should recommend sports based on both categories:

- Similarities to a sports that user has done in the past
- Sports that receive the most points during the matchmaking from similar users

2D Embedded

Suppose we assign to each sport a scalar in [0,1] that describes whether the sport is individual (1) or teamsport (0) as well as outdoor(0) or indoor(1). Suppose we also assign a scalar to each user in [0,1] that describes the user's interest in individual (closer to 1) or teamsports (closer to 0) as well as to outdoor's or indoor's categories. The product of the sport embedding and the user embedding should be matched between each other for sports that we expect the user to like. (Collaborative Filtering , 2021) (Build a Recommendation Engine With Collaborative Filtering, n.d.)
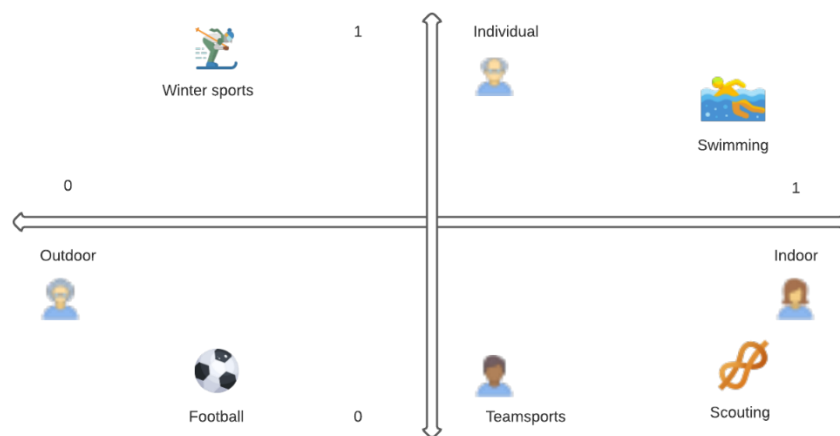
Figure 35 2D embedded conclusion

5.3 CONCLUSION

During the research of this sub question, Datastic made a prototype for how a rule engine could look like. While an expanded version of this prototype could be used for the first phase of the app, is the use of a rule engine not future proof. With this method, no recommendations can be given based on the sports that a user has chosen to perform previous. Also, no recommendations can be made based on other users' similar experiences.

To resolve this is it advised that when municipality Eindhoven goes to the next phase and start looking at more advanced technologies, to start storing data that is necessary for machine learning. This will improve the recommendation system in the app. Because with this stored data the collaborative filtering method can be used.

6 ETHICAL RESPONSIBILITY (Q5)

In chapter six, sub question five, *'How can the matchmaking be ethically responsible and transparent?'* will be answered. To answer this, Datastic used the ICT research approach which is shown in Figure 36.
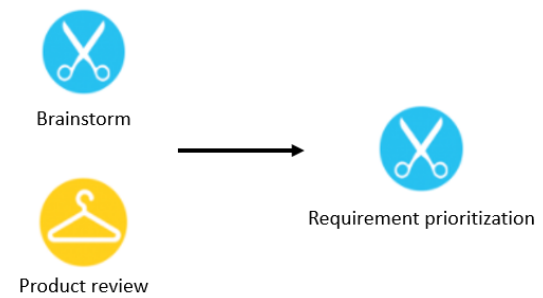
The ethical responsibility from the vitality platform will be discussed. This is done with help of a counselling ethics discussion with use of a brainstorm. Another way the ethical responsibility will be discussed is with help of the TICT-tool, where the product, in this case the vitality platform, will be reviewed.

Figure 36 research methods ethical

6.1 COUNSELLING ETHICS

Counselling ethics is ethics that focusses on the questions such as what the position of humans in this environment is, what is the current situation regarding data and privacy, what are the algorithms doing and that do the end-users want. Counselling ethics is based on technology philosophy.

During this session, participants were involved in a dialog about the effects of new technology and the values that came to question. Different stakeholders participated, leading to different views. Together, values that are important for the digital domain were defined and guaranteed. In appendix 8.7, the full counselling ethics workshop can be found.

6.1.1 Context and reason case

Ethics will be in general more and more important. Also in Eindhoven, the municipality thinks that ethics is important. Recently the local council has established an ethical assessment framework because over the time a lot more ethical cases will be coming. With that coming, it is getting more and more important that things will be done with a certain responsibility. People are asking themselves how they will come to an ethical discussion within a production environment. At this point the approach of counseling ethics will take part. Humans are making technique and technique makes and changes the human. There is a continues cycle between human and technique, which is now more than ever. The approach of counseling ethics is a dialog and way of working that's connected to modern ethics but fulfills a more prominent position next to classical ethics. Underneath in Table 4 the differences/comparisons between classical and modern ethics are shown:

Classical Ethics	Modern Ethics
Focus on technology	Focus on technology and environment
Judge: Is it allowed?	Accompany: How something can go well
Top-down (Frames)	Bottom-up (practical)
Experts	Parties concerned

Table 4 differences classical and modern ethics

oOo DATASTIC

6.1.2 Approach

During the implementation of the Counceling ethics, an approach is used in which phases are used.

The counceling ethics approach is divided in a total of 3 phases. An overview of this can be seen in the figure.
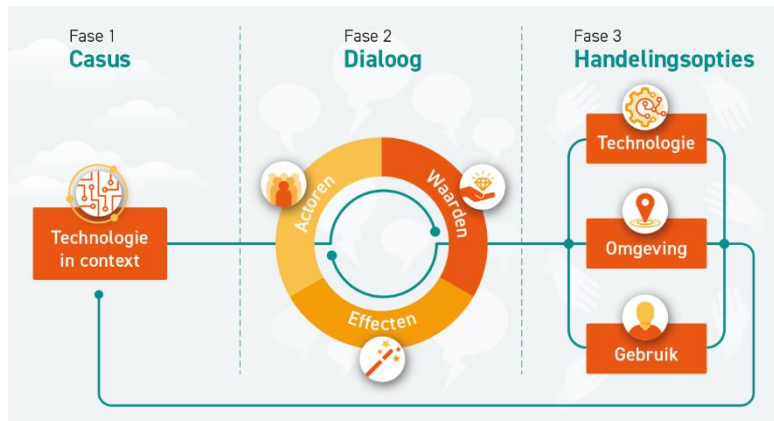
Figure 37 Approach Ethics

Table 5 Approach

Phase 1	Phase 2	Phase 3
Explanation of what the technology looks like and in what context is it applied	1. The actors are defined with a short session/discussion with the group. 2. The group will discuss the possible negative and positive effects. 3. The effects are assigned to a value.	In subgroups, the participants look for action options based on technology, the environment, and the individual.

6.1.3 Phase 1

With the platform for vitality the municipality aims to an integral solution which is a central platform for vitality. The platform needs to offer users suggested sport activities within the city of Eindhoven based on their matchmade profile. To gather the information that must be offered within the platform, the municipality can make use of a couple of modern techniques like:

- Read sources by API-connections
- Web-crawling and scraping websites
- Curation and manual classification by the municipality themselves

For analysis from user perspective, an explicit and implicit user profile can be aligned:

- Explicit: Build a profile by engaging
- Implicit: Share profile details for improvement suggestions (e.g., ask users to share their preferences with peers)

Still there must be decided if machine learning or something easier will be used to create the matchmade profile. Attributes will be used to do the matchmaking, which means that some specifics things will not be counted in the offered activity's based on the preferences the user gave. Also, environment data like flora and fauna, alcohol restricted areas and works of art can be used. These kinds of data can be combined with things like:

- Data about sport associations (e.g., their locations)
- Content-based cognitive filtering (Make use of what the user had red, recommended to the user)
- Collaborative filtering (Make use of interest of comparable users)

Aiming for that the profile data will only stay on the user's device and the matchmaking will take place on their device. Only the process will be monitored to look how successful the platform is.

6.1.4 Phase 2

In this second phase, the participants discuss who is involved in the deployment and use of the Vitality Platform. This resulted in a list of actors which can be seen in Table 6.

Table 6 Actors

Actors
Sportclubs within Eindhoven, but also just outside Eindhoven.
Sport- and event organizers
Associations, unfettered sports groups and communities.
Sponsors of sports activities and sports clubs.
Sports directors.
College and council.
Municipality of Eindhoven's sports policy and implementation.
The great potential of people who are not currently active in sports, but who might want to be but don't know where to start.
(Un)sporting expats who want to practice a (new) sport.
Specific target groups, such as people who are not digitally skilled, do not have a smartphone, but may want to be matched.
People who want to move to Eindhoven and people who are new to the city, including status holders and expats and their spouses.
Employers.
General practitioners.
Healthcare institutions.
Health insurers.
Education.
Social neighbourhood teams and community centres.
Infrastructure.
Algorithm designers, technology partners and platform administrator.
The current platforms from which we get data.
Media.
Other parties that stimulate sport and vitality.

They also consider the positive and negative effects of this deployment. This can be seen in Table 7.

Table 7 Positive and negative effects

Positive effects	Negative effects
The platform maps out a lot of up-to-date information about sports and informs residents.	Risk of interference care.
The platform offers 'support' for newcomers to the city.	Chance of manipulation by companies.
Better insight is gained into who is and who is not reached.	Diversity of sports on offer is declining.
A low-threshold entry into (more) sports.	Data requires maintenance.
A wider reach for sports providers, who are better able to find their target groups via the platform.	Target groups get biased.
It will also be easier for sports clubs to find 'customers'.	Digi bête older generation not connected.
The platform may provide new 'ideas' for activities.	Invasion of privacy, personal data.
Opportunity to profile Eindhoven with innovation and sport (tourism).	Tunnel view by algorithm: you 'limit' people from the algorithm view. The user does not know what is available, only what suits you according to the platform.
Opportunity for health insurers to encourage healthy behaviour.	Health insurer: unhealthy people punished.

