University of
South-Eastern Norway

www.usn.no

FM4017 Project 2018

Management of Environmental and Health related historical data in Grenland, using Machine Learning

MP-11-18

Faculty of Technology, Natural Sciences and Maritime Sciences
Campus Porsgrunn

University of South-Eastern Norway

www.usn.no

Course: FM4017 Project 2018
Title: *Management of Environmental and Health related historical data in Grenland, using Machine Learning*
Pages: *91*
Keywords: ...

Project group: *MP-11-18*
Group participants: *Slim Ayache*
Terje Eikerol
Mahdi Khoshbakhtian
Jørund Martinsen
Supervisor: *Hans-Petter Halvorsen*
External partner: *Hærøya Industripark, Sykehuset Telemark, Porsgrunn Kommune, Folkehelseinstituttet, Miljødirektoratet, Prosjekt kompetansesenter miljø og helse*
Availability: *Open*

Summary:
The goal of this project is to gather environmental and health-related data from the Grenland area, store this in a reliable, expandable and searchable solution and make this accessible to the general public, as well as researchers. Some sample data has been gathered and stored in a solution, as previously mentioned. Some data analysis, in the form of machine learning has been performed in this solution. In addition, contact data has been collected and is available to the customer. A discussion of how Machine learning can further improve the system is included along with several other suggestions for future projects.

Preface

This project is the master student project of University of South-Eastern Norway, in cooperation with Herøya industripark, Sykehuset Telemark, Porsgrunn kommune, Folkehelseinstituttet, Miljødirektoratet og Prosjekt kompetansetjenester miljø og helse. The purpose is to understand the basics of project management and team work in practice. In this project, two IT and Automation students and two Energy and Environmental students are involved.
This project is part of a larger project which will work as a reference data bank for gathering historic environmental and health related data of Grenland that can be used by ordinary people as well as researchers.
We discuss in detail the best technology available to deal with big data and analytic approaches. The result is shown in a website temporarily accessible by the address:
http://mehdwebapp.azurewebsites.net
We also thank professor Saba Mylvaganam, Martin Veel Svendsen and Tone Rabe for providing data for our project and special thanks to Hans Petter Halvorsen for supervising and assisting the project, and his work on the Project Management tool.

Porsgrunn, 20th November 2018

Slim Ayache
Terje Eikerol
Mahdi Khoshbakhtian
Jørund Martinsen

Contents

List of Figures

List of Tables

Nomenclature

Symbol	**Explanation**
CFC	Chlorofluorocarbon
CO	Carbon monoxide
CO_2	Carbon dioxide
DBMS	Database Management System
EDMS	Environmental Data Management system
EMS	Environmental Management system
FAO	Food and Agriculture Organization of the United Nations
H_2	Hydrogen
H_2S	Hydrogen sulfide
HBFC	hydrobromoflurocarbons
HCFC	hydrochlorofluorocarbons
HDFS	HADOOP Data File System
NO	Nitric oxide
NO_2	Nitrogen dioxide
NoSQL	Not Only SQL
NOx	Nitrogen oxide
O_2	Oxygen
O_3	Ozone
OH	Hydroxide
PAH	Polycyclic aromatic hydrocarbon
PCDD	Poly Chlorinated Dibenzo Dioxin
PCDF	Poly Chlorinated Dibenzo Furan
PM	Particulate matter
PM_{10}	Particle matter, smaller than 10 micrometer
$PM_2 \cdot 5$	Particle matter, smaller than 2.5 micrometer
POP	Persistent Organic Material
PPMv	Part Per Million Volume
PVC	Polyvinyl chloride
RDBMS	Relational Database Management System
SCF	Scientific Committee for Food
SO_2	Sulphur dioxide
SO_3	Sulfur trioxide

Symbol	**Explanation**
SQL	Structured Query Language
STP	Standard Temperature and Pressure
VOC	Volatile organic compound

Introduction

In recent years, there has been an increasing focus on environmental aspects around the world. This is a large topic, ranging from local pollution to global climate change and health related issues. The interest in this subject is increasing probably not only because of the media displaying stories related to pollution impact but also because people, now, are taking more notice of how it affects them personally. Millions of peoples are now affected by climate change consequences such as hurricanes and long droughts and biodiversity in animals and plants are now in danger.
Fortunately awareness has been arisen and scientists succeeded to gather politicians together in conventions and agreements such as Paris 2015 climate change to legislate environmental solutions to save the earth. Norway takes the initiative to the natural life and environment protection. The emission limits and standards are one of the most severe in the world.

Grenland, Norway, is a district in the county of Telemark, in the south of Norway, and is composed of the municipalities Skien, Porsgrunn, Bamble and Siljan. The area is intensive with heavy industries since the beginning of 20's century so associated with large emissions to the environment.
Herøya Industrial Park is located on an island in Porsgrunn, and has hosted industry from the 1920s. This resulted in substantial pollution in the first decades, with little efforts made to limit or minimize the emissions. Having said that some environmental activities have been done in the 1950s, it was not until the 1970s, when 7000 people signed a paper demanding more focus on solving smog problem in the area and the environmental conditions came into focus. Over the Grenland area, it was common to have a yellow smoke lying over the area, covering the sun, and with a distinct smell. The pilots were flewing along, reported that they were using it as a sign to verify where they are!
Even though a lot of effort has been put to face these problems, in Grenland, people are still concerned that more, needs to be done. One particular example is POPs such as dioxins which are deposited on the bottom of the fjord waters. Norsk Hydro has been ordered by the environmental directorate to clean up the Telemark Fjords even though the responsible industries have been shut down several decades ago.
Also, we can notice that the collective awareness of the citizens is playing an effective role in monitoring the situation and standing up against some hasty or even corrupted political decisions. We may talk about the movement ”Nei til deponi i Brevik” that is

gathering normal citizens, experts and politicians to face a governmental decision that they think will be very harmful to the environment.

To this end, USN is collaborating with Herøya industripark, Sykehuset Telemark, Porsgrunn kommune, Folkehelseinstituttet, Miljødirektoratet and Prosjekt kompetansetjenester miljø og helse on this project, to look into what impacts these emissions may have had, by gathering historical environmental and health related data and make them available for every interested person. The proposed technical solution is to collect and systematize the data (Database), and make it possible to search for and analyze them using advanced techniques such as machine learning.

Earlier work has been made on this subject, in a collaboration between USN, Tel-Tek, Porsgrunn municipality and Telemark hospital - Department of Occupational Medicine, where systems were made to collect and present environmental data from Grenland, Norway.
In the Grenland area, Norsk Institutt for luftforskning (NILU) retrieves data from five stations. These stations measure the following parameters:

- NO_2.
- PM_{10}.
- $PM_{2.5}$.
- O_3
- CO

Four projects have been carried out so far. They are all centered on gathering and displaying some of the data from the Grenland area. Regarding the data, the projects differ by some of them displaying only historical data (with manual import of data), to displaying live data (updated from other sources or measured directly). As for the overall system, they all include a web solution and a database. One project also included a mobile app, and another developed a measurement tool to directly gather data and send over GSM. Some projects were hosted on a live server (like Azure), while other were hosted locally. The projects also differ in regard to which type of data is included, and which options the user has in the web application. These projects are:

- Development of a Database System for Environmental and Public Health Information - Master project 2016. Alexander Zhang Gjerseth, Lucille Ang. [1]
- Environmental Public Health Information Management System - Masters' Thesis 2017. Artem Chynchenko. [2]

- Information Management System for Environmental and Public Health Information - Bachelor Thesis 2017. Hans-Martin L. Kristensen, Henrik Mølmen, Joakim Johansson, Kjetil Berg Skjelbred, Thomas Prestvik. [3]
- Environmental Public Health Information Management System - Master's Thesis 2018. Ola Anton Grytten [4]

Part I

Data Management

1 Environmental Pollutants

1.1 Quick Overview

The environmental pollutants are chemicals exhausted by any type of source (industries, vehicles, nature,...) and that, under certain conditions (temperature, quantity, reaction,...), may cause damage to the environment and to human health or wildlife.
The environmental pollutants can be grouped into three main categories:

- Air
- Soil
- Water

In this chapter we will focus on the air pollutants with the most significant impact on the environment and briefly about water pollutants as well as soil.

1.2 Emission Measurements

Any data recorded, must be converted to standard condition (STP), meaning zero Celsius temperature and 1.0 atm pressure. The O_2 content should also be converted to the reference value based on the formula 1.1.

$$Ref.C_i = \frac{21 - RefO_2}{21 - \%O_2} C_i \tag{1.1}$$

Where RefO_2 can be read from tables such as Table 1.1 and % O_2 is a process value.
Before applying this, the water content must also be zero, in other words, dry condition.
In the presence of water vapor, use the following formula to convert it into dry condition. [5]

$$DryC_i = \frac{1}{1 - y_{H_2O}} * wetC_i \tag{1.2}$$

C_i is the concentration of species i

1.3 Sulphur

Perhaps the easiest pollutant to deal with is sulphur. By 1990's a significant reduction in sulphur happened due to legislation. As an air pollutant, sulphur can be seen in forms of SO_2, SO_3 and H_2S.[5]
Based on EU large combustion plants directive 2001/80/EC, table 1.1 the emission standards for sulphur varies between 5 to 2000 mg/m^3.

1.4 Nitrogen Oxides

Nitrogen oxides are harmful gases, which at high concentration may cause or be part of causing different types of diseases. The nitrogen oxides known as NOx enroll the nitric oxides NO, nitrogen dioxides NO_2 and other oxides of nitrogen. For the formation of nitrogen oxides to happen, nitrogen and oxygen need to react. However, this reaction only takes place at high temperature. That's why this reaction mostly happens in internal combustion engines, power plants and other high temperature combustion. There is a reverse relation between CO and nitrogen oxides, so optimum combustion temperature must be considered. [5]

$$\mathrm{NO} + \mathrm{O_3} \longrightarrow \mathrm{NO_2} + \mathrm{O_2} \tag{1.3}$$

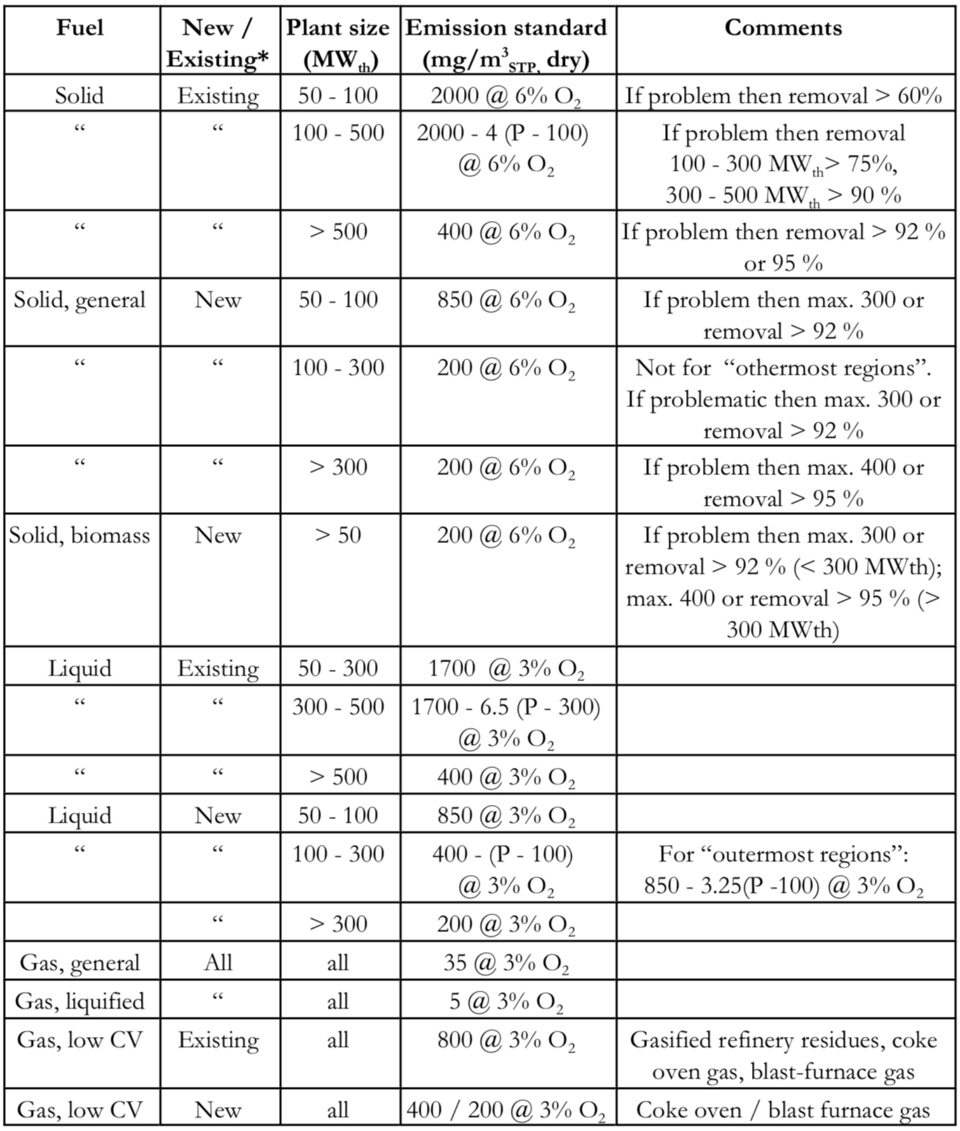
Table 1.2 displays the emission values that should not be exceeded, to avoid harm to human health or the environment.

1.5 Particulate Emissions

There are three main reasons why PM (particulate Mater) control is important.

- They can be seen by naked eye!
- PM settles down shortly, so it has more local aspects
- They have much higher concentration than any other pollutant

Table 1.1: Emission Standards Sulphur for heavy combustion plants[5]



Fuel	New / Existing*	Plant size (MW_{th})	Emission standard ($mg/m^3_{STP,}$ dry)	Comments
Solid	Existing	50 - 100	2000 @ 6% O_2	If problem then removal > 60%
"	"	100 - 500	2000 - 4 (P - 100) @ 6% O_2	If problem then removal 100 - 300 MW_{th}> 75%, 300 - 500 MW_{th} > 90 %
"	"	> 500	400 @ 6% O_2	If problem then removal > 92 % or 95 %
Solid, general	New	50 - 100	850 @ 6% O_2	If problem then max. 300 or removal > 92 %
"	"	100 - 300	200 @ 6% O_2	Not for "othermost regions". If problematic then max. 300 or removal > 92 %
"	"	> 300	200 @ 6% O_2	If problem then max. 400 or removal > 95 %
Solid, biomass	New	> 50	200 @ 6% O_2	If problem then max. 300 or removal > 92 % (< 300 MWth); max. 400 or removal > 95 % (> 300 MWth)
Liquid	Existing	50 - 300	1700 @ 3% O_2	
"	"	300 - 500	1700 - 6.5 (P - 300) @ 3% O_2	
"	"	> 500	400 @ 3% O_2	
Liquid	New	50 - 100	850 @ 3% O_2	
"	"	100 - 300	400 - (P - 100) @ 3% O_2	For "outermost regions": 850 - 3.25(P -100) @ 3% O_2
	"	> 300	200 @ 3% O_2	
Gas, general	All	all	35 @ 3% O_2	
Gas, liquified	"	all	5 @ 3% O_2	
Gas, low CV	Existing	all	800 @ 3% O_2	Gasified refinery residues, coke oven gas, blast-furnace gas
Gas, low CV	New	all	400 / 200 @ 3% O_2	Coke oven / blast furnace gas

*: Existing = plant existing on Nov. 27, 2002 ; or license for new plant requested before that date and plant entering operation before Nov. 27, 2003

Table 1.2: Assessment Nitrogen limit values[6]

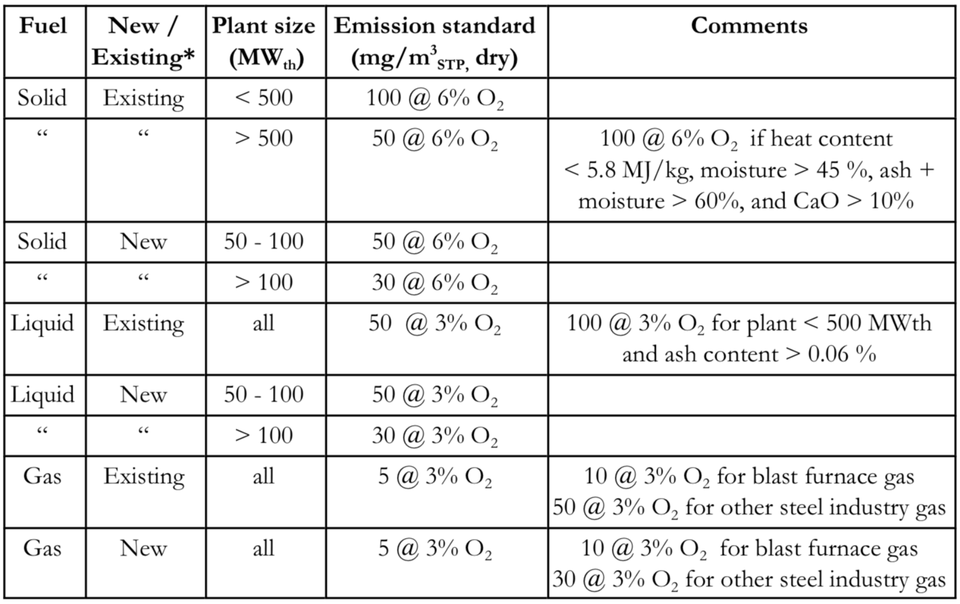
$\mu g/m^3$	Protection of human health	Protection of vegetation and nature
Upper limit	32	24
Lower limit	26	19.5

Fortunately the emission control is relatively easy and since 1920's it has been controlled. Coal-fired power plants, powder plants, diesel engines and wood burning stoves are the main sources of this pollutant. The particulate emission standards depend on several factors such as:

- Size of the industry
- Type of fuel
- New or existing plants
- Oxygen reference value
- The flue gas must be considered dry

Based on EU large combustion plants directive 2001/80/EC, the emission standards vary between 5 to 100 m g/m^3 for 3 to 6 % of reference O_2, shown in table 1.3. This is just an example and for other industries one can search through other directives.[5]

Table 1.3: PM standards for large combustion plant in EU except new gas turbine plants[5]

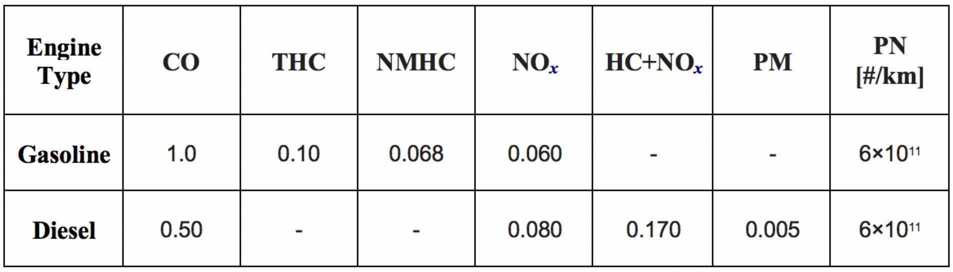


Fuel	New / Existing*	Plant size (MW_{th})	Emission standard (mg/m^3_{STP}, dry)	Comments
Solid	Existing	< 500	100 @ 6% O_2	
"	"	> 500	50 @ 6% O_2	100 @ 6% O_2 if heat content < 5.8 MJ/kg, moisture > 45 %, ash + moisture > 60%, and CaO > 10%
Solid	New	50 - 100	50 @ 6% O_2	
"	"	> 100	30 @ 6% O_2	
Liquid	Existing	all	50 @ 3% O_2	100 @ 3% O_2 for plant < 500 MWth and ash content > 0.06 %
Liquid	New	50 - 100	50 @ 3% O_2	
"	"	> 100	30 @ 3% O_2	
Gas	Existing	all	5 @ 3% O_2	10 @ 3% O_2 for blast furnace gas 50 @ 3% O_2 for other steel industry gas
Gas	New	all	5 @ 3% O_2	10 @ 3% O_2 for blast furnace gas 30 @ 3% O_2 for other steel industry gas

* : Existing = plant existing on Nov. 27, 2002 ; or license for new plant requested before that date and plant entering operation before Nov. 27, 2003

The ambient air quality standard (Annual mean concentration) for Norway is 20 $\mu g/m^3$ for PM_{10} and 8 $\mu g/m^3$ for $PM_{2.5}$ and 40 $\mu g/m^3$ for NO_2 [7]
Emission standards for cars based on Euro 6 Standard: (g/Km)

Table 1.4: Cars Emission Standards[8]



Engine Type	CO	THC	NMHC	NO_x	HC+NO_x	PM	PN [#/km]
Gasoline	1.0	0.10	0.068	0.060	-	-	6×10^{11}
Diesel	0.50	-	-	0.080	0.170	0.005	6×10^{11}

In Table 1.4, PN represents particle number.

1.6 VOCs,PAHs,CO

1.6.1 VOCs

VOCs refers to ”Volatile Organic Compounds” and is a generic name for every anthropogenic nature organic compound, except methane, that can react with other chemicals like nitrogen in presence of sunlight. [5] VOCs are varied and complex and they may have harmful effect on human health and on the environment. However, studying these effects is quite difficult because VOCs are not very toxic and show their effects over a long period.

Table 1.5: Emission of VOCs for the oil industry for 2014 in Norway by main sources in tonnes [9]

Main sources	VOC	%
Dry compressor seals	3600	19%
Vent header	3300	17%
Produced water treatment	2900	15%
HC purge and blanket gas	2400	12%
Gas leaks/fugitives	2200	11%
Flare gas not burnt	2100	11%
Glycol regeneration	1550	8%
Compressor wet seals	1200	6%
Other sources	550	3%
Total	19800	100%

Table 1.5 is showing the main sources of the emission of VOCs in Norway for the oil industry, which is very important. Hence, monitoring the VOCs emissions is necessary, especially because it’s not only a regional problem but can also spread.

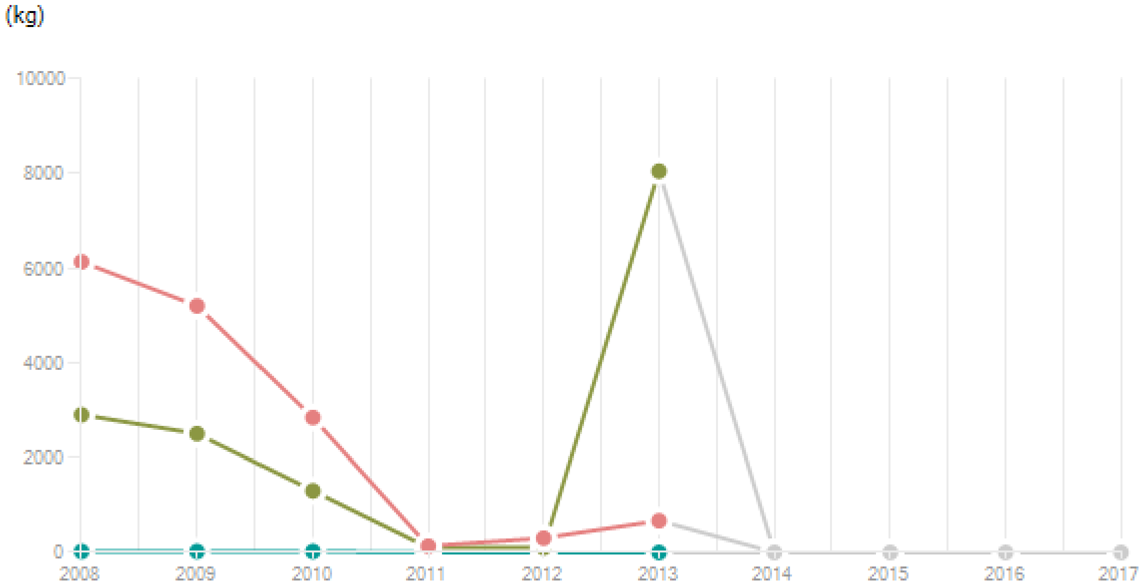


Figure 1.1: Emissions of PAHs in Norway(in kg per year)[11]

1.6.2 PAHs

The polycyclic aromatic hydrocarbons form a particular class of organic compounds that can be studied because of its carcinogenic character. The PAHs are relatively stable molecules formed by carbon and hydrogen atoms. These atoms are organized in aromatic cycles. Considering their stability in the environment and their toxicity, many of them have been declared as major pollutants by the EU Scientific Committee for Food (SCF), the European Union (EU), and US Environmental Protection Agency (EPA).[10]

In figure 1.1, the green line represents the emission to air, the blue line the emission to water and the pink line is the emission to soil.