If the platform is very effective, it contributes to the vitality of the city and helps to counteract downward spirals in some neighbourhoods.	Sports director offside.
Better overview for the municipality.	Energy consumption of technology.
Broader than sport: exercising together means meeting each other and making more contacts.	Media Suspicion
The platform gets more people moving.	Not in line with need
It increases pro-activity with the government (as long as the resident stays in the lead).	Possible vitality sports providers are excluded
There is a good business model behind it.	Competition between cities, sports tourism.
	Hassle about the border, who can/may/wants to join?
	Costs for the municipality.
	Data leaks/security risks.
	Spiral: bad neighbourhoods even worse, etc.
	If it doesn't work/is effective.

And the last step is to identify important values that must be taken into account when deploying the Vitality application. The relevant values were inventoried by the facilitators while the participants were busy identifying effects. Values are hidden behind many effects. This list was shared with the participants and discussed and supplemented by them. This resulted in the following list of values in Table 8, in which the bold values were identified as core values for the solution.

Table 8 Values

Values	
Autonomy & privacy	Equality of opportunity (playing field)
affordability	Customer service
Reliability	Positive Imaging
Durability	Renewal
Effectiveness	Care & health
Inclusivity	

The Core values will be further discussed in Phase 3 in the following chapter.

6.1.5   Phase 3

During the third phase, action options were introduced. For these action options, the three core values that were introduced in phase two are used. The participants were divided in three different subgroups to discuss the different topics, namely design of technology, the adjustment from the environment it operates, and the behaviour of humans. The outcomes and action options from these conversations can be found in Table 9.

	reliability	Effectiveness	Inclusivity
Technology	• Reliable algorithm • Ensure data security • To what extend is data shared (un)wittingly	• Rapidness application • Matchmaking • User friendly	• Depth/width offer • Minimal question set • More connection with healthcare, education

	• User profiling • Explaining certain outcomes	• Quality factor to measure quality • Technology doesn't solve everything	• Involving digital illiterates • Unexpected matches • Sport directors
Environment	• Law and regulations • Openness algorithms and data • train sport associations in data quality • shared ownership, independent control	• communication • vitality cargo bike • guidance from healthcare and neighbourhood directors	• overarching coalitions and partner programs • explanation criteria and reviewing body • insights about membership from sport associations • use additional catches, contribute to social card
Human	• early inclusion media • privacy assessment • testing algorithm for functioning and responsibility	• campaign for expats • conversations with sport directors	• conversations with different stakeholders, sharing experiences • keeping an overview about supervision app • helping digital illiterates • different types of campaigns • neighbourhood specific contact • trial lessons • supply and demand

Table 9 outcomes topic conversations

6.2 TICT TOOL:

The Technology Impact Cycle Tool (TICT) is used by Datastic to dive deeper into the value inclusivity. During the framework of counselling ethics, a couple of values were mentioned but inclusivity was not discussed. The Technology Impact Cycle Tool helps you to estimate the impact of (new) technology. This tool does that by asking questions. By answering the questions, you will gain insight into a lot of facets that are involved in determining the impact of technology. This tool will not tell you what is right or wrong or better. It will help you think about the impact of a (new) technology by asking the right questions so you can make better decisions (Fontys University, 2019).

oOo DATASTIC

Figure 38 TICT-tool municipality

The value of inclusivity dives into built-in biases. With a brainstorm, the built-in biases can be found. The questions to keep the brainstorm going were: Can you find a built-in bias in this technology? Maybe because of the way the data was collected, either by personal bias, historical bias, political bias or a lack of diversity in the people responsible for the design of the technology? How do you know this is not the case?

The built-in biases are show into a mindmap. The built-in biases are from different perspectives:

- Sport facilitators (shown in green)
- The municipality of Eindhoven (shown in red)
- Users of the Vitality application (shown in blue)

Mapping these biases will ensures that you can take certain aspects into account and not exclude anyone and thereby take the value Inclusivity into account.

6.3 ETHICAL ADVICE:

During the development of the Vitality platform application, three core values must be taken into account. The three values have been established in the counselling ethics. Datastic made an in depth research with the TICT tool of one of the three values.

The TICT tool was used for the core value "inclusivity". This core value was looking at built-in biases. A couple of these were formulated during a brainstorm session. Every target group was colored different to show the diversity in built-In biases between the target groups.

Datastic advice to do this kind of research with the other values as well. The outcome of the research must be included in the making the application to take everything into consideration.

7   LAW

Datastic has consulted an external group to discuss the aspects of law. The external group comes from the law education at Avans university of applied sciences. They did a research which resulted in an advice on how to take the aspects of law into account during the making of the vitality application.

For developing an AVG compliant app, Datastic needs to proces the personal data that are actually required so that the app can function properly. At the moment the name, address, date of birth, gender, and the email are significant. Before taking any personal Datastic must ask permission prior before personal data are taken. It is advised to remove the data (or have it removed) if the user has not used the app over a year or if the user deletes the account. The users must be able to withdraw their consent with a click of a button. The default settings of the app must be set privacy friendly, and the users should be able to change these settings if they wish to.

The cookies can be divided in two types, cookies that requires permission and cookies that not require any permission. Functional and analytical cookies can be placed without the consent from the user. However, it must state the use of the cookies in the cookie statement. Furthermore, the use of cookies that use tracking or third-party cookies requires permission from its user. This can be done by using a banner, where the cookie banner has not be checked beforehand. To make the app cookie compliant it must have a cookie statement and a privacy statement.

When developing the app, the copyrights apply on the source code and the design of the app. These two parts meet the requirement which has been stated in paragraph 5.1. In regard to the swipe function: it can be considered as a functionality. Functionality is not protected under the copyright, because it cannot be said that creative choices were made by the functionality. If the swipe function is protected by the patent law, it is possible that Datastic is violating the patent right of a different user by using of the swipe function. Therefore, it is preferable to require an investigation by a specialist advocate or patent attorney.

An open source code is a cluster of all the types of software whose source code is made available freely. The person who changes the source code is known as the creator. for an open source code there are no limits by law and the creator has allowed to use code under certain conditions. This means that no breach has been made against the copyright towards previous users. Given that an third party is developing the app, they get the copyrights. The municipality of Eindhoven can take over the open source code only if the external party has granted a license. Therefore, the external party gives permission to the municipality of Eindhoven to use of software development.

The basic rule is that a work cannot be made public without the permission from the author. If it is about hyperlinks none of the work will be made public. By embedding it involves of replication/publication. The concept of public is important, because of the following question: "Is an announcement being made to a 'new' public?". If that is so, then there is a breach on the copyrights of the author. The works are available on the internet. The public "internet users" has already been reached. For this reason, Datastic is free to scrape the websites, as the information is not planned for specific user and is legal.

Datastic can scrape both www.eskbvimpact.nl and www.balans-yoga.com without any infringement on the dabase, simply because there is no database. Whereas, www.mygymclub.nl, all the basic information can be used from the site, however, when a user needs to log in, the information cannot be used.

Datastic makes no infringement on privacy law when the websites are being scraped, since there are no personal data that are being scraped. There is no violation made on the terms and conditions of

these websites when they are being scraped as well. This is because it is not excluded from the conditions. It is suggested to ask permission from the administrators of the website for web scraping. Since there is a chance in the future that web scarping can be excluded in the terms and conditions A trademark can be protected by third parties once it has been registered. Before a trademark is registered, it must be controlled by the trademark register to see if it did not violate an existing trademark. The fact that the app is focused on Eindhoven, it is therefore suggested to do a Benelux-registration at the office of Benelux Bureau for Intellectual Property (BOIP) by a professional.

The municipality of Eindhoven needs to make it clear to the user that there is interaction with an AI system. The conditions of legality and fairness should be recognized by asking approval from its user for taking personal data. It must be noted clearly how the algorithm works. Also, the municipality of Eindhoven and Datastic needs to control the results on fairness, and if the system has any unjustified animosity. Furthermore, they only need to use applicable data for the period which data are required. They need to astute and note how the algorithm works and which factors are important for the result.

8 ADVISE AND RECOMMENDATIONS

The municipality asked Datastic to create a prototype based on the webscraping and matchmaking part of the platform of vitality. Datastic divided this project in separate phases where each phase had their own outcome.

Firstly, during the first phase end-users liked to see the description, timetable, place, location, recent activity, age, experience level and price for each activity within the application. In the second phase of the project, the web scraping took place. During this phase Datastic figured out that almost all of the websites didn't contain open activities. There is no content to scrape and the HTML structure for each website diverse a lot compared to others. Also, the content that end-users wanted to see in the application, is not available at the activities. Datastic advice to do one of the following things:

- Create your own websites where associations or commercial companies can put their activity into a form. This means every activity will be putted into the Municipality's site in the same structure, so it is easy usable for their platform (this will reduce the costs) and every needed content can be asked to the organization.
- Contact the organizations about your plans and perform research if they are willing to update their websites with the content the municipality likes to scrape.

In the third phase matchmaking took place. There were a few options available to do the matchmaking with, like machine learning and a rule engine. Datastic chose for a rule engine because the machine learning wasn't possible with the available information. Datastic advice the municipality to keep using the rule engine at the first phase of the platform. This rule engine fits perfectly to the desired swiping solution the municipality wants. In a later stage of the platform machine learning could be added by the municipality which can offer activities to the users based on their previous preferences. The municipality must consider which solution they will use because of the complexity of the model.

During the phases also the ethical part was taken in mind by Datastic. Based on different core values a couple of target groups were determined. These target groups all have different expatiations about the application. The result of these expectation can be found in the ethical chapter. Datastic advice the municipality to use these expectations when they are developing the platform. Also take in mind to look further then the used core values.