Figure 1.1 shows the PAHs emissions to air, water and soil in Norway from 2008 to 2013 in kg.
We can notice that for soil emissions, the values are always zero. This is due to the fact that data is not available for soil pollution. This lack of data for emission to soil represents a major problem for monitoring soil pollution everywhere in the world.
Also, we see that that the PAHs emissions to air and water have been drastically reduced. For the year 2013 we observe that the emissions to air abnormally increased. The reason is unknown but it is most likely due to some errors.

1.6.3 CO

Carbon monoxide, ”CO”, is a toxic and environmental pollutant gas. It is produced during incomplete combustion of carbon, when it is burnt with limited oxygen.
The oxidation of the carbon monoxide occurs, following the reaction in eq. 1.4 [5]

$$CO + OH \rightleftharpoons CO_2 + H_2 \tag{1.4}$$

Carbon monoxide may be exhausted by several type of sources[12]:

- Power stations.
- Waste incinerators.
- Petrol engines.

However, the petrol engines, like cars, are considered as the major source of carbon monoxide. Fortunately, the emissions due to these sources have been reduced significantly thanks to catalytic converters.
As any other gas pollutant, CO has an impact on the environment and human health. Ground-level ozone can be produced while CO reacts with other pollutants, leading to damage to peoples health and buildings. Also, inhaling it causes harm to the heart, blood and brain as well as to unborn children.

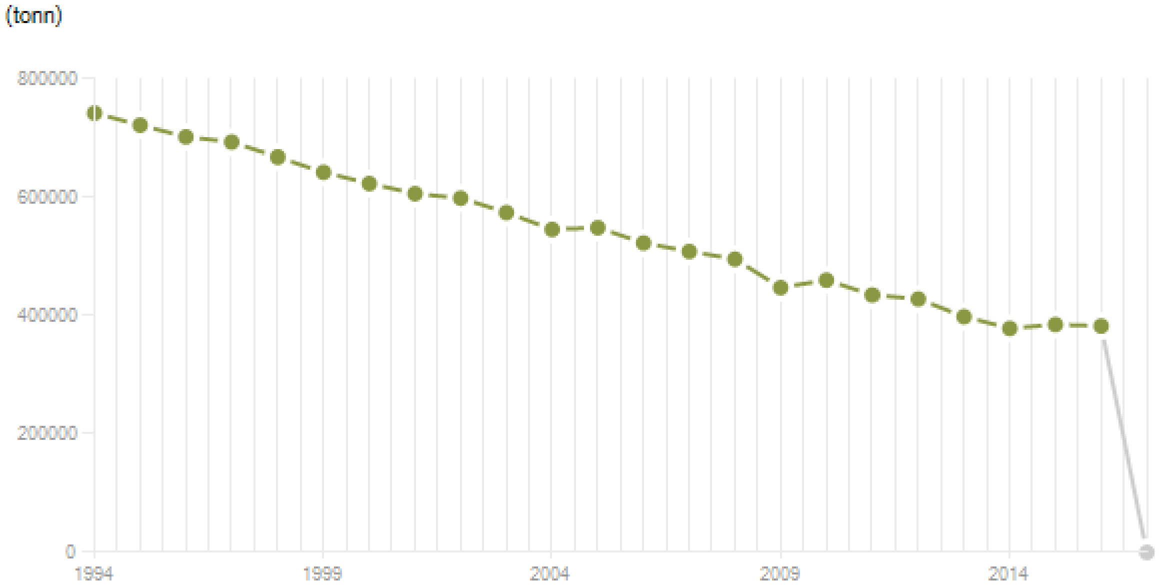


Figure 1.2: Emissions of carbon monoxide in Norway from 1994 to 2016 (in tons per year)[12]

Figure 1.2 shows the CO emissions tendency in Norway, to air, from 1994 to 2016. We note that the emissions are being reduced gradually.

1.7 Halogens, Dioxins/Furans

Halogens are F, Cl, Br, I. Dioxins and Furans are cyclic hydrocarbon compounds of halogens. HF has widely usage in glass industry. Cl is mainly used as NaCl in food salt and in plastic industry such as *PVC*. Dioxins are discussed in the following sections. [5]

Table 1.6: Halogens, Dioxins/Furans, Emission Standards, MSW incinerator, EU 2000[5]

$mg/m^3_{STP}@11\%O_2, Dry$	MSW (Municipal solid waste)	Comment
HCl	10	
HF	1	
PCDD	0.1	Dioxin $ng/m^3_{STP}TEQ$

1.8 Dioxins

Dioxins are chloro-organic aromatic chemicals that consist of PCDD (poly-chlorinated di-benzo dioxins) and PCDF (poly-chlorinated di-benzo furans) compounds. They are also a group of POPs. Persistent organic pollutants (POPs) are organic materials which slowly degrade in the environment.
PCDD/Fs are formed as unwanted byproducts of industrial and combustion processes. Within incinerators dioxins form in different mechanisms.[5] For example:

- Combustion waste incineration and backyard burning
- Metals melting and refining.
- Chemical manufacturing unwanted byproduct.
- Biological and photo-chemical processes.

Ambient Air Quality Standard for Dioxins based on Stockholm convection is $0.6pg/Nm^3$ or $0.610^{-12}g/Nm^3$ [13]

1.9 Trace Elements

By definition any chemical element at concentration less than 0.1 %W is a trace element. The main sources are wastes and waste-derived fuels. The european union has concerns about 13 elements and there are regulations about their emission in cement plants and waste incinerators. These elements are: $As, Cd, Co, Cr, Cu, Hg, Mn, Ni, Pb, Sn, Ti, V$
Among these, Hg, Br, Cl, F are volatile, meaning that during the combustion, they will appear in the flue gas rather then the ash, so they are air pollutants.[5]

Table 1.7: Trace Elements, Emission Standards, MSW incinerator, EU 2000[5]

$mg/m^3_{STP}@11\%O_2, Dry$	MSW (Municipal solid waste)	Comment
Hg	0.05	
Cd+Ti	0.05	
As+Co+Cr+Cu+Mn+Ni+Pb+Sb+Sn+V	0.5	

1.10 Greenhouse Gases and Ozone-depleting gases

The greenhouse gases absorb and emits the sun radiant energy. Without it, the earth's average temperature would be more or less -18°C. However, since the industrial revolution, the concentration of the greenhouse gases in the atmosphere increased largely leading to to higher temperature of the earth. [14]
The most famous greenhouse gases are:

- Water vapor, H_2O.
- Carbon dioxide, CO_2.
- Methane, CH_4.
- Nitrous oxide, N_2O.
- Ozone, O_3.
- Chlorofluorocarbons, CFCs.
- Hydrofluorocarbons, HCFCs.

Moreover, the ozone depleting substances are anthropogenic gases that may destroy the ozone layer causing damage to human health and the environment, since this layer protects the earth from the sun's ultra violet radiation.[15]
The major sources for these substances are:

- Chlorofluorocarbons, CFCs.
- Hydrochlorofluorocarbons, HCFCs.
- Hydrobromoflurocarbons, HBFCs.
- Halons.
- Methyl bromide.
- Carbon tetrachloride.
- Methyl chloroform.

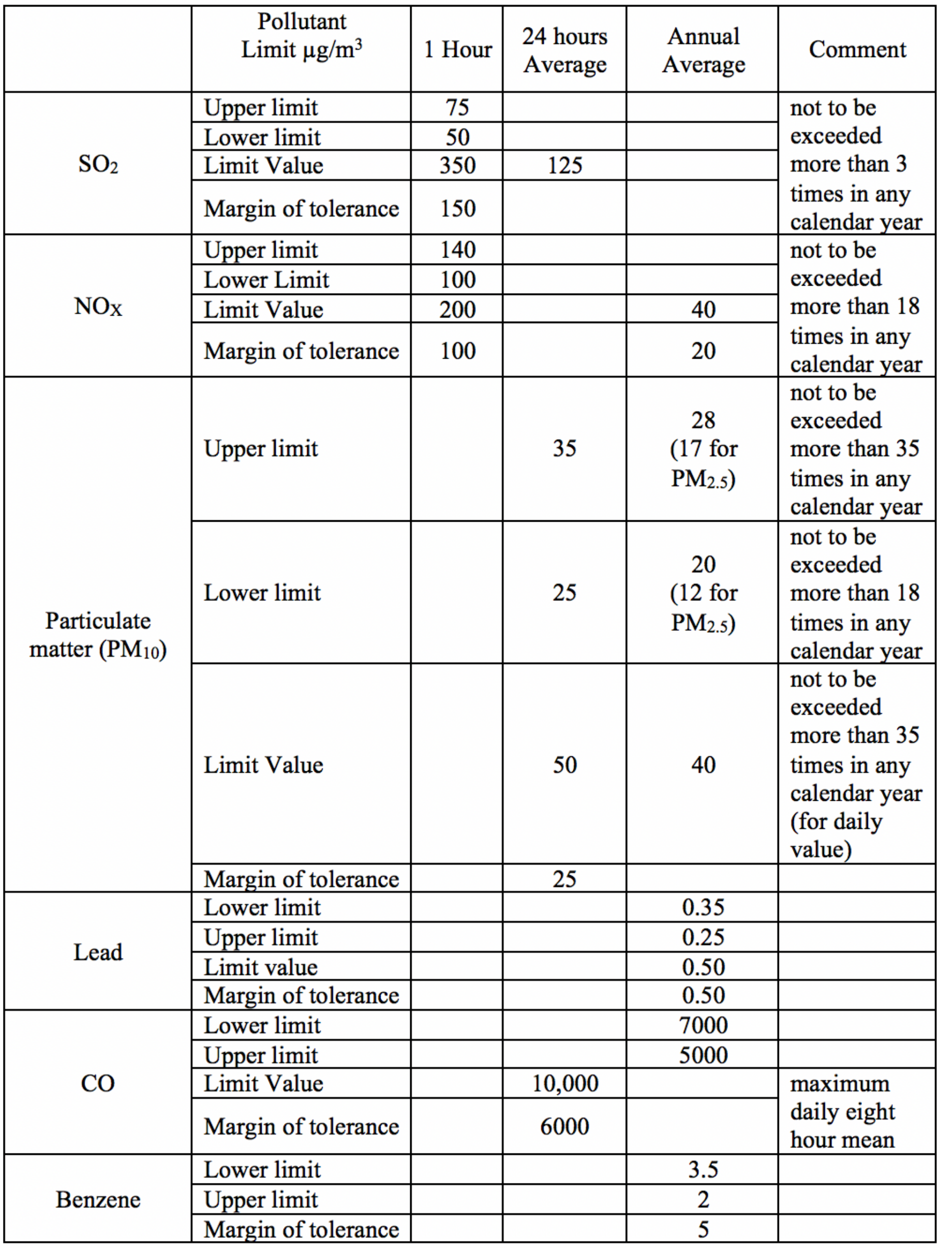
1.11 Air quality Standards

This section is used as the basis of our data analysis. The reference for regulation is Directive 2008/50/CE of the European Parliament and of the conceal of 21 may 2008 on ambient air quality and cleaner air for Europe. The measuring intervals are average per:

- Annual
- 24 Hours
- Maximum daily 8 hrs mean value: Starting from 17:00, splitting the day into three 8 hrs intervals. The largest mean value will be used.
- Number of occurrence

Figure 1.3 describes the action plan for three groups of measured conditions based on table 1.8 .

Table 1.8: Air Quality Regulations[6]



	Pollutant Limit µg/m^3	1 Hour	24 hours Average	Annual Average	Comment
SO_2	Upper limit	75			not to be exceeded more than 3 times in any calendar year
	Lower limit	50			
	Limit Value	350	125		
	Margin of tolerance	150			
NO_X	Upper limit	140			not to be exceeded more than 18 times in any calendar year
	Lower Limit	100			
	Limit Value	200		40	
	Margin of tolerance	100		20	
Particulate matter (PM_{10})	Upper limit		35	28 (17 for $PM_{2.5}$)	not to be exceeded more than 35 times in any calendar year
	Lower limit		25	20 (12 for $PM_{2.5}$)	not to be exceeded more than 18 times in any calendar year
	Limit Value		50	40	not to be exceeded more than 35 times in any calendar year (for daily value)
	Margin of tolerance		25		
Lead	Lower limit			0.35	
	Upper limit			0.25	
	Limit value			0.50	
	Margin of tolerance			0.50	
CO	Lower limit			7000	
	Upper limit			5000	
	Limit Value		10,000		maximum daily eight hour mean
	Margin of tolerance		6000		
Benzene	Lower limit			3.5	
	Upper limit			2	
	Margin of tolerance			5	

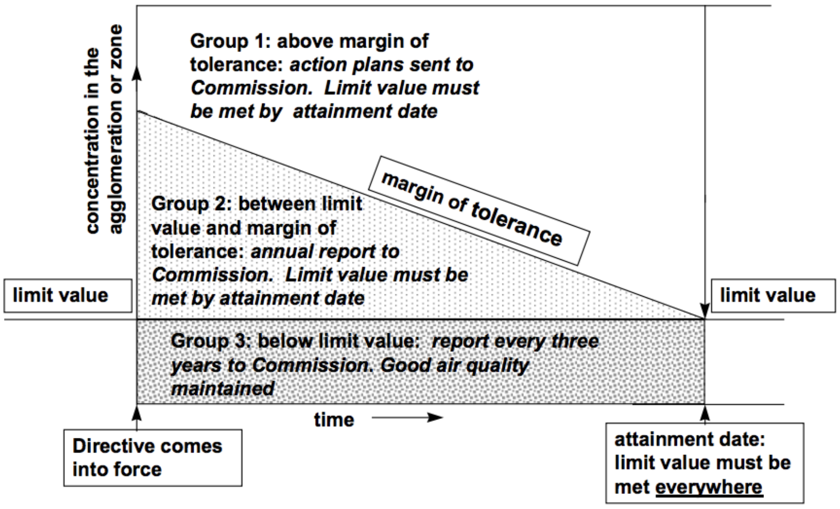

concentration in the agglomeration or zone
Group 1: above margin of tolerance: *action plans sent to Commission. Limit value must be met by attainment date*
margin of tolerance
Group 2: between limit value and margin of tolerance: *annual report to Commission. Limit value must be met by attainment date*
limit value
limit value
Group 3: below limit value: *report every three years to Commission. Good air quality maintained*
time
Directive comes into force
attainment date: limit value must be met everywhere

Figure 1.3: Action plan and limit value plus the margin of tolerance [16]

1.12 Water pollutants

Based on EU water directive 2000/60/EC when studying water, one has to narrow the topic to surface waters, ground waters, inland waters, rivers, lakes and so on. As our study has a local target, we focus on fjords and lakes in particular.

The pollutants must be measured and monitored based on relevant directives. One of these main pollutants is persistent hydrocarbons and persistent and bio-accumulated organic toxic substances such as Dioxins and furans, discussed in section 1.8. One of the most concerning Grenland regions is Herøya (Gunneklevfjorden) where industrial activities in 1970’s produced a lot of Dioxin deposits at the bottom of the lake. They are persistent and have entered the food cycles and are being carried around.

1.13 Soil Pollution

Pollution is not just restricted to what is emitted to air or dropped in water. A lot of people seem to forget that pollution exists in soil, becasue it’s not visible; unfortunately it’s there and may be very dangerous. That’s why the last FAO report is titled ”Soil pollution: a **hidden** reality.”

By soil pollution, we mean that a harmful substance is found, at very abnormal concentration, where it is not supposed to be. [17]

In Norway, the sources of soil pollution are mainly related to industries, mining or closed contaminated landfills. The Norwegian Environment Agency declared that more than 5000 sites are contaminated and that actions have been taken for more than 2200 of them. However, in May 2017, they also declared that 441 sites are heavily contaminated. [18]

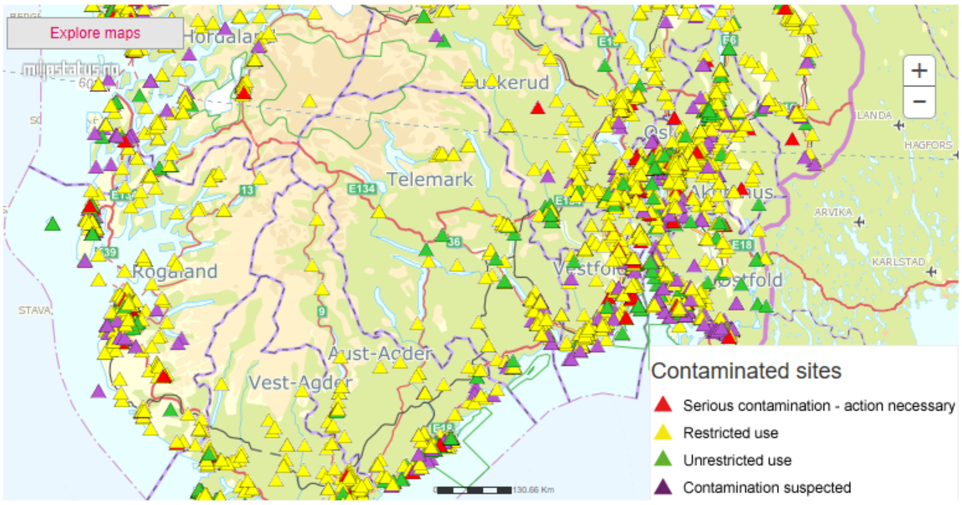
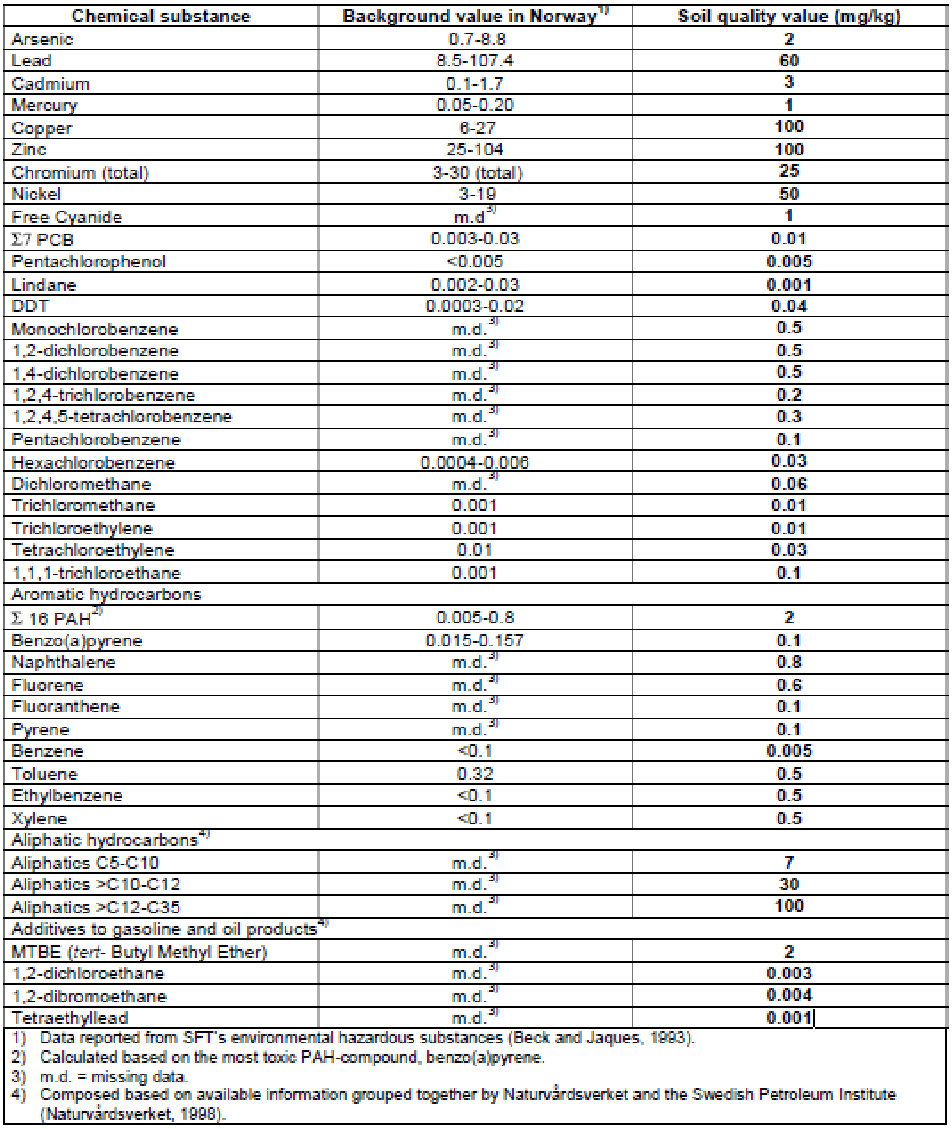


Figure 1.4: Contaminated sites next to Oslo[18]

We can notice from figure 1.4 that Telemark is one of the least contaminated regions even though it’s known for its historical active industries.
The main soil pollutants can be divided in: [17]

- Heavy metals and metalloids.
- Nitrogen and phosphorus.
- Pesticides.
- PAHs.
- POPs, (persistent and accumulate throught food chain).
- Radionuclides, (due to nuclear pollution).
- Emerging microorganisms, (chemicals that recently appeared).
- Pathogenic microorganisms, (organisms with chemical defense).
- Antimicrobial resistant bacteria and genes, (bacteria resistant to antibiotics).

Table 1.9: Soil quality guideline [19]



Chemical substance	Background value in Norway[1]	Soil quality value (mg/kg)
Arsenic	0.7-8.8	**2**
Lead	8.5-107.4	**60**
Cadmium	0.1-1.7	**3**
Mercury	0.05-0.20	**1**
Copper	6-27	**100**
Zinc	25-104	**100**
Chromium (total)	3-30 (total)	**25**
Nickel	3-19	**50**
Free Cyanide	m.d[3]	**1**
Σ7 PCB	0.003-0.03	**0.01**
Pentachlorophenol	<0.005	**0.005**
Lindane	0.002-0.03	**0.001**
DDT	0.0003-0.02	**0.04**
Monochlorobenzene	m.d.[3]	**0.5**
1,2-dichlorobenzene	m.d.[3]	**0.5**
1,4-dichlorobenzene	m.d.[3]	**0.5**
1,2,4-trichlorobenzene	m.d.[3]	**0.2**
1,2,4,5-tetrachlorobenzene	m.d.[3]	**0.3**
Pentachlorobenzene	m.d.[3]	**0.1**
Hexachlorobenzene	0.0004-0.006	**0.03**
Dichloromethane	m.d.[3]	**0.06**
Trichloromethane	0.001	**0.01**
Trichloroethylene	0.001	**0.01**
Tetrachloroethylene	0.01	**0.03**
1,1,1-trichloroethane	0.001	**0.1**
Aromatic hydrocarbons		
Σ 16 PAH[2]	0.005-0.8	**2**
Benzo(a)pyrene	0.015-0.157	**0.1**
Naphthalene	m.d.[3]	**0.8**
Fluorene	m.d.[3]	**0.6**
Fluoranthene	m.d.[3]	**0.1**
Pyrene	m.d.[3]	**0.1**
Benzene	<0.1	**0.005**
Toluene	0.32	**0.5**
Ethylbenzene	<0.1	**0.5**
Xylene	<0.1	**0.5**
Aliphatic hydrocarbons[4]		
Aliphatics C5-C10	m.d.[3]	**7**
Aliphatics >C10-C12	m.d.[3]	**30**
Aliphatics >C12-C35	m.d.[3]	**100**
Additives to gasoline and oil products[4]		
MTBE (*tert*- Butyl Methyl Ether)	m.d.[3]	**2**
1,2-dichloroethane	m.d.[3]	**0.003**
1,2-dibromoethane	m.d.[3]	**0.004**
Tetraethyllead	m.d.[3]	**0.001**

1) Data reported from SFT's environmental hazardous substances (Beck and Jaques, 1993).
2) Calculated based on the most toxic PAH-compound, benzo(a)pyrene.
3) m.d. = missing data.
4) Composed based on available information grouped together by Naturvårdsverket and the Swedish Petroleum Institute (Naturvårdsverket, 1998).