7. BIBLIOGRAPHY

Berlioz, C. (2019, 1 25). *Best Practices Series: Rules Writing Pitfalls*. Retrieved from Sparkling Logic: https://www.sparklinglogic.com/best-practices-series-do-not-write-rules-with/

Build a Recommendation Engine With Collaborative Filtering. (n.d.). Retrieved from Realpython: https://realpython.com/build-recommendation-engine-collaborative-filtering/

Collaborative Filtering . (2021). Retrieved from https://developers.google.com/machine-learning/recommendation/collaborative/basics

Developer. (2021, 09 20). *Parse - MDN Web Docs Glossary: Definitions of Web-related terms | MDN*. Retrieved from Developer: https://developer.mozilla.org/en-US/docs/Glossary/Parse?retiredLocale=nl

Elite Data Science. (n.d.). *5 Tasty Python Web Scarping Libraries*. Retrieved 2021, from Elite Data Science: https://elitedatascience.com/python-web-scraping-libraries#beautiful-soup

elitedatascience. (2020, 06 09). *5 Tasty Python Web Scraping Libraries*. Retrieved from elitedatascience: https://elitedatascience.com/python-web-scraping-libraries

F.Marchioni. (2012). *What is a rule engine?* Retrieved from Mastertheboss: http://www.mastertheboss.com/bpm/drools/what-is-a-rule-engine/

Fontys University. (2019, 08 12). *Technology Impact Cycle Tool :: About*. Retrieved from Technology Impact Cycle Tool: https://www.tict.io/about

Fuchs, K. (2014, 4 18). *Best Practices When Creating A Decision Tree For Customer Support*. Retrieved from Yonyx: https://corp.yonyx.com/automate-support/best-practices-creating-decision-tree-customer-support/

Gainsight. (2020, 7 27). *Best Practices of Rules Engine*. Retrieved from Gainsight: https://support.gainsight.com/SFDC_Edition/Rules_Engine/Examples_and_Tutorials/Best_Practices_of_Rules_Engine

Geeks for Geeks. (2021). *Desision Tree*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/decision-tree/

html5rocks. (n.d.). *How Browsers Work: Behind the scenes of modern web browsers - HTML5 Rocks*. Retrieved from html5rocks: https://www.html5rocks.com/en/tutorials/internals/howbrowserswork/#HTML_Parser

JobsPikr. (18, 08 2020). *Top Programming Languages For Job Data Scraping*. Retrieved from JobsPikr: https://www.jobspikr.com/blog/top-programming-languages-for-job-data-scraping/

Katiyar, R. (2019). *What is Rule-Engine?* Retrieved from Medium: https://medium.com/@er.rameshkatiyar/what-is-rule-engine-86ea759ad97d

Koersen, W. (2018, 06 21). *Web Scraping, Regular Expressions, and Data Visualization: Doing it all in Python*. Retrieved from Towardsdatascience: https://towardsdatascience.com/web-scraping-regular-expressions-and-data-visualization-doing-it-all-in-python-37a1aade7924

MHerman. (2012, 05 10). *Web Scraping with Regular Expressions*. Retrieved from MHerman: https://mherman.org/blog/web-scraping-with-regular-expressions/

MoSCoW Prioritization. (n.d.). Retrieved from ProductPlan: https://www.productplan.com/glossary/moscow-prioritization/

Nayak, R. (2019, February 16). *Web Scraping using Beautiful Soup and Selenium for dynamic page*. Retrieved from www.medium.com: https://medium.com/ymedialabs-innovation/web-scraping-using-beautiful-soup-and-selenium-for-dynamic-page-2f8ad15efe25

Octoparse. (n.d.). *Using This RegEx Tool to Match HTML Tags*. Retrieved from Octoparse: https://www.octoparse.com/blog/using-regular-expression-to-match-html#

Paul, T. (2021, 06 24). *Datahut*. Retrieved from 13 Web Scraping Best Practices and Tips: https://www.blog.datahut.co/post/web-scraping-best-practices-tips

ProWebScraper. (2021). *12 Web Scraping Best Practices You Should Follow in 2021*. Retrieved from https://prowebscraper.com/blog/web-scraping-best-practices-you-should-follow/

Prowebscraper. (2021, 10 14). *The 5 Best Programming Languages for Web Scraping – ProWebScraper*. Retrieved from Prowebscraper: https://prowebscraper.com/blog/best-programming-language-for-web-scraping/

Rizvi, M. (2021, 07 26). *Web Scraping in Python | Python Scrapy Tutorial*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/#:~:text=Scrapy%20is%20a%20Python%20framework,your%20preferred%20structure%20and%20format.

Sharma, A. (2020, April 24). *5 Popular Python Libraries to Perform Web Scraping*. Retrieved from analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2020/04/5-popular-python-libraries-web-scraping/

Sharma, A. (2020, 04 28). *Python Library For Web Scraping | Web Scraping Tools Python*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/04/5-popular-python-libraries-web-scraping/

Stackabuse. (2019, 04 10). *Introduction to the Python lxml Library*. Retrieved from Stackabuse: https://stackabuse.com/introduction-to-the-python-lxml-library/

Study tonight. (n.d.). *Drools: When not to use a Rule Engine*. Retrieved from Study tonight: https://www.studytonight.com/drools/not-use-rule-engine

Summit. (n.d.). *Reading Data from the Web: Web Scraping & Regular Expressions*. Retrieved from Summit: https://www.summet.com/dmsi/html/readingTheWeb.html

Techopedia. (2017, 03 23). *Parse*. Retrieved from techopedia: https://www.techopedia.com/definition/3853/parse

Tippett, D. (2020, 07 22). *If You are Web Scraping Don't Do These Things - Level Up Coding*. Retrieved from Medium: https://levelup.gitconnected.com/if-you-are-web-scraping-dont-do-these-things-2cba2ebe5b29

Zyte. (14, 10 2021). *Best Practices For Web Scraping*. Retrieved from Zyte: https://www.zyte.com/learn/web-scraping-best-practices/

8. APPENDIX

8.1 INTERVIEWS USER REQUIREMENTS

interview A

1. What is your age?
   21
2. Where are you from?
   Eindhoven (used to live in Steenbergen)
3. Are you attending (practicing) a sport now?
   Yes
4. Would you like to assay (practice) new sport activities?
   No, it is already quite busy with my current sport activities.
5. Do you know where to look for new sport activities in Eindhoven?
   Yes, I think for students the SSCE offers a lot of activities.
6. How do you usually find out about sport activities/events in Eindhoven?
   SSCE, social media or via other friends that do a certain sport.
7. Would you get motivated to sport if you could find all sports in one place?
   Personally not, but I think if you are still finding out which sports you like it would be very helpful to have this overview.
8. What kind of sport websites do you know?
   I look at the website of the SSCE
   a. What do you think about the existing websites? (eindhovensport.nl)
      I hadn't heard from it till now. However, the website seems well organized and you can find any sport you like.
9. Would you like to get a suggestion by Eindhoven sports website which sport activity/event is suitable for you?
   Yes maybe. I would certainly be considering it more than when I have to search myself.
10. Would you like to have a personalized sports profile?
    Not now, but perhaps when I am older and looking for some new sports/activities.
    a. Are you willing to share personal data for the personalized sports profile?
       Yes
11. Which kind of information about the sport activities would you like to see on the platform?
    Description and perhaps pictures from previous similar activities to get an idea of what the event looks like.

Interview B

1. <u>What is your age?</u>

23

2. <u>Where are you from?</u>

The Netherlands, Almere/ Eindhoven

3. <u>Are you attending (practicing) a sport now?</u>

Yes, gymnastics

4. <u>Would you like to assay (practice) new sport activities?</u>

Yes

5. <u>Do you know where to look for new sport activities in Eindhoven?</u>

No, only on the SSC site.

6. <u>How do you usually find out about sport activities/events in Eindhoven?</u>
My roommate

7. <u>Would you get motivated to sport if you could find all sports in one place?</u>

Yes, it will trigger me to try a new sport.

8. <u>What kind of sport websites do you know?</u>

https://ssceindhoven.tue.nl/

9. <u>What do you think about the existing websites? (eindhovensport.nl)</u>
Clear. Didn't hear from eindhovensport.nl before, but it looks nice.

10. <u>Would you like to get a suggestion by Eindhoven sports website which sport activity/event is suitable for you?</u>

Yes, that will be a reason to try a new sport/ activity.

11. <u>Would you like to have a personalized sports profile?</u>

Yes

12. <u>Are you willing to share personal data for the personalized sports profile?</u>

Yes.

13. <u>Which kind of information about the sport activities would you like to see on the platform?</u>
- Timetable
- Place
- Information about the activity
- What you need to bring to the activity

Interview C

1. What is your age?

23

2. Where are you from?

Maastricht

3. Are you attending (practicing) a sport now?

Yes, Hockey

4. Would you like to assay (practice) new sport activities?

No

5. Do you know where to look for new sport activities in Eindhoven?

Yes, at the SSC. (= Student sport centrum)

6. How do you usually find out about sport activities/events in Eindhoven?

From new letters from the SSC or from friends.

7. Would you get motivated to sport if you could find all sports in one place?

No, I only want to do my own sport (hockey).

8. What kind of sport websites do you know?

I do not really use sport websites. I only use the website of SSC for assigning to group lessons or fitness sometimes.

   a. What do you think about the existing websites? (eindhovensport.nl)

I do not use eindhovensport.nl but I looked at it and I think that on the website the overview of what sports you can do is not really clear but when you are searching for a specific sport it is clear where you can do that sport.

9. Would you like to get a suggestion by Eindhoven sports website which sport activity/event is suitable for you?

Yes

10. Would you like to have a personalized sports profile?

No, I do not want to sport that much.

   a. Are you willing to share personal data for the personalized sports profile?

Depends on the data you need from me.

11. Which kind of information about the sport activities would you like to see on the platform?

Just the time, the location and a brief introduction about the sports. (a few sentences).

Interview D

1. What is your age? 22
2. Where are you from? Nederland
3. Are you attending (practicing) a sport now? Yes
4. Would you like to assay (practice) new sport activities? Not eager to, but if I had the time I would have tried some new sports.
5. Do you know where to look for new sport activities in Eindhoven? Yes
6. How do you usually find out about sport activities/events in Eindhoven? SSC or facebook
7. Would you get motivated to sport if you could find all sports in one place? Yes
8. What kind of sport websites do you know? SSC
    a. What do you think about the existing websites? (eindhovensport.nl) Overzichtelijk
9. Would you like to get a suggestion by Eindhoven sports website which sport activity/event is suitable for you? No
10. Would you like to have a personalized sports profile? Yes
    a. Are you willing to share personal data for the personalized sports profile? Yes
11. Which kind of information about the sport activities would you like to see on the platform? When the next lesson is (plus information about that lesson). My recent activity.

Results survey user requirements

Age	Origin	Current sports	New sports	Find out about new sports	Current source of information	Motivated for new sport if info easily available	Known sport websites	Suggestion which sports are suitable	Personalized profile	Information available on platform
46	Eindhoven	Fitness, skating	No	No	Internet	Maybe	No specific one. Just start googling	Could be usefull. Don't know what I am missing now.	No, would like to keep it to myself.	Information about attending sport on different age level and also about sport events as a spectator.
22	miedzyrzecz, Poland	Squas, Fitness	Yes	no	Facebook	Yes	Non	Yes, this would decrease the barrier of entry. This makes me look at sports and have the feeling of oh hey, I wanted to try XYZ let me try it.	no, They do not need that much information of mine	Different sports highlighted, where we can also have a free session to try it once
23	Lithuania	Yes, gym	Yes	University boards / online	Online	Yes	Dont know any specific		Yes, as it would make finding more relevant offers easier and it would be more attractive for individual users	Time, what is needed to prepare, etc
23	Bulgaria	Yes, I am going to the gym	No	Yes	Through gym ads, facebook ads, ads in general	Yes	Trainmore, TUe's gym website	No, I dont like my mail to be filled with ads regarding those kind of things	That sounds like a good idea, because you'd be able to track your progression	Schedule, type of sport, and description to it

8.2 WEBSCRAPING TECHNIQUES

Web scraping tools are designed to extract, collect any public information from websites. These resources are needed when you need to quickly receive and store any data from the Internet in a structured form. This kind of software searches for information under the control of the user or automatically by selecting new or updated data and storing it in such a way that the user has quick access to it. For example, using web scraping, you can collect information about products and their prices on the website. Web scraping tools can be used for different purposes and in different scenarios such as collecting data for market research, retrieving contact information, finding jobs or employees, track prices across stores.

8.2.1 HTML Parsing

Parsing means analyzing and converting a program into an internal format that a runtime environment can run (Techopedia, 2017). When parsing is done sentences or group of words are breaking up into separate components, including the definition of each part's function or form (Developer, 2021).

The exact definition of html parsing is to parse the HTML mark up into a parse tree. HTML parsing involves:

- Tokenization: HTML tokens include start and end tags, as well as attribute names and values. The parser parses tokenized input into the document, building up the document tree.
- Tree construction: The output tree (the "parse tree") is a tree of DOM element and attribute nodes. DOM is short for Document Object Model. It is the object presentation of the HTML document and the interface of HTML elements to the outside world (html5rocks, n.d.).
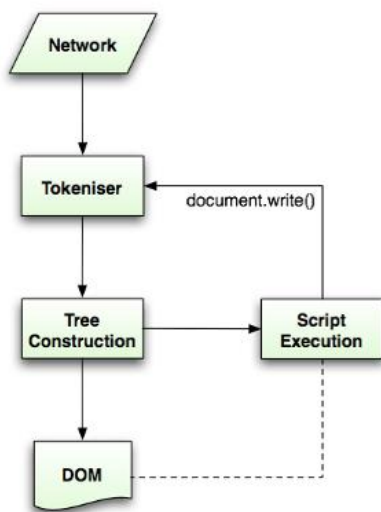
Figure 39 HTML parsing (html5rocks, n.d.)

8.2.2 HTTP Programming

Hypertext Transfer Protocol, or HTTP, is an application-layer protocol for collaborative, distributed, hypermedia information systems. HTTP is used for data communication on the internet, and TCP/I based. When a web client requests something, the HTTP protocol responds to that. The basic architecture is visualized in Figure 1. The client is somebody that is for example visiting the website. (Tutorials Point, sd)
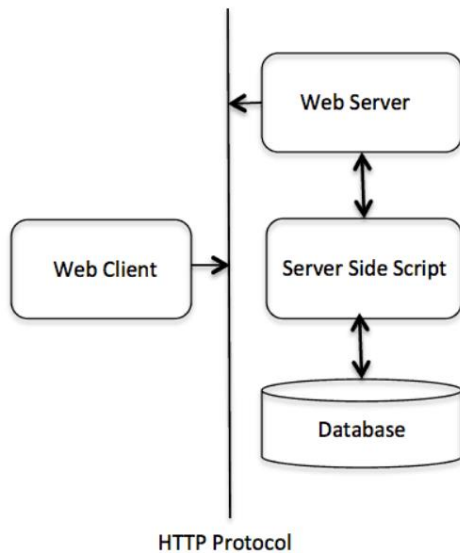
Figure 1 Basic architecture HTTP

8.2.3 Text Pattern Matching

Text Pattern Matching is a technique of web scraping that can be used by regular expressions matching facilities of programming languages. The most common used languages that are used with this technique are Perl or Python (Koersen, 2018). Regular Expression (RE) allows the user to find specific patterns of text and extract data they want more easily than manually searching for specific characters in the webpage. RE are very powerful and can be used to solve many parsing problems (Summit, sd). In the example below there is shown how the title is scraped from the web with the link attached to it (MHerman, 2012):

```
xmlTitle = re.compile("&lt;title&gt;(.*)&lt;/title&gt;")
xmlLink = re.compile("&lt;link&gt;(.*)&lt;/link&gt;")
```

This is used within Python with the re. () Expression. As there is shown, specific HTML part are written inside the brackets, so the scraper knows which kind of pattern it should get from the website. Below in table 1 are Pros & Cons of Regular Expressions (Octoparse, sd):

Pros	Cons
Can find multiple strings	When missing closing brackets in HTML code, it is encountering problems
Easy to use	Experienced programmers should use other tools

Table 1 - Pros & Cons Regular Expressions

8.2.4 DOM parser

The DOMParser interface provides the ability to parse XML or HTML source code from a string into a DOM Document. You can perform the opposite operation—converting a DOM tree into XML or HTML source—using the XMLSerializer interface. The Document Object Model provides APIs that let you create, modify, delete, and rearrange nodes.
The W3C Document Object Model (DOM) is a platform and language-neutral interface that allows programs and scripts to dynamically access and update the content, structure, and style of a document."

The Document Object Model (DOM) is an official recommendation of the World Wide Web Consortium (W3C). It defines an interface that enables programs to access and update the style, structure, and contents of XML documents. XML parsers that support DOM implement this interface.

8.2.5   WebScraper.io

WebScraper.io is a free Chrome extension. This tool runs right in your Chrome browser, has detailed documentation and is straight forward to use. After you build out how you want to navigate and scrape the website you can watch the scraping take place in a secondary browser. After it's done scraping you can export the data to a CSV file.

oՍ0 DATASTIC

8.3 WEB SCRAPING LIBRARIES TO USE

In this paragraph we will research different web scraping libraries in Python. Based on research of (Prowebscraper, 2021) and (JobsPikr, 18) Python is the best option for web scraping. The research explains that Python is easiest to master with a gentler learning curve. Its statements and commands are very similar to the English language. And someone with a bit of coding experience can master the language in a matter of a week.

In python you can use dozens of libraries for web scraping. 5 libraries are researched to look for the best option. This is based on research of (elitedatascience, 2020) In this research they decided to feature the 5 Python libraries for web scraping that they love most. Together, they cover all the important bases, and they are well-documented.

8.3.1 Beautiful soup

Beautiful Soup is one of the more popular libraries that is used for webscaping. One of the reasons that Beautiful Soup is so popular is because it is easier to work with and well suited for beginners. Beautiful Soup's default parser comes from Python's standard library. Furthermore, Beautiful Soup can also be used with different parsers such as lxml or request. Unlike lxml is Beautiful Soup not as fast even while using lxml as a parser (Elite Data Science, sd).

An example of Beautiful Soup for scraping all the links for a website:

```Python
from bs4 import BeautifulSoup
soup = BeautifulSoup(contents, 'html.parser')
soup.find_all('a')
```

Advantages	Disadvantages
Requires a few lines of code	Slower than lxml
Great documentation	
Easy to learn for beginners	
Robust	
Automatic encoding detection	

Table 10 overview Beautiful Soup (Sharma A. , 2020)

8.3.2   Requests

The Requests library is vital to add to your data science toolkit. It's a simple yet powerful HTTP library, which means you can use it to access web pages. Moreover, it can access API's, post to forms, and much more. Its simplicity is its greatest strength.

Code example:

1)import requests

2)page = requests.get('http://examplesite.com')

3)print(page.text)

1) request library into python environment

2) gets web page from the brackets, "page" in this example is response object

3)show the results

Why use it?

Do not have to manually add query strings to URLs, or form-encode Post data

Advantages	Disadvantages
Simple	Retrieves only static content of a page
Basic/digest authentication	Can't be used for parsing HTML
International domains and URLs	Can't handle websites made purely with JavaScript
Chunked requests	
HTTP(S) proxy support	

Table 11 overview Requests

8.3.3   lxml library

lxml is a Python library which allows for easy handling of XML and HTML files, and can also be used for web scraping.
There are a lot of off-the-shelf XML parsers out there, but for better results, developers sometimes prefer to write their own XML and HTML parsers. This is when the lxml library comes to play. (Stackabuse, 2019)

Benefits:

Advantage	Disadvantages
It's easy of use	Does not work well with poorly designed HTML
Extremely fast when parsing large documents	The official documentation is not very beginner-friendly
Uses element trees	
provides easy conversion of data to Python data types, resulting in easier file manipulation.	
Pythonic API	

8.3.4 Selenium

Selenium Library can be used for web scraping in combination with another library like Beautiful soup (Nayak, 2019). The original goal of Selenium was automated testing of web applications. However, nowadays it is a helpful tool for web scraping (Sharma A. , 2020). One of the biggest advantages of Selenium is that it capable of running JavaScript. This ensures that for example clicks on pages, fill forms and scrolls on pages can be done. This however does take a lot of time, since it is being done for every web page. In Table 12 an overview of the advantages and disadvantages from the Selenium library has been displayed.