Table 1.9 shows the most important soil contaminants, their concentration and the values that are considered to be the best for the soil quality.

This table is more an indicator of the soil quality requirements than a reference since it is from 1999 and that, as stated above, soil pollutant data is very rare.

1.14 Environmental Management System

EMS is a tool or a systematic way of approach for management to look for the most sustainable solution for the organization. By reducing the environmental impacts, the organization can become more competitive and can better handle future environmental regulations. Currently there are more than 200 legal acts and more than 500 directives enforced at the EU level. This is one of the most dense and complex policy areas, so many organizations prefer to have web-based solutions named EDMS (Environmental Data Management system)
There are two main guidelines for EMS:

- ISO 14001:2015: This guideline is presented by International Organization for Standardization for companies and organizations.
- EMAS: EU Eco-Management and Audit Scheme is also another management instrument developed by the EU commission.

These guidelines will help the organization to gain: [20]

- Control of the environmental impact of the organization
- Reduction of total cost in a long term
- An easier control to limit the environmental impact
- Better control over use of raw material and energy
- Less waste of materials
- Less wastes and dangerous by-products
- Efficient environmental investments

2 Data Collection

2.1 Choices

As mentioned in the first chapter, it was decided to mainly concentrate, for this work, on air pollution. The report goes briefly through many different types of air pollutants, whereas in this solution (website), a special interest was given to some pollutants. This choice has been made accordingly to what ”NILU” is monitoring and considering to be the most harmful pollutants like:

- NO_x
- O_3
- PM_{10}
- $PM_{2.5}$
- SO_x

However, other types of data, like studies and reports, can be about any subject; pollution (air, water, soil ...) or/and health related.

2.2 Data sources

The data used in this project are collected from different sources. It has either been found or given by external partners.

The found data are data found by the students on trusted websites, books, online libraries, etc.
Some of these sources are:

- NILU (Norsk institutt for luftforskning).
- SSB (Statistisk Sentralbyrå).
- NIVA (Norsk Institutt for Vannforskning).
- Miljødirektoratet.

Also, some data has been given by professionals working in related industries or agencies. A non-exhaustive list of contacts has been prepared, but for privacy concerns, it is kept confidential.
An important part of 'Big data' is to clean and organize the collected data. Actually, "washing" the data is considered to represent 80% of the job.[21]
Thus, as soon as data was received or collected, it was organized by type (numeric, document), impact (environment, health), pollutants (PM, NO_x, ...).

There were some problems gathering data. Some messages have not gotten an reply. The question to ask is why this happened. There could be several issues with this, discussed in chapter 11.

2.3 Analysis

Data refining and numeric data analysis are an important part of every environmental management system. Figure 2.1 shows 5 years trend of NO_2 pollutant for five stations located in Grenland area. Although the pollutant value is far below the limits, Lensmannsdalen station located between Skien and Porsgrunn has the worst condition. Moreover, the worst air condition is during Autumn, specifically November.

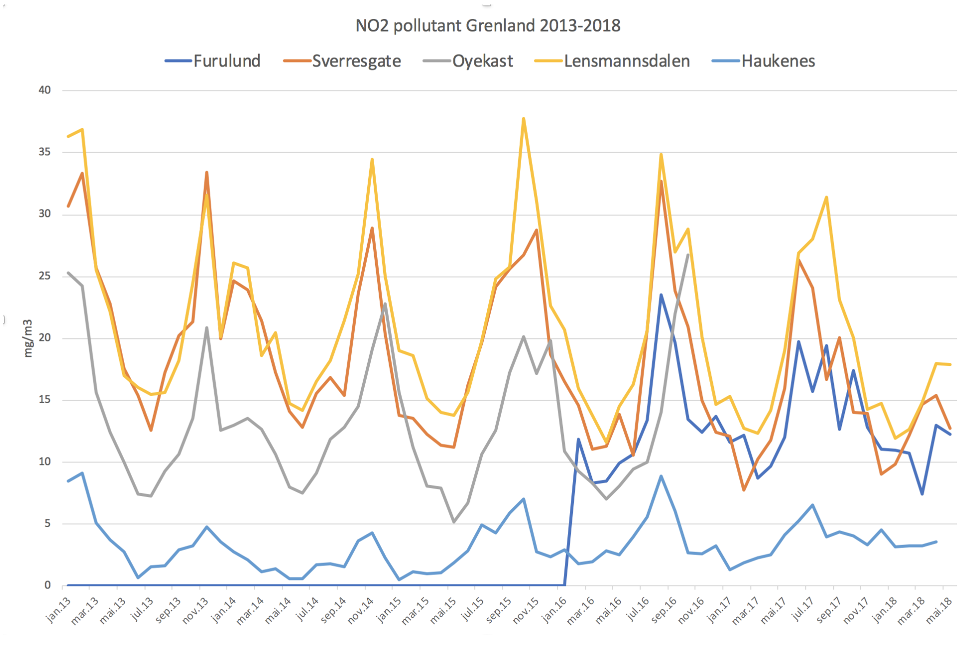


Figure 2.1: NO_2 pollutant trend. [22]

Part II

System

3 Risk Management

3.1 Risk Identification

To identify the risk associated with the project, a brainstorm meeting was performed, and the whole project was divided into different categories:

1. Data Collection
2. Data Analysis
3. Data
4. Website
5. System
6. Project Management

The project consist of different sections, each with different risks. Table 3.2 was used as a guideline to develop risk evaluation presented in table 3.1 as explained in section 3.2. The probability was graded from 1, which is ”unlikely”, to 5, which is ”very likely”. The impact of the risk is graded from 1, which is minor impact, to 5 which is extreme.

Table 3.1: Project Related Risk

#	Risk Item	P (1-5)	I (1-5)	Priority
1	**Data Collection**			
1.1	No response from contacts	5	3	15
1.2	Misleading Data	2	3	6
1.3	Not credible sources (eg: internet)	4	1	4
1.4	Conditions for data (constraints)	4	4	16
1.5	Data get lost (post, ...)	2	2	4
2	**Data Analysis**			
2.1	Wrong Data	2	4	8

Continued on next page

Table 3.1 – *Continued from previous page*

#	Risk Item	P (1-5)	I (1-5)	Priority
2.2	Wrong units and scales for numeric data	2	4	8
2.3	Using wrong formulas	1	4	4
3	**Database**			
3.1	Crashed database	2	5	10
3.2	Hacked	1	5	5
3.3	Dataloss in database	3	3	9
3.4	Data ownership	1	5	5
3.5	Database back-up crash	1	5	5
3.6	Not enough knowledge of proper database tools	3	2	6
3.7	Bad coding for database	2	3	6
3.8	Database size choosing is not correct	1	1	1
3.9	Database is not expandable	1	3	3
4	**Website**			
4.1	Website maybe Hacked	1	5	5
4.2	Website may crash	2	3	6
4.3	Website bandwidth shortage	1	3	3
4.4	Loose connection between the website and database	4	3	12
4.5	Using cookies permission	1	2	2
5	**System**			
5.1	Lack of server storage	1	4	4
5.2	System ownership	1	5	5
5.3	Lacking of functionality of the system	4	2	8
6	**Project**			
6.1	Signature of final project description	5	4	20
6.2	Team member quit	2	3	6
6.3	Overleaf crash	1	5	5
6.4	Supervisor quit	1	3	6
6.5	Customer terminate contract	1	3	3
6.6	Oral presentation booking room	1	4	4

$$Priority = Probability \times Impact \tag{3.1}$$

3.2 Risk Evaluation

After The risks were identified, they must be quantified by the probability factor multiplied by the impact or severity equation (3.1). The combination of these two will define the value or priority of the risk:
In order to prioritize the risks, we implement 80-20 rule to focus on the 20 percent which has the highest priorities.

Table 3.2: Risk Matrix[23]

Probability ↑	**Very Likely**	Medium	High	Extreme
	Likely	Low	Medium	High
	Unlikely	Low	Low	Medium
		Minor	**Moderate**	**Major**

Impact →

3.3 Risk Treatment

Based on previous section and Table 3.1, the top 20% risks with high priorities are listed below:

1. Signature of project description, 6.1: We avoided this risk by talking to the project supervisor and he managed the issue.
2. Conditions for data (constraints), 1.4: This risk eliminated because we used the publicly published data without any confidentiality.
3. No response from contacts, 1.1: Using dummy data instead of the real data caused the risk to be migrated by reducing the advert effects. Being Perseverance and by following up, we could make valuable connections and eliminated the risk.
4. Loose connection with database in the website, 4.4: Technically by using the proper tools, the risk is reduced to this value and it is accepted.
5. Database Crash, 3.1: The risk is reduced by using HADOOP
6. Data loss in the database, 3.3: By using back up feature, this risk is avoided.

There are 6 major risks to manage. There are several solutions, first try to avoid it, then try to transfer it, attempt to migrate it and potentially accept the risk.[23]

4 Machine Learning

4.1 Intro

Data analysis and Machine Learning has become very popular in recent years, and it should be possible to apply some of these techniques. As opposed to more specified data analysis, one of the interesting aspects of machine learning is that you can apply a more generic algorithm to a data set, and find relations that you did not even know you looked for. Typically, you want to apply these techniques to be able to make predictions about the future or to extract unknown information and connections.
Machine Learning algorithms be divided into three classes: Supervised Learning, Unsupervised Learning and Optimization. Unlike Unsupervised Learning, in Supervised Learning the corresponding outputs are known, and the system can be trained to fit the outputs. In Unsupervised Learning, the goal is to locate unknown patters and clusters. In Optimization, the goal is to find the optimal set of parameters which minimize a predefined cost function. This cost-function is also found when training the machine learning applications. [24] With Unsupervised Learning there are no known outputs, so the goal is to extract hidden information and patterns, or clusters, in the data.

4.2 Machine Learning Methods

4.2.1 Regression

Regression methods have been known for a long time, and can be used to predict the output by giving correct weights to different inputs. It can be used in many ways, from finding more simple linear relationships to classifying inputs into different classes. One very simple form is linear regression, where the error is minimized using least squared error. This can then be used to predict outputs from new inputs.[24]

4.2.2 Classification and Regression Trees

Classification and Regression Trees can be thought of as a line of nodes with different questions, and depending on the input (answer) you are either sent this or the other way. (Is it cloudy? Yes - bring an umbrella). They are fast, and can be applied to a broad range of problems without much data cleaning beforehand.[24]

4.2.3 Naive Bayes

Naive Bayes is a technique that calculate two probabilities from your data: The probability of each class, and the conditional probability for each class given each x value. Using the Bayes Theorem, you can then make predictions for new data using this, assuming that each input variable is independent. (This is not always true for real data, hence ”Naive”.) [25]

4.2.4 Principal Component Analysis

PCA is about finding the underlying structure of a data set, and how to effectively represent this data in a compressed format. PCA can be thought of as reducing the dimensionality of the data to compress it, while maintaining its structure and usefulness. This is unsupervised learning, as it has no known outputs.