Advantages	Disadvantages
Beginner-friendly	Very slow
Automated web scraping	Difficult to setup
Can scrape dynamically populated websites	High CPI and memory usage
Automates web browsers	Not ideal for large projects
Can do anything on a web page similar to a person	

Table 12 overview Selenium

8.3.5 Scrapy

Scrapy is a Python framework for large scale web scraping. It gives all tools that are needed to extract the data, process them, and store them in a structure the user wants. With Scrapy you don't need to reinvent the wheel again (Rizvi, 2021).

Scrapy is using a crawler that download all the content it browses from a website. Underneath an example is shown when the subreddit "Game of Thrones" is set in the crawler:

```
fetch("https://www.reddit.com/r/gameofthrones/")
```

It will be put in a response and when you open the response, the downloaded webpage will be shown within the browser:

```
view(response)
```
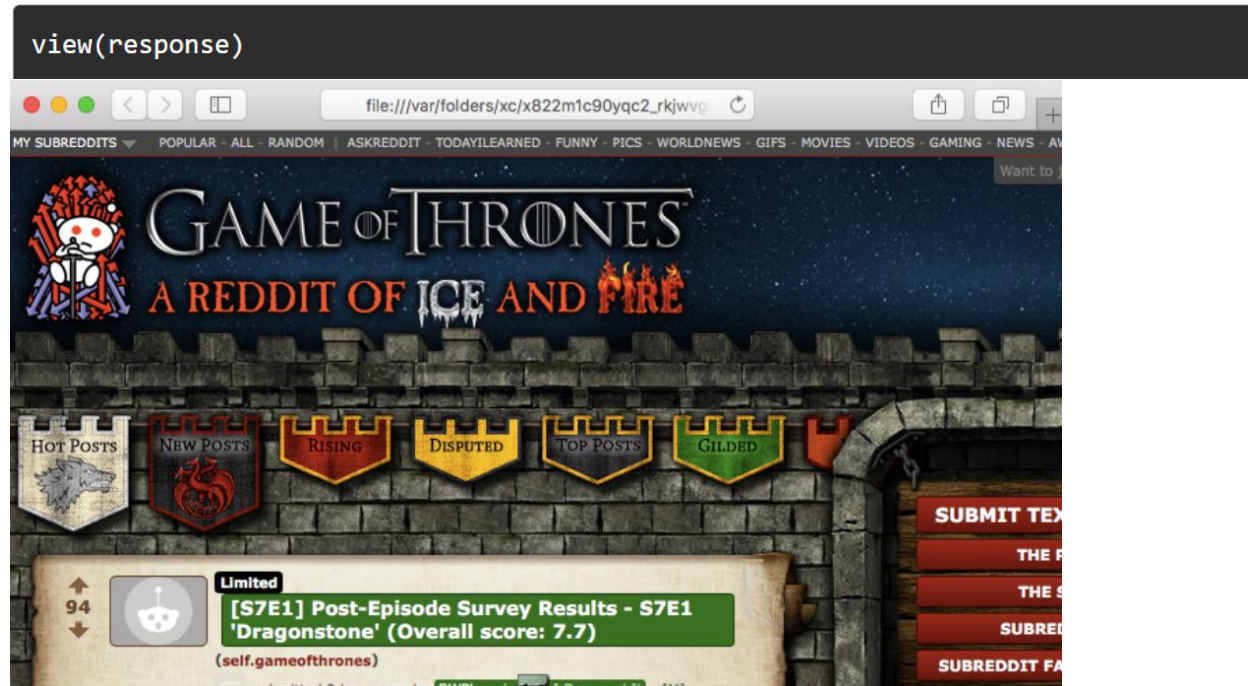

There will be a lot of content in the raw file that is scraped. To select the files the user needs, they have to take a look at the HTML part of the website. In Table 13- Pros & Cons Scrapy the advantages/disadvantages of scrapy are shown (Sharma A. , 2020):

Table 13- Pros & Cons Scrapy

Pros	Cons
Asynchronous, it can make multiple HTTP requests simultaneously. This saves a lot of time and increases efficiency.	Steep learning curve
Excellent documentation	Overkill for easy jobs
Various plugins to use	Not beginner-friendly
Create custom pipelines and middleware's	
Low CPU and memory usage	
Well-designed architecture	
A plethora of available online resources	

8.3.6 Conclusion

Based on the research Datastic found that a couple of libraries would be useful to use for the project. There were a few requirements that the tool should need:

- Easy to learn
- Able to scrape from multiple websites at once
- Free accessible to everyone

Based on these requirements the library's Scrapy, lxml and request didn't work for our project. They weren't beginner friendly, had difficult documentation or were not able to do HTML parsing. The most common used library was beautiful soup and based on the advantages Datastic is going to use this library.

8.4 BEST, GOOD AND BAD PRACTICES WEB SCRAPING

In this paragraph the best, good and bad practices are researched. This allows for various aspects to be considered before the actual web scraping.

Best practices

- API: IF an API is available, the query can pass into the API. If this is not available, Scraping is the only way to gather the information you need.

- Be gentle: Hitting the servers too often with web scraping affects the user experience of the target website. To handle this there are a few options available:
  - You can scrape during off-peak hours when the server load will be minimal compared to the peak hour.
  - Limit the number of parallel / concurrent requests to the target website.
  - Spread the requests across multiple IP's
  - Add delays to successive requests

- Respect robots: If you're attempting web scraping - it is probably a good idea to look at the robot.text file first. The robot.txt is a text file the website administrators create to instruct web scrapers on how to crawl pages on their website.

- Don't follow the same crawling pattern: Anti-scraping technologies on the website can block web scraping. To work around this it is a good idea to incorporate some random actions that confuse the anti-scraping technology.

- Route your request: When your request hits the server of a target website - they'll know and log it. The website will have a record of every activity you are doing on the website. Websites will be having an acceptable threshold on the rate of requests they can receive from a single IP address. Once the request rate crosses this threshold - the website will block the IP. There are several methods that can be used to change the outgoing of API:
  - VPN
  - TOR
  - Proxy services

(Paul, 2021)

- Do not harm the website: the volume and frequency of queries you make should not burden the website's servers or interfere with the website's normal operation. This can be done in a few ways:
  - Limit the number of concurrent requests to the same website from a single IP.
  - Respect the delay that crawlers should wait between requests by following the crawl-delay directive outlined in the robots.txt file.
  - If possible it is more respectful if you can schedule your crawls to take place at the website's off-peak hours.

- Don't violate copyright: When scraping a website you should always consider whether the web data you are planning to extract is copyrighted.

- Beware of login and website terms and conditions: When you log in and/or explicitly agree to a website's terms and conditions you are entering into a contract with the website owner, thereby agreeing to their rules regarding web scraping. Which can explicitly state that you aren't allowed to scrape any data on the website.

(Zyte, 14)

Good practices:

- Be clear and transparent: When asked to create a login and password be honest about who you are. If you hide who you are the website can block you from using their data.

- Use captcha solving services: some websites that you want to scrap have anti-webscraping measures in place. When you scrape data from these websites you will receive captcha pages. To bypass this there are captcha solving services to use.

- Use headless browser: When using a real browser there is a change that the website you want to scrap deploys bots' detection tools to prevent users from webscraping their website. The anti-bot tools check whether:
  - There exists support for nonstandard browser features.
  - Any bot-specific signature is present or not.
  - Any automation tool like Selenium or Playwright is present or not.
  - There are any random mouse movements, scrolls, clicks, tab changes, etc.
  To prevent this there are multiple headless browsers programs that can be used when webscraping.

(ProWebScraper, 2021)

- Watch out for honeypot traps: honeypot traps are links that are placed on a website to detect web scrapers. When a honeypot trap is activated, the server can confirm it is not a real human and start blocking the IPs or put the scraper into a wild goose chase draining the resources. These honeypots links usually have their background-color CSS property set to None to mask it from users.

(Zyte, 14)

Bad practices:

Underneath a couple of bad practices are written down where Datastic must look at before they start with the web scraping part of the project (Tippett, 2020):

1. **Don't hard code session cookies**

Anything that will be hard coded has the potential to fail miserably. For example: the client has a site that they want to be scraped, that requires a login. The time of life for that session is not clear so it can happen that the next time someone wants to access the site, an error occurs.

To tackle this problem, code your program to login and use the sessions to be sure they will be sent for every request.

2. **Don't DOS websites**

DOS means Denial of Service. When writing a for loop for a website, there is a chance your IP is getting banned because it will spam request in a short period of time. To be sure this will not happen, just create a simple delay, that will prevent you from getting banned.

3. **Don't copy and paste reusable code**

It's not good to copy and paste reusable code into the scraper. It is important to keep every code unique so it will be easier to track while scraping and you can come back to correct easier if necessary.

4. **Don't write single threaded scrapers**

If there is an array with a website that has multiple pages, make it in the same thread. This will have your scraper get better performance. However, if you have a different website, place it in a different thread.

5. **Don't use the same pattern for scraping**

Some websites are blocking visitors based on the same type of visiting. It's important to build something in your scraper that is accessing websites in a random order. This means that the detection of a website doesn't see it as the same type of visiting and prevent you from getting blocked from the website.

8.5 Expert interviews

8.5.1 Interview Pepijn from BMD

1. Have you ever used web scraping library such as beautiful soup?
   They did not implement a web scraper, it was the next step for this project
   They used web crawling
   Scrapy - parse the page itself, blacklist, whitelist specific id
   Mix of crawling and scrapy
2. What is your opinion about beautiful soup. Is it worth to use it in such project?
3. We have multiple sport providers, so we are curious how to make pattern to scrape multiple pages?
   Parce html in a nice way
4. Legal part of web scraping. Do we need permission? How do we avoid being blocked by the website?
   > Best practices when you r allowed to scrape and crawling
   > When you actually use to display this data from the beginning. Some news web sites let u use some parts of web page with reference on the link
   > We can add a lot of ethical stuff
   > Robots.txt files
   > Html headers
5. In what format can we get the output? (Best way to store it)
   > Format of output - sql database but some of the data can be not structured (Mango, db) JSON format - main, XML
   > Depends on what type of storage we need
   > Specific format wasn't set on our project
6. Can we use any templates for web scraping?
   > HTML standardization to adapt more web sites.
7. Difference between web scraping and web crawling and can we use web crawling for this matter?
8. Do we violate copyrights while web scraping logos?
9. What if the website changes the HTML structure? Will scraper stop working?
   Crawler will still work as it goes through links.
10. How not to harm website with queries?
    how long you should not visit web site, checking cache headers. Slow down the process

11. Some websites are using JavaScript to display content, BS doesn't allow to scrape this script content, what can we do in this case?
    https://pptr.dev/

12. There also extensions for browser like webscraper.io, is this also good to use?

Main conclusion: some of the content sources you have to make an agreement. Crawl all through the content you can find a lot of new sources. Make sure that they let us post it on the platform.

1. Best practices scope
2. Key words selectors

Can we proof this concept?
Put the number
Quality of the content is matter
It can be more manual process + automatic process
Think of tools what can help us to automate manual work

8.5.2 Interview Marco Langhuizen
1. How can sport activities that exist on Eindhoven Sport website be combined?

In the next 5 months, Datastic will help the municipality of Eindhoven with a project regarding sports in Eindhoven. The municipality of Eindhoven has commissioned research for a sports application. In this application, residents of Eindhoven will find information about sports. To obtain and store data we decided to use web scraping technique.

Right now, we have a list of websites such as eindhovensport.nl. We are looking for specific information such as name of the sport, name of the sport provider (171), opening hours, location, contact details, short description, target audience, logo image. We did research about web scraping tools and techniques, also good and bad practices.

2. List of questions

- Have you ever used web scraping library such as beautiful soup?

Yes, I used Rvest, Beautifull soup and Scrapy.

- What is your opinion about beautiful soup. Is it worth to use it in such project?

I think I t is the most popular library, the most comprehensive one. Good documentations. Recommendation to look at Scrapy. The whole idea is to scrape HTLM pages, containing elements. The whole idea of scraping is searching for tags, once one has been identified, search for the content in the tag. You have two ways of searching, CSS matching and Xpath matching.\
I thought beautiful soup was the most extended one and scrapy was a less-functionality version of beautiful soup. I thought Scrapy was easier to use then beautiful soup. My knowledge is outdated.

- We have multiple sport providers, so we are curious how to make pattern to scrape multiple pages?

You need to have an idea what the structure of the site is like. The only thing what you can do is, make a list of all the pages containing interesting information. We must know where we are looking for. Is it always in the same HTML element? It is not straight forward if everything is not in the same structure.

- Legal part of web scraping. Do we need permission? How do we avoid being blocked by the website?

Depending on the website. I don't think a regular soccer club in Eindhoven that any of them do mind scraping their site. The sites are always hosted by their IT-provider. Sometimes they don't like it that you are creating such a load on the servers. They might have a ban on it. You should look up how to get this solved. I don't think that would be a problem for small sport clubs. I don't know about the legal part of web scraping, you should ask this to an expert.

- In what format can we get the output? (Best way to store it)

Depends on what you want. I would say in some kind of database.

- Can we use any templates for web scraping?

I'm sure you can but probably not for the sites you want to scrape.

- Difference between web scraping and web crawling and can we use web crawling for this matter?

Web crawling is more like some type that agents, a program that on its own, search the web. If it finds an interesting page and follow the links on this page. I don't think this could be very interesting for you. It is more used in a way more open question. Not if you are looking in a very specific type of data.

- Do we violate copyrights while web scraping logos?

I think if you just scrape them, that won't be a violation but if you use them, that could be a problem.

- What if the website changes the HTML structure? Will scraper stop working?

Then you're screwed, that is the common problem. Depending on the structure, what they change. If they change the tags you are scraping, it won't work anymore.

You can try to get rid of all the html tagging and retain the raw text. Then you can use regular expression to search for specific keywords.

- How not to harm website with queries?

You can't, you can give some additional load, but I don't think you will run this in script.

- Some websites are using JavaScript to display content, BS doesn't allow to scrape this script content, what can we do in this case?

I don't know, I don't assume most of the use dynamic HTML. If they do, use something else (bypass).

- There also extensions for browser like webscraper.io, is this also good to use?

I Think to get known what is available on a website, I recommend always using something like that. I recommend starting with a small number of websites to test if its working and see later if its expendable.

8.6 Q & A FOCUS GROUP

- Do you think it is currently easy to find different kinds of sports?

As a student it is easy, if you are not a student it is more difficult.

- Is the main source of information indeed online?

Mostly google, sometimes also from friends.

- Would you be more likely to try out a new sport if it is recommended?

It depends, if I like it I would look into it and try it, if not then not. When it is more based on my preferences then I am definitely more likely. This goes for everybody.

- How much data are you willing to share for a personalized profile?

It depends, if I am not willing to share it I also would not like to get a new suggested sport.

It depends on the data they want and the way they use it. I mostly only want to share my preferences and some personal data but not to much. I would also like to see your recent activity

It would also be nice to see my progress. This also would keep me motivated to sport. Push notifications would also be nice to have.

- Are this the things you would like to see? Is there something missing?

Experience level would also be nice to see. If I want to join a football training, I want to know what the experience level is.

If there is a limit to the amount of people, I would like to know.

EINDHOVEN

Platform
Begeleidings
ethiek

Verslag workshop
Aanpak
Begeleidingsethiek
**Vitaliteitsplatf
orm**
Gemeente Eindhoven in
samenwerking met Fontys

ECP
Platform voor de
InformatieSamenleving

● **Voorwoord**

De toenemende digitalisering brengt ingrijpende veranderingen met zich mee, in allerlei sectoren. Nieuwe mogelijkheden voor communicatie, monitoring en analyse roepen vragen op. Wat is de positie van de mens in deze omgeving, hoe staat het met data, met privacy, wat doen algoritmes en wat willen gebruikers eigenlijk? Begeleidingsethiek is ethiek die zich specifiek met dit soort vragen bezighoudt. Hij is gebaseerd op techniekfilosofie, die uitgaat van eeuwenoude verwevenheid tussen mens en technologie.

De Aanpak Begeleidingsethiek is een concrete aanpak waarin betrokkenen met elkaar in dialoog gaan over de effecten van de nieuwe technologie én de waarden die daarbij in het geding komen. Dat zijn vaak waarden gelieerd aan autonomie van de gebruiker, efficiëntie van het proces, transparantie van het algoritme, privacy, et cetera.

Tijdens de sessie komen verschillende stakeholders als gebruikers, ontwikkelaars, beleidsvormers en beslissers met elkaar in gesprek. Na de dialoog hebben de deelnemers ethische handelingsopties gegenereerd, waarvan verschillende vaak direct opgepakt kunnen worden. Er is gezamenlijk gekeken welke waarden we in het digitale domein belangrijk vinden en hoe we die willen verankeren en borgen in digitale processen en handelingen. Dit alles om ook bij verdere digitalisering op het vertrouwen van de samenleving en de participanten kunnen blijven rekenen. Dit is niet in één stap te realiseren het is een continu proces waarin deze workshop een schakel is.

• Workshop aanpak begeleidingsethiek

Initiatiefnemers: Ran Haase en Michiel Oomen (gemeente Eindhoven)
Moderatoren: Daniel Tijink (ECP), Maike Popma en Robert van Rijssel (VNG)

Op 15 december 2021 organiseerde de gemeente Eindhoven op initiatief van Ran Haase en Michiel Oomen (gemeente Eindhoven) in samenwerking met Jo-Ann Kamp (Fontys) en Pepijn Verburg (Bureau Moeilijke Dingen) een sessie begeleidingsethiek over de casus 'Vitaliteitsplatform'. Aan de workshop namen 21 deelnemers (zie bijlage voor specificatie) deel onder leiding van Daniel Tijink (Ecp), Maike Popma en Robert van Rijssel (VNG).

De doelstelling is tweeledig:

▷ Leren tijdens de workshop
- Welke effecten, waarden en actoren worden genoemd.
- Welke handelingsopties zijn er om de toepassing te verbeteren.
▷ Leren over de Aanpak Begeleidingsethiek
- Hoe kan de organisatie deze methodiek gebruiken rondom ethische kwesties en technologie-implementaties.

Het verslag bevat de volgende elementen:

▷ Toelichting op de Aanpak Begeleidingsethiek.
▷ Weerslag van de workshop.

Aanpak begeleidingsethiek
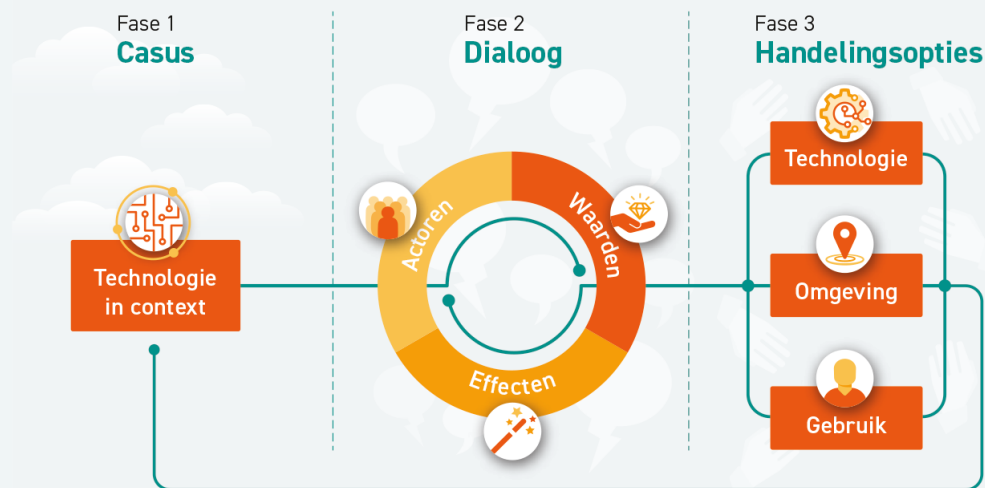
De aanpak bevat de volgende fasen:

- **Fase 0 Introductie**

Introductie over de doelstelling en een toelichting op het model en het gedachtengoed.