The goal is to find the basis vectors, called principal components. With these basis vectors, the data can be reduced but much of the complexity as possible is still stored. The significant principal components are identified by how much of the data's variance they capture, and then order them by that metric.[24]

4.2.5 Neural Networks

Neural Networks can be thought of as an attempt to copy how the brain processes information. They contains a set of connected neurons, or nodes, which again can be divided into an input layer, a hidden layer and an output layer. The number of neurons wary highly depending on the application, and there can also be several layers of neurons within the hidden layer. Neural networks can be used for supervised and unsupervised learning, but often the goal is to learn a function to map the input to the output.

A simple example is using a neural network to estimate which digit that is written in an image. The input layer could then be all the pixels in the image. The hidden layer could consist maybe of two or three layers with the same numbers of neurons, and the output layer would have ten neurons: one for each digit (or more general: label). After training the network, the output neurons would present a value, indicating the network's probability that the image consists of the corresponding digit, with the probabilities ranging from 0 to 1 in each output.

Neural networks are really good at deep learning, particularly in situations where the data is complex. They are known as universal function approximators, because they are able to learn any function, with just one single hidden layer.

In the neural network, each node has an activation function and a bias that the input value is processed through before the value is passed further. In order to find nonlinear connections in the data, these functions has to be nonlinear. The network is then trained using gradient descent, in an iterative process to find the parameters in the activation function, along with the bias, that minimizes a loss function. This, in short, is deep learning.[24]

4.2.6 Support Vector Machines

Support Vector Machines are one of the most popular Machine Learning algorithms, and is a powerful classifier that can be used on a wide range of problems. Examples of problems SVMs can solve are:

- Is this an image of a cat or a dog?
- Is this review positive or negative?
- Is this document related to health or environment?

If you have a data set, the goal is to create a divider, called a hyperplane, that separates the two classes in the best possible way. With a 2D space, this can be thought of as a line that divides all the points as best it can into their correct class. In 3D, the divider can be thought of as a plane, but as it moves into higher dimensions this becomes more abstract.

This is an optimization problem, where the goal is to maximize the margin on each side of the divider, while constraining the divider to separate the classes. It is not always possible to separate the data cleanly, and there are two ways of dealing with this problem. The first is to soften the definition of separate, so that some points fall on the wrong side. The other solution is to throw the data into higher dimensions, where nonlinear classifiers can be created. Then the dividers can be brought back to the lower dimension. Here they will often look more ”random”.[24]

4.2.7 Bootstrap Aggregation

Bootstrap Aggregation is a statistical method for estimating quantity from a data sample. In stead of averaging the whole sample to find the mean, you take lots of smaller samples and find the mean in each, and then average from there. This can often lead to a better estimation of the true value. In Random Forest, the same approach is used, but with decision trees, and then the average is found from all the models.[25]

4.2.8 Case-Based Reasoning

CBR can be used both for classification and regression. The concept is to compare cases, and when given a new case, find the case that resembles that one the most, and present the solution (s) to that case, and in the process updating the cases based on the feedback.[26] This requires that the data can be modelled so that a distance metric can be measured. For example, in a case with two parameters, the distance between two cases is calculated using the Euclidean distance, which in older literature is referred to the Pythagorean metric.

Case-based reasoning consists of a cycle of the following steps[27]:

- Retrieve: Given a new case, retrieve similar cases from case base.
- Reuse: Adapt the retrieved case to fit new case.
- Revise: Evaluate solution, revise on how well it works.
- Retain: Decide whether to retain this new case in case base.

4.2.9 Genetic Algorithm

Optimization is the technique of finding an optimal set of parameters that typically minimize (i.e. cost) or maximize (i.e. production) a function.

The genetic algorithm is based on principles of natural evolution and Darwian Science. First, a parent generation is randomly created. This generation then combines and produces ”offspring”, whose features are a mixture of the parent generation’s, as well as some ”mutation”, which is to replace a random part of the formula with another randomly selected part. The next generation then have characteristics inherited from their parents, as well as a random part.

After reproduction, the offspring is tested for their ”fitness”: how they perform according to a task’s end criteria. The iterations usually stop after a certain number of iterations, or when a stop criteria is met. The generation with the highest performance is presented as the model result. [28]

4.3 Applications in this Project

It can be reasonable to divide the Machine Learning applications in this project into two parts: machine learning applied to the data itself, and machine learning applied to the meta-data, used to increase the user experience of the site.
Machine learning applied to the data could extract more information, and create extra value for scientists and others, who can be assumed to be more interested in specific data. Typically applications for this would be predictions and multivariate analysis to identify the important and non-important parameters. Case-based reasoning can also be used for this.
Machine Learning can also be used to boost the user experience. This could mean to apply methods that gives the user inputs and suggestions that they themselves may not would have thought of. This could be to present popular topics, or to present the user with data that may be of special interest to the user. It could also be to present some of the data analysis on the front page to attract interest. Case-based reasoning can be thought of as a well suited application for a chat-bot or similar: that comes up with suggestions and answers for the users. One benefit of CBR is that it does not need extensive training in order to work.
In this project, both Supervised and Unsupervised learning can be applied. Supervised Learning would be those applications where historical data is used to predict future outputs. An example of this can be to present suggestions to the user, using a neural network for predictions. A downside of this would be that in the beginning, there would be no historical data to draw information from, so the suggestions would probably be not that good. This can be solved by turning off suggestions in the beginning, until the model has started behaving somewhat well. Another example is to present pollution predictions, drawing from historical data.
Unsupervised Learning would more typically be applied to the data sets, using PCA, where the goal of the analysis can is unknown. To be able to perform Machine Learning on the (multiple) data sets, they need to be in a form where data analysis is possible, in well structured data sets. That means that .pdf documents and similar are unwanted for this application.

There are several applications that can be integrated in this solution:

- Make recommendations to the user based on previous searches and/or popular items.
- Present latest/most popular topics/inputs that the user may like.

- Present charts/plots with historical data and short-time predictions, and display on front pages. Typically pollution, but also diseases/health issues?
- Make predictions on data sets. Both from user input and automatically on chosen data sets to present on front page.
- Help complete text inputs to search and meta-data when entering (auto-correct).
- Multivariate data analysis on data sets. Find and present the contributions from each parameter on the output.
- Use Machine Learning and data analysis to extract information that is not ”looked for”: are there unknown connections between parameters, possibly between parameters in different data sets, i.e. looking for links between diseases and pollution’s. Are there links between number of cars, population, temperature, emission levels, treatments etc. and diseases?
- Group users by search history,input history and behaviour, possibly suggest contacts.
- Group data sets with the help of Machine Learning; find unknown groups.
- Find popular/non-popular items/pages by counting how long users spend at each place.
- Identify bottle-necks: Find out if users often type in the same word at the same place, identify if (and why) people leave the site at the same place.
- Give ratings to data with stars, or another way. This way the user can get an overview of popular items both from ratings, and other ways the site recognizes popular items.
- Scrape/monitor web sites with similar content/news, and use this to i.e. present popular topics.
- Connect to open sources to link to and possibly perform analysis on.
- Chat-bot or similar. Come with hints, give users advice based on errors, if they are idle for a long time.
- Identify the best approaches/solutions/treatments to apply after accidents/spills based on previous accidents/incidents.
- Use machine learning to identify the need for access limitations on new data sets. This would require a model that does not mark restricted data as non-restricted.

4.3.1 Implemented Applications

4.4 Machine Learning in Environmental Engineering

4.4.1 Wastewater Treatment

Wastewater treatment plants are necessary for environmental protection, buffering nature from nature and industrial waste. These plants consists of a wide variety of operations, including mechanical, electrical, chemical, biological and physical. In order for the plant to function, the failures of these systems must be dealt with quickly and correctly.
in one study, researchers set up a plant so that the evaluation, preventing and handling of the failures was handled using Case Based Reasoning. For each case, the system found the most similar one, and the solution, or an adjusted one, could be used. The system would then learn from the results, optimizing the evaluation, preventing and solving of failures in the waste water treatment. [28]

4.4.2 Water resources management

State Vector Machine was used for prediction of the lake volume of The Great Salt Lake. Horologic systems are highly complex and variable, and it was of great interest to the researchers to look into SVS's ability to function in sparse, chaotic systems. In order for the system to work, full knowledge of the system had to be acquired.
The researchers combined chaos theory with SVM. The goal was to develop techniques to identify model parameters.
The study showed that SVM provided accurate prediction results, that could be utilized in developing water resources management strategies for the lake. [28]

4.4.3 Rainfall-runoff modelling

Machine learning can be applied to establish the relationship between rainfall and runoff. This is known to be non-linear and complex. Neural networks and genetic algorithms to create models and to identify optimal model structure and optimum coefficients. [28]

4.4.4 Anticipating environmental threats

Machine learning is used to evaluate possible risk in terms of hazard or toxic exposure of chemicals, to humans, animals and entire ecosystems. A research group in United States developed prediction models that could estimate the physio chemical properties of environmental chemicals. These models are freely available, and can be used by anyone to make predictions on new chemical sets, to improve toxicity models and inform about hazard/risk characterization.
In University of Queensland, Australia, a model was developed that could predict deforestation in Mexico and Madagascar. These models could then also be applied to other regions. [29]

4.4.5 Water Quality Analysis

Several machine learning techniques can be applied to analyze water quality data. Examples of its use are correlations between components using Random Forest, and classification models for identifying season of sample and the use of land where samples were taken. Clusterization techniques can be used to classify given data by various chemical, biological and physical parameters, into classes such as good, average and poor conditions. [30]

4.5 Machine Learning in Microsoft Azure

Microsoft Azure is an internet platform hosted on Microsoft's data centers, that provides an operation system and a set cloud services.
Azure provides multiple machine learning services, many of which are automated. Using Machine Learning Studio, machine learning applications can be implemented in the browser using drag-and-drop with no code writing needed. Azure supports open source libraries such as Tensorflow, PyTorch and scikit-learn. Azure also gives access to data storages such as Hadoop, that are well suited for big data and deep learning. These models can then be deployed as a web service.
All the supervised and unsupervised learning techniques mentioned earlier in the chapter are supported, and could, if data is stored on Azure, be applied quite easily on the data. [31]

4.6 Machine Learning in combination with ”Information Management System for Environmental and Public Health Information”

In the bachelor thesis ”Information Management System for Environmental and Public Health Information”[3], a system was developed for live data monitoring. By accessing this system, it could be possible to develop machine learning models on the historical data, which then could be used to predict the future development based on the live data. By analyzing other data sets from the same area, it would be possible to look for hidden and unknown connections, and the live data could then possibly be used for predicting other variables that are not monitored live.

5 Technical solution

We present the systems chosen for the database, for applying the big data and machine learning techniques, as well as how we are going to establish a website that will host these information.

The tools chosen for this project are shown in table 5.1.

Table 5.1: Technical solutions chosen

Area	Solution	Comment
Database	MongoDB [32] and Azure Blob [33]	MongoDB was chosen for the compatibility with HADOOP server, scalability and speed. This is explained further in section 5.1. Azure Blob is for file storage.
Server	Azure [34]	Azure was chosen for its high uptime and accessibility. It also has the required services available. There are also free credits provided by USN
Website	ASP.net MVC [35]	The authors have experience in C#. This is the programming languge for .net. The MVC pattern is also well suited for the chosen database
Website style	Bootstrap 4 [36]	Bootstrap 4 is the style used for previous projects.

5.1 Database

There are several requirements for the database structure. These govern the choice of database. Several structures have been considered. These are; MS SQL [37], MongoDB

[32] and FEDORA [38].
The data structure of FEDORA was used to structure article data. However, FEDORA is a database system designed to handle reports and articles. This is only a part of the requirements, and therefore, FEDORA does not cover the requirements. SQL has a major benefit of being very structured. This makes the data inherently searchable. This is also SQL's drawback, in addition to its lack of scalability. The data is highly un-structured, so SQL is not suitable.
MongoDB is a NoSQL database [39]. This means that it is capable of handling unstructured data and is simple to extend. Another benefit of NoSQL over SQL is its capability of scalability. For smaller data, SQL is faster than NoSQL, but for big data, NoSQL maintains its speed. This is illustrated in figure 5.1.