- **Fase 1 Toelichting**

Hoe ziet de technologie eruit en in welke context wordt deze toegepast

- **Fase 2 Dialoog**

▷ Een korte ronde waarin de deelnemers aan de workshop de betrokken actoren benoemen.

▷ Brainwrite waar deelnemers mogelijke effecten benoemen en bespreken.

▷ Benoemen van waarden die een rol spelen bij die effecten.

- **Fase 3 Handelingsopties**

In subgroepen gaan de deelnemers op zoek naar handelingsopties vanuit de technologie, de omgeving en het individu.

Technologie en context

Inleiding over de context en aanleiding van de casus

Welke verhouding en relatie ontwikkelen mensen tot techniek?
Op die ethische vraag wordt tijdens de workshop ingegaan. Recente voorbeelden, waaronder de inzet van een robot op het inwonersplein van de gemeente Leidschendam-Voorburg, die de bezoekers welkom heet, laat zien hoe snel bezoekers toch een vorm van relatie met de robot opbouwen.
Dat is een mooie analogie voor het thema waar we het vandaag over gaan hebben in het kader van het Vitaliteitsplatform.

Ethiek staat steeds meer in de belangstelling.
Ook in Eindhoven wordt ethiek belangrijk gevonden, de gemeenteraad heeft onlangs een ethisch afwegingskader vastgesteld. Er gaat ook een ethische commissie ingesteld worden. Nu we steeds vaker met ethische afwegingen te maken krijgen, is het belangrijke om na te gaan hoe je dat verantwoord doet.
Kom je wel toe aan een ethische discussie in een productieomgeving? De vraag die dan immers snel op komt is: **"Maar hoe dan wel?"**. In het gezamenlijk beantwoorden van die vraag doet de aanpak begeleidingsethiek een voorzet.

Hoe kun je de kracht van technologie inzetten om maatschappelijke opgaven te realiseren? Vertrekpunt voor het matchen van inwoners op het sportaanbod via een Vitaliteitsplatform is het gebruikersperspectief. Dus de gebruiker staat centraal. Er zijn klantreizen op het thema sport gemaakt, er is bekeken hoe die reizen eruitzien en waar er technische hickups zitten. Daaruit is een prototype en een projectvoorstel gekomen.

Door het lokaal sportakkoord is nu ruimte ontstaan om eerste stap in het plan te gaan zetten. Voornemen is om voor maart de contouren te verkennen, om dan een klap op de ontwikkeling van een Vitaliteitsplatform te kunnen geven. Recht en ethiek zou wel eens in die opdracht van doorslaggevend belang kunnen worden. De workshop is zeker ook bedoeld om vooral veel te leren, het vormt voor de initiatiefnemers een mooie testcase om met elkaar te experimenteren in dit vraagstuk.

Toelichting op de techniekfilosofie: klassiek versus modern

De lijn tussen mens en machine is dun, vloeiend.

De mens maakt techniek maar de techniek maakt en verandert ook de mens: er is sprake van continue wederzijdse aanpassing tussen mens en techniek, nu meer (digitaal) dan ooit. De aanpak begeleidingsethiek is een dialoog- en werkvorm die thuishoort bij de moderne ethiek, die een steeds prominentere positie **naast** de klassieke ethiek vervult.

Klassieke ethiek	Moderne ethiek
Focus technologie	Focus technologie & omgeving
Oordelen: of iets mag	Begeleiden: hoe iets goed kan
Top-down (kaders)	Bottom-up (praktijk)
Experts	Betrokkenen

Uitleg over de toepassing/technologie

Het platform wil bijdragen aan vitaliteit met behulp van sport.

Met het Platform voor Vitaliteit streeft de gemeente Eindhoven naar een integrale oplossing, naar een centraal platform voor vitaliteit. Het platform dient gebruikers, door de combinatie van het inventariseren van aanbod, het analyseren van gebruikers(gedrag) en matchmaking, persoonlijke suggesties te bieden voor sportactiviteiten in de stad.
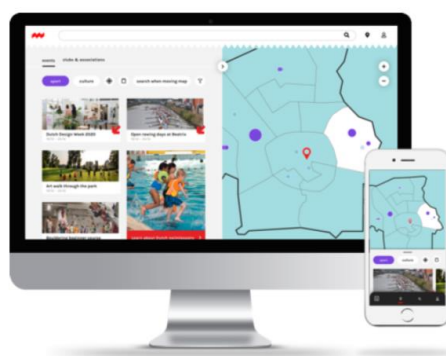
Hoe kun je het leven en bruisen van de stad ook plek geven in je aanbod?

Op het vlak van het in kaart brengen van aanbod liggen er kansen voor automatiseren, de stad heeft veel te bieden, de stad broeit, een deel is bekend en beschikbaar, een deel nog nauwelijks tot niet. Voor de inventarisatie van aanbod worden meerdere methoden ingezet, bruikbare content ontstaat door:

- Bronnen via API-koppelingen uit te lezen.
- Crawling en scraping van activiteitenkalenders zoals Meetup en Eventbrite.
- Curatie en handmatige classificatie (door de gemeente).

Voor de analyse van gebruikers(gedrag) zal zowel expliciet als impliciet het gebruikersprofiel in kaart gebracht worden.

- Explicit: op een engaging manier een profiel opbouwen.
- Implicit: delen van profielgegevens voor verbetersuggesties, bijvoorbeeld door de gebruiker te vragen om voorkeuren te delen met leeftijdsgenoten.

Er is nog geen keuze gemaakt of er gebruik gemaakt gaat worden van geavanceerde machine learning of een simpelere vorm van filtering. Wel is de verwachting dat een vrij eenvoudige semantische analyse zou kunnen voldoen.

Voor matchmaking worden attributen gedestilleerd, waarop je kunt afstrepen en matchen. De daarbij bruikbare data over de stad is breder dan de sportactiviteiten. Denk ook aan gebiedsdata over flora en fauna, alcoholverbodsgebieden, real-time fijnstofmonitoring, kunstwerken, locaties van geluidssensoren. Dat soort data wordt via data fusion gecombineerd met:

- Data over sportverenigingen, zoals de locatiegegevens van verenigingen.
- Content-based cognitive filtering (gebruik maken van wat de gebruiker gelezen heeft, recommended to user).
- Collaborative filtering (gebruik maken van interesses van vergelijkbare gebruikers, read by her recommended to him).

Uitgangspunt is dat profieldata standaard op het device van de gebruiker blijft en dat daar de matchmaking plaatsvindt. Om te meten hoe succesvol het platform is, wordt puur het proces gemonitord.

Het platform doet suggesties voor vitaliteit. Het genereert een persoonlijk aanbod, eigenlijk op dezelfde manier zoals iedereen er op dit moment al in aanraking mee komen op internetplatforms zoals bol.com en coolblue.

Om tot de belofte van meer vitaliteit te komen, zal ook community building nodig zijn. Zo kwam in de klantreizen die met expats zijn gemaakt naar boven dat er behoorlijk veel aanbod voor hen is, maar dat het niet makkelijk gevonden wordt.

Kan ik via het platform alleen een tennisbaan vinden of ook een tennismaatje?

Vanuit de technologie kunnen we het verbinden van mensen via het platform

vergemakkelijken, die functie is echter op dit moment nog niet in scope, maar is wel nadrukkelijk een wens voor de toekomst.

Dialoog

In deze tweede fase gaan de deelnemers in gesprek over wie er betrokken zijn bij de inzet en het gebruik van het Vitaliteitsplatform. Ook buigen zij zich over de positieve en negatieve effecten van deze inzet, en benoemen zij belangrijke waarden waar rekening mee gehouden moet worden bij de inzet van een dergelijke toepassing.

- ## Actoren:

Bij de actoren is de vraag wie er betrokken is of geraakt wordt door de case. De deelnemers aan tafel vertegenwoordigen al een deel van die actoren. Ze noemen de volgende betrokkenen:

Sportclubs binnen Eindhoven maar ook net buiten Eindhoven.
Sport- en evenementorganisatoren.
Verenigingen, ongebonden sportgroepen en communities.
Sponsors van sportactiviteiten en sportclubs.
Sportregisseurs.
College en raad.
Beleid en uitvoering sport gemeente Eindhoven.
Het grote potentieel aan mensen die nu niet sportief actief zijn, maar dat wellicht wel zouden willen zijn maar niet weten waar te beginnen.
(On-)sportieve expats die een (nieuwe) sport willen beoefenen.
Specifieke doelgroepen, zoals mensen die niet digitaal vaardig zijn, geen smartphone hebben, maar wellicht wel gematcht willen worden.
Mensen die naar Eindhoven willen verhuizen en mensen die nieuw zijn in de stad, waaronder statushouders en expats en hun spouses.
Werkgevers.
Huisartsen.
Zorginstellingen.
Zorgverzekeraars.
Onderwijs.
Sociale wijkteams en buurthuizen.
Infrastructuur.
Ontwerpers algoritme, technologiepartners en beheerder platform.
De huidige platforms waar we data vandaan halen.
Media.
Andere partijen die sport en vitaliteit stimuleren.
De vraag aan de deelnemers was om in het vervolg van de sessie ook te proberen vanuit deze perspectieven hun inbreng te geven.

- **Effecten:**

De vraag is hier welke positieve en negatieve effecten de invoering en het gebruik van het Vitaliteitsplatform hebben.

Mogelijke positieve effecten

- Het platform brengt veel up to date informatie over sport in kaart en informeert inwoners.
- Het platform biedt 'steun' voor nieuwkomers in de stad.
- Er ontstaat beter inzicht in wie wel en wie niet bereikt wordt.
- Een laagdrempelige instap om (meer) te gaan sporten.
- Een groter bereik voor sport aanbieders, die via het platform hun doelgroepen beter weten te vinden.
- Ook voor sportclubs wordt het makkelijker om 'klanten' te vinden.
- Het platform levert mogelijk nieuwe 'ideeën' op voor activiteiten.
- Kans om Eindhoven te profileren met innovatie en sport(toerisme).
- Kans voor zorgverzekeraars om gezond gedrag te stimuleren.
- Als het platform heel effectief is, draagt het bij aan de vitaliteit van de stad en helpt het neerwaartse spiralen in sommige wijken tegen te gaan.
- Beter overzicht voor gemeente.
- Breder dan sport: samen bewegen is elkaar ontmoeten en meer contacten.
- Het platform brengt meer mensen in beweging.
- Het vergroot de pro-activiteit bij de overheid (als de inwoner maar in de lead blijft).
- Er zit een goed businessmodel achter.

Mogelijke negatieve effecten

- Risico op bemoeizorg.
- Kans op manipuleren door bedrijven.
- Diversiteit sportaanbod neemt af.
- Data vergt onderhoud.
- Doelgroepen raken gebiast.
- Digibete oudere generatie niet aangesloten.
- Inbreuk op privacy, persoonlijke data.
- Tunnelzicht by algoritme: je 'beperkt' mensen vanuit het algoritme beeld. De gebruiker weet niet wat er allemaal is, alleen wat volgens het platform bij je past.
- Zorgverzekeraar: ongezonde mensen gestraft.
- Sportregisseur buitenspel.
- Energieverbruik van technologie.
- Media-achterdocht, zwart-wit
- Niet aansluit bij behoefte
- Mogelijk vitaliteitsportaanbieders worden uitgesloten
- Concurrentie tussen steden, sporttoerisme.
- Gedoe om de grens, wie kan/mag/wil erbij?
- Kosten voor de gemeente.
- Datalekken/veiligheidsrisico's.
- Spiral: slechte buurten nog slechter etc.
- Als het niet werkt/effectief is.

Waarden:

De relevante waarden zijn door de facilitators geïnventariseerd terwijl de deelnemers bezig waren met het benoemen van effecten. Achter veel effecten gaan waarden schuil. Die lijst is gedeeld met de deelnemers en door hen besproken en aangevuld. Dat leverde de volgende lijst met waarden op, waarin de vetgedrukte waarden als kernwaarden voor de oplossing werden aangewezen:

Waarden	
Autonomie & privacy	Kansengelijkheid (speelveld)
Betaalbaarheid	Klantvriendelijkheid
Betrouwbaarheid	Positieve beeldvorming
Duurzaamheid	Vernieuwing
Effectiviteit	Zorgzaamheid & gezondheid
Inclusiviteit	

Handelingsopties

In het proces van het opstellen van de handelingsopties zijn de waarden uit de vorige fase meegenomen. De deelnemers kregen een uitleg over de verschillende categorieën binnen de handelingsopties. Het kan bijvoorbeeld gaan over het ontwerp van de technologie, het aanpassen van de omgeving waarbinnen de toepassing functioneert en het gedrag van mensen. De deelnemers worden in drie werkgroepen verdeeld en gaan ieder aan de slag met een van deze drie gespreksonderwerpen. Dat leidt tot handelingsopties per categorie.

- ## Technologie:

  - Handelingsopties bij de kernwaarde *Betrouwbaarheid*:
    - Betrouwbaar algoritme (rechtmatigheid, behoorlijkheid, transparantie).
    - Dataveiligheid waarborgen.
    - Hoe (ver) worden data (on)bewust gedeeld/beschikbaar?
    - Gegevens veilig, data veiligheid.
    - Zichtbaar maken, uitleggen hoe je als platform rekening houdt met 'waarden'.
    - User profiling, gebruikersattributen/classificaties transparant laten zijn.
    - Vertellen laten zien waarom recommandaties komen.

  - Handelingsopties bij de kernwaarde *Effectiviteit*:
    - Snelheid applicatie.
    - Matchmaking: grootte vragenlijst t.o.v. afhaken.
    - Gebruiksvriendelijk.
    - Kwaliteitsfactor om te meten (geen like/star maar metric, meeteenheid). Hoe goed doet een match het?
    - Techniek lost niet alles op!

  - Handelingsopties bij de kernwaarde *Inclusiviteit*:
    - Diepte/breedte aanbod (wat wel niet gescraped).
    - Minimale vragenset, geen onnodige koppelingen.
    - Meer in contact komen met platform: zorg/onderwijs, niet digitaal.
    - 'Bewuste digibeten' ook meenemen.
    - Sportregisseurs bij de juiste plek in te zetten.
    - Ook tool voor sportregisseur.
    - 'Onverwachte matches'.

- **Omgeving:**

  - Handelingsopties bij de kernwaarde *Betrouwbaarheid*:
    - o Volgen van wet en regelgeving, openheid over algoritmes en data.
    - o Trainen van sportclubs in kwaliteit van data.
    - o Gedeeld eigenaarschap, onafhankelijke sturing.

  - Handelingsopties bij de kernwaarde *Effectiviteit*:
    - o Communicatie, ook op papier.
    - o Vitaliteitsbakfiets.
    - o Toeleiding/begeleiding door zorg en wijkregisseurs.

  - Handelingsopties bij de kernwaarde *Inclusiviteit*:
    - o Overkoepelende coalities en partnerprogramma.
    - o Toetsingscriteria en toetsingsinstantie.
    - o Inzichten rondom lidmaatschap bij sportclubs verzamelen.
    - o Bijvangsten benutten: bijdragen aan sociale kaart.

- **Mens:**

Waarde	Actoren	Ondersteuning
Effectiviteit	Sporters	Campagne, voor expats.
	Data analisten	Met de sportregisseur regelmatig in gesprek.
Betrouwbaarheid	Media	Vroeg betrekken in de pilot gedachte, nieuwsgierigheid naar de innovatie.
	Juristen	Privacy assessment, informatie over cloud of device opslag.
	App-ontwikkelaars	Toetsen van het algoritme en de werking daarvan en verantwoording.
Inclusiviteit	Sporters	In gesprek met anderen en ervaringen delen, iemand meenemen die te weinig beweegt.
	Sportregisseurs	Overzicht bewaren hoe deze app uitpakt en toezicht op app. Ouderen helpen met de digivaardige kant en zij

		blijven aanspreekpunt.
	Inwoners	Campagne, bijvoorbeeld buurtblaadjes of posters bij abri's.
	Buurthuizen	Buurtgericht contact maken, inloop voor inwoners en sportregisseurs.
	Sportclubs	Wie willen ze aantrekken en daar goed aanbod op maken, proeflessen bij een nieuwe sport.
	Zorg	In gesprek met anderen en ervaringen delen, iemand meenemen die te weinig beweegt.

● Terugblik
en afronding

Aan het einde van de workshop wordt teruggekeken en besproken wat de Aanpak Begeleidingsethiek concreet heeft opgeleverd. De deelnemers geven aan dat het veel nieuwe inzichten oplevert en dat er in korte tijd veel wordt gerealiseerd. Daarnaast wordt de brede groep van deelnemers als meerwaarde ervaren. De combinatie van een duidelijke casus, enthousiaste deelnemers, en een aanpak die zorgt voor duidelijke uitkomsten in een korte tijd was dus een succes.

● **De deelnemers**

1. **Ran Haase** jurist Recht en Ethiek, werkgebied smart urban planning (gemeente Eindhoven)

2. **Michiel Oomen** programma manager digitale innovatie (gemeente Eindhoven)

3. **Jo-Ann Kamp** docent/begeleider & ontwikkelaar technology impact cycle tool (Fontys)

4. **Pepijn Verburg** directeur/eigenaar (bureau Moeilijke Dingen)

5. **Claire Voskuijlen** communicatieadviseur Eindhoven Sport (gemeente Eindhoven)

6. **Louella Verleg** beleidsadviseur sport en bewegen (gemeente Eindhoven)

7. **Ben Wiermans** sportregisseur (gemeente Eindhoven)

8. **Bouke de Boer** fervent sporter (werkzaam bij Philips Medical Systems in Best)

9. **Yaniek Martens** student datadriven businesslab (Fontys)

10. **Lei Nelissen** UX-designer (bureau Moeilijke Dingen)

11. **Remco Bisschops** student datadriven businesslab (Fontys)

12. **Rik van Stiphout** programma-adviseur innovatie (gemeente Eindhoven)

13. **Pim te Loeke** trainee, onderzoeker sensortechnologie (Toekomst van Brabant)

14. **Denise Ortan** online marketingspecialist (Eindhoven Sport)

15. **Esma Vurgun** student Recht & Ethiek (Juridische Hogeschool Avans & Fontys)

16. **Maaike van Loosbroek** student Recht & Ethiek (Juridische Hogeschool Avans & Fontys)

17. **Rosa van den Boogaard** student Recht & Ethiek (Juridische Hogeschool Avans & Fontys)

18. **Shankari Mahendran** student Recht & Ethiek (Juridische Hogeschool Avans & Fontys)

19. **Lieke Nijs** student datadriven businesslab (Fontys)

20. **Femke Boogerd** student technische bedrijfskunde (Fontys)

21. **Edwin Borst** bedrijfs- en technologie ethicus (ECP)

Moderatoren:
Daniel Tijink (ECP), Maike Popma en Robert van Rijssel (VNG)