5.1.1 Non-relational database

NoSQL is a non-relational database system. There are several NoSQL databases. One of these is MongoDB. MongoDB is capable of running in an HADOOP environment[39][40], see section 5.1.3. An important part of non-relational database system is to define the structure of input data. This is done as a protocol shown in Paper B.

5.1.2 Relational

SQL is a traditional RDBMS. RDBMS is a highly structured way to store data. Everything is organized in tables with strict relationships. This is the strength of all RDBMS. The structure makes it simple to search, querying data is fast and output is well structured.
The major reason SQL is not used in big data, is the performance. When the tables become long and filled with a large amount of data, SQL's latency and vulnerability increases.

5.1.3 HADOOP

Not a database, but rather an environment, HADOOP is a logical system. Based on a scalable file system, HDFS, HADOOP can exist across several clusters. These clusters are all part of the same system, working together to process map-reduce jobs, ensure stability by redundacy and optimizing usage across nodes.
A cluster can be made of one or several nodes. The cluster has a master node. This master recieves a map-reduce job, divides it into several smaller jobs and passes them to each slave-node. This is done using YARN. [41] The slave nodes then processes their part

of the job and returns the result to the master node.
There is no limit to number of nodes, or node location. If the system latency is inclining, more nodes are needed. This makes HADOOP a highly scalable solution. The combination of HADOOP and MongoDB then makes for a highly scalable system capable of handling unstructured data, with reliability and high performance.

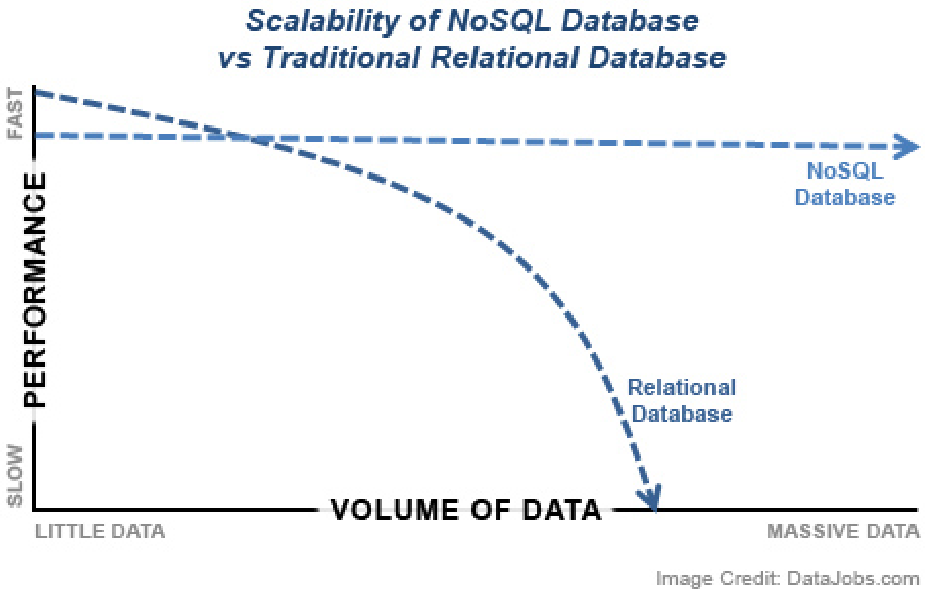


Figure 5.1: NoSQL vs SQL performance. This is only an illustration. [39]

5.1.4 Comparison of NoSQL and SQL performance

When comparing relational and non-relational databases, speed and scalability are the two main factors to consider. Several papers have been written, comparing various non-relational databases. These often conlude that non-relational databases, in general, are faster than relational. [42][43][44]. Of these papers, one compares MongoDB to SQL. There, benchmark results are included. These benchmarks are shown in tables 5.3, 5.4 and 5.5.
These benchmark tests, and the notion big data, type of data and system requirements lead to the selection of database.

Table 5.3: Benchmark test READ (ms) [42]

Database	Number of operations					
	10	50	100	1000	10000	100000
MongoDB	8	14	23	138	1085	10201
RavenDB	140	351	539	4730	47459	426505
CouchDB	23	101	196	1819	19508	176098
Cassandra	115	230	354	2385	19758	228096
Hypertable	60	83	103	420	3427	63036
Couchbase	15	22	23	86	811	7244
MS SQL Express	13	23	46	277	1968	17214

Table 5.4: Benchmark test WRITE (ms) [42]

Database	Number of operations					
	10	50	100	1000	10000	100000
MongoDB	61	75	84	387	2693	23354
RavenDB	570	898	1213	6939	71343	740450
CouchDB	90	374	616	6211	67216	932038
Cassandra	117	160	212	1200	9801	88197
Hypertable	55	90	184	1035	10938	114872
Couchbase	60	76	63	142	936	8492
MS SQL Express	30	94	129	1790	15588	216479

Table 5.5: Benchmark test FETCH ALL KEYS (ms) [42]

Database	Number of operations					
	10	50	100	1000	10000	100000
MongoDB	4	4	5	19	98	702
RavenDB	101	113	115	116	136	591
CouchDB	67	196	19	173	1063	9512
Cassandra	47	50	55	76	237	709
Hypertable	3	3	3	5	25	159
MS SQL Express	4	4	4	4	11	76

5.2 Big Data and Machine Learning

To apply big data and machine learning techniques, the requirement is a software/framework with built-in solutions for easy implementation. There is no requirement for a real-time solution, and no requirement for a in-built solution that can directly access the data.

5.2.1 Apache Spark

Apache Spark is a open source unified analytics engine for large-scale data processing. With the MongoDB Connector for Apache Spark, all of Sparks libraries can be accessed. It supports both batch data and Real-Time data processing.

5.2.2 MATLAB - Statistics and Machine Learning Toolbox

This is a toolbox for MathWorks MATLAB with functions to describe, analyze and model data. Students at USN have have access to MATLAB and this toolbox. MongoDB can be accessed from MATLAB.

5.2.3 TensorFlow

Originating at Google, TensorFLow is an open source software library for a range of tasks, including machine learning system based on neural networks. It provides a Python API, with third party packages for many other languages.

5.3 Website Design

The website design determines every aspects of the website. Website design is a process that gives :

- The layout.
- The colors.
- The text styles.
- the structure.
- graphics.

- the images.

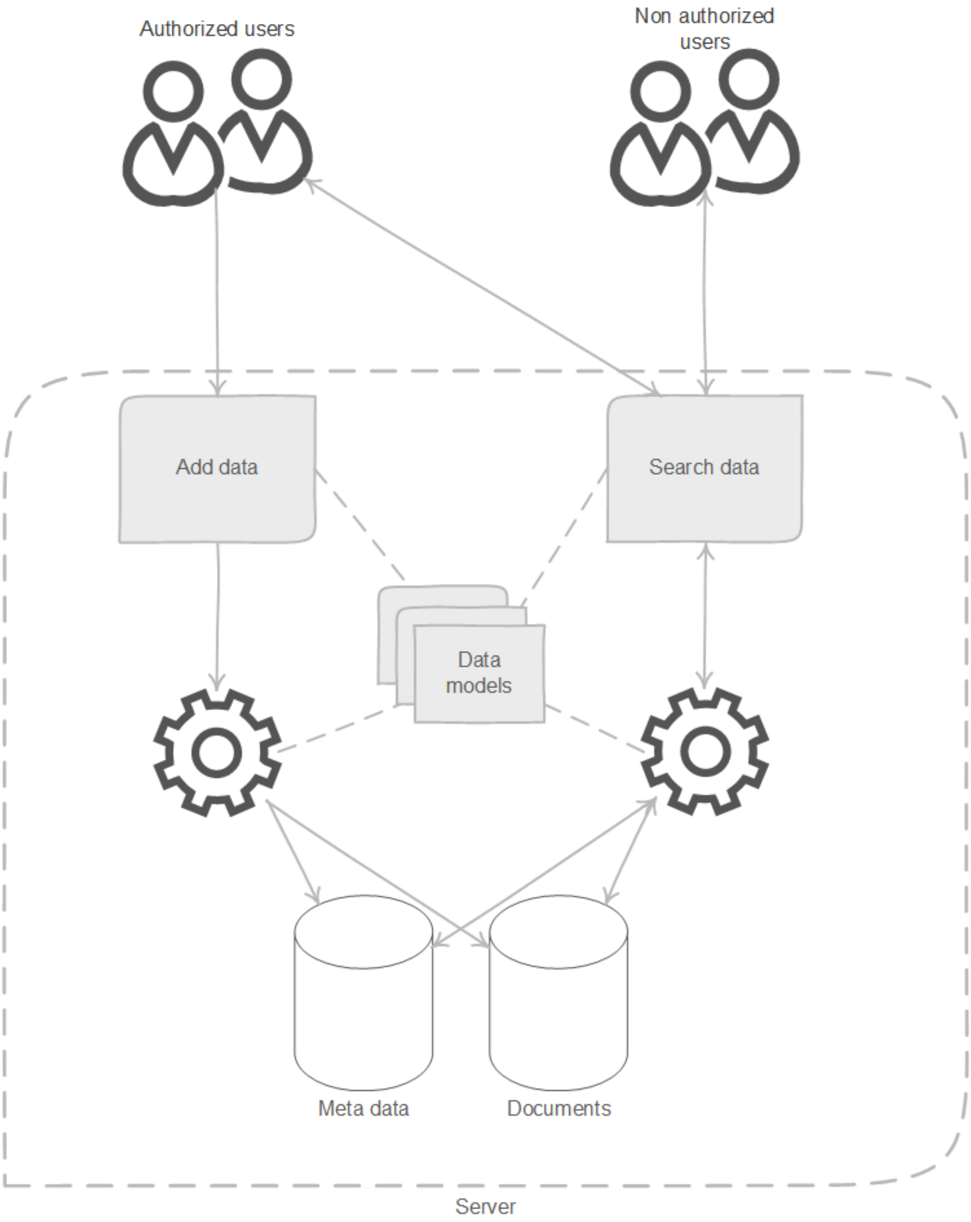


Figure 5.2: Overall system sketch

For our system, we used several tools to get the result displayed in our website.

5.3.1 ASP.NET

ASP.NET is a Microsoft technology made to create websites. ASP.NET is formed by two terms. the first one is ”ASP” (Active server page) which is the first technology made by Microsoft to create websites at 1996. The second term is ”.NET” which comes from ”framework.NET”.
The integration of ”framework.NET” to ”ASP” allowed to create ASP.NET and make it one of the most used technologies to create websites.

5.3.2 C#

C# or C-sharp is the computing language used to develop our website. It have been created by Microsoft. C# is very well suited to work with framework.NET and hence to work with ASP.NET.

5.3.3 MVC: Model-View-Controller

When developing a code, the developer often encounters the same problems. Thus, the right actions that were taken to solve these problems have been reunited under the name ”design pattern”.
One of the most famous design pattern is MVC. MVC seperates the system into three major parts:

- Model.
- View.
- Controller.

The goal of this separation is to make the information displayed for the more friendly and easier to understand than the way it is organized internally in the system.

Concretely, the ”controller” is a kind of bandmaster. It receives the request from the user and contact the ”model” and the ”view” to exchange the information and display them properly for the user.

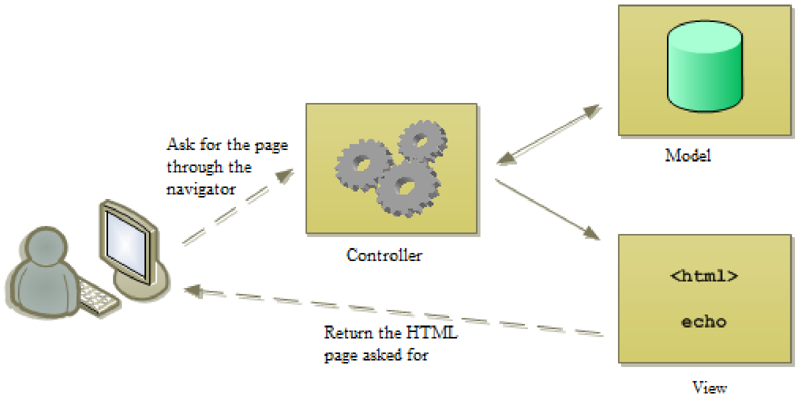
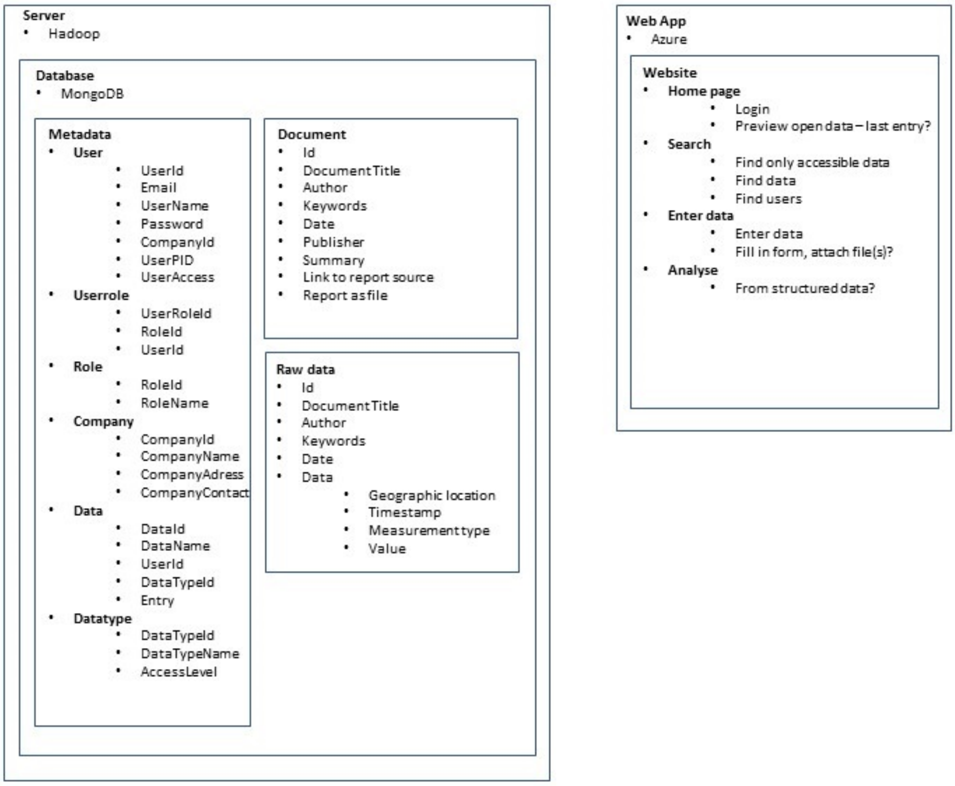


Figure 5.3: Client's request path[45]

5.4 System Overview



Server
- Hadoop

Database
- MongoDB

Metadata
- **User**
  - UserId
  - Email
  - UserName
  - Password
  - CompanyId
  - UserPID
  - UserAccess
- **Userrole**
  - UserRoleId
  - RoleId
  - UserId
- **Role**
  - RoleId
  - RoleName
- **Company**
  - CompanyId
  - CompanyName
  - CompanyAdress
  - CompanyContact
- **Data**
  - DataId
  - DataName
  - UserId
  - DataTypeId
  - Entry
- **Datatype**
  - DataTypeId
  - DataTypeName
  - AccessLevel

Document
- Id
- DocumentTitle
- Author
- Keywords
- Date
- Publisher
- Summary
- Link to report source
- Report as file

Raw data
- Id
- DocumentTitle
- Author
- Keywords
- Date
- Data
  - Geographic location
  - Timestamp
  - Measurement type
  - Value

Web App
- Azure

Website
- **Home page**
  - Login
  - Preview open data – last entry?
- **Search**
  - Find only accessible data
  - Find data
  - Find users
- **Enter data**
  - Enter data
  - Fill in form, attach file(s)?
- **Analyse**
  - From structured data?

Figure 5.4: System Overview

5.5 Web app - Database Communication Protocol

The MongoDB Wire Protocol is a simple socket-based, request-response style protocol, where clients communicate with the database server through a regular TCP/IP socket [32]. There is no connection handshake. Connection is made using a connection string with the Tls12 Secure Sockets Layer protocol.

Both for read and write, the first steps of accessing the data is similar. The client gets access to the database using the ’GetDatabase’ method. Then, the user gets access to a specified collection in the database using the ’GetCollection’ method.

This collection is in a BsonDocument-format. A BsonDocument is a binary encoding of documents that MongoDB uses when storing documents in collections. The BsonDocument is linked to a model (class) of the selected type, and will be structured as the model in the database when inserted. Similarly, the BsonDocument from the database will be related to model fields when retrieved. Because the BsonDocument is so flexible, it is easy to create many different types of models to store in the database. It can handle documents of any complexity, and is therefore well suited for this type of solution. It is on this collection that a request is made, for example to insert or to search for data.

To insert a document, the ’InsertOne’ method is used, along with the BsonDocument. Searching for one or more items in the database is done using a filter and the ’Find’ method. Multiple filters can be added together, using logic operators.

The protocol consists of two types of messages: client requests and database responses. In general, each message consists of a standard message header followed by request-specific data. The standard message header is structured as follows:

- messageLength: the total message size, including the header
- requestID: the identifier for this message
- responseTo: requestID from the original request (used in responses from db)
- opCode: request type

The most common operation commands are:

- delete: Deletes one or more documents
- find: Selects documents in a collection or a view
- insert: Inserts one or more documents
- update: Updates one or more documents

The model classes are shown in appendix C. These classes are mapped to the document structures shown in appendix B.

6 Requirements

This chapter is not definitive. It contains, for now, the initial requirements for the project. This is shown in figure 5.4

6.1 Data Collection

- The system must support different kinds of pollutants.[46]
- A list of people to contact to collect necessary data.
- Visit the websites of the most influential parties looking for data that may be already present.
- Maybe split these contacts by field of activity: industry, environment monitoring agencies, hospitals and healths responsibles...
- Maybe split these contacts by field of activity: industry, environment monitoring agencies, hospitals and healths responsibles...
- Decide what area are we going to concentrate on following our resources and the amount of data collected (if we don't have enough data on a particular region to make a proper analysis, maybe it would be best to not include it)

6.2 Data Analysis

- What data to focus on: air, water, health, polloutans (CO2, CO, ...), parameters (temperature, pH, ...) and what health aspect to study.
- The impact and improvement due to legislation and technologies. (Maybe a comparison with other countries).
- Do we have the right to make conclusions/analysis from these data.

6.3 Database

- Robust structure for storing different types of data.
- The Data need to be searchable.
- Backup.
- Test.
- Which information are we allowed to put in the database. (data, personal information, ...)
- How do we want to organize the database to make it easier to go through it.
- The database needs to have an open access for future implementation, but maybe this access should be under certain condition.

6.4 System

- Machine Learning should be used
- The system should fulfill proper ISO standard. We are only fulfilling the base for this standard. Later projects should build upon this to fulfill the full standard.
- Follow only the reccomendation, not be certified.
- The site should have some way to easily supliment and update data in the future.

6.5 Web Site

- Some of the functionality need a Login.
- Should the website be only in Norwegian, only in english or both of them. Also, what language start with. (Maybe for our group work with one language and leave the development in other languages for next groups).
- What information can we put on the website, maybe some of them need to be kept confidential. Moreover, what information to let people see without login and what to keep in the login section.
- Give a special interset on keeping the data safe and secure espacially the person related ones.

- Try to decide if we are going to need an immediate or future financial aid to maintain the website running and put in place all the requirements needed. Of course do an estimation.
- The code needs to be documented properly.
- The Web Site need to be easy to maintain by the supervisor or future students maintaining the solution in new student projects.

7 GDPR

A new legislation has been made at EU-level. This is known as the General Data Protection Regulation. This was implemented by Norwegian law on the 25 of May, 2018.[47] The essence of this law is that all data that can be traced to an individual is protected.

7.1 About GDPR

There are several definitions in this law that are essential to this project. These are:

- Applicability
  - This law applies to any automatic, partially automatic or non-automatic gathering of information, if the information is to be registered. Exceptions are if the data is to be used for personal or family affairs, while gathered by a physical person. In addition, it does not apply to cases of justice law.
- Area
  - Applies to Norwegian companies or legal bodies.
  - Applies to data gathering of individuals in Norway.

This means that this project is subject to this law.

7.2 Impact from GDPR

There are several ways that GDPR affects this project and system.

The main area that is affected, is health data. This is typically related to an individual or a group, and therefore, the storage of this data is subject to the law. Also, environmental data can be traced back to an individual or a group.
Because of this, a few measurements must be taken to be compliant to Norwegian law. For instance, the users must confirm that this is in fact public data. In addition, there is a data field for specifying if the data is public or not. Data that is not public is not

available, without logging in.

Future improvements of the system should move to a local server. Currently, the server is hosted in a cloud solution. This means the data is stored in a hard drive, owned by a different company. Because of this, only public data is allowed in the database.

When storing user information, this must first be approved by the user. There must also be a function, enabling a user to delete themselves, or part of the information about themselves. In addition, there must be a check, to verify that the user has agreed to this, typically implemented as a confirmation email.

When users log in, cookies are used to verify the connection. Users must be made aware of this. This does not apply to unregistered users, as there are no cookie solutions affecting them.

Part III

Future work and recommendations

8 Database

This and the following chapters contain suggestions for future work. These are specific suggestions.

Storage of documents could be moved to MongoDB, instead of having it in a separate blob service. Big files can be stored in a GridFS document. This splits the files and gives the added benefit of searching and retrieving parts of a document.[32]

Use an HADOOP Server. This enables clustering, which in turn allows for faster computations and more storage. HADOOP also implements machine learning tools.

9 Website

9.1 Adding data

Login function must be implemented. This function enables control of users, and the system can be used to store private data.

Making the categories dynamic. An administrator could be given the ability to add categories beyond 'environment' and 'health'. The database will accept any input, and can retrieve it. If this is to be implemented, dynamic classes is one way. Then, using reflection, the system would create classes based on a 'master' class, containing the definition of a category(name, values, author, etc.). [48]

Store the user that inserted the data. This gives control of which user adds what data, and makes it a searchable feature.

9.2 Search

Filtering options. This could be implemented on the same page as the results. Some checkboxes and fields to narrow a general search.

Search subscription. A user could subscribe to a specific search. This search is then run at a fixed interval, or every time something is added.

9.3 General

More information on the website should be included. The information about different pollutants and effects given in this report could be used.

Feedback and comments. There should be a function for giving feedback, reporting bugs etc.

10 Machine Learning

Giving location in document. If documents are stored in MongoDB, section 8, the location of the search term could be given, and the part of the document that contains the term, could be made available.

Generate reports based on the information in the database. Using the documents and raw data, reports on trends and issues can be generated.

Rating system. As a means to rate data, several functions could be used. This function is partly implemented, but could be exteded upon. Currently, the function rates a document by how many times it is searched for and opened. In addition a 'star' sytem could be implemented. The important thing is then to decide if this is individual ratings or a weighted average.

11 Data Collection

As stated in section 2.2, some questions need to get answers for facilitate the data collection.
We may suggest for the next groups to:

- Try to use the Norwegian language to contact people.
- Use a professional mailbox instead of a personal one like gmail or outlook.
- Try to phone call, if possible, the contacts; at least to introduce the project and then continue the discussions through mails.
- Concentrate the research on the Telemark region more than on the pollutants type.
- When talking to the contacts, ask for data by relatively short period of time (5 or 10 years); because asking for data for the last 50 years may be frightening.

Discussion

The background for this project was to focus on Grenland, Telemark: one of the areas in Norway with the highest density of heavy industry and associated emissions to the environment. Throughout the last 50 years, the awareness of emissions and the effects of those became increasingly significant. With an increasing focus over the years, both from government and the public, the emissions have now been severely limited both by public regulations and by the industries own efforts. Environmental challenges is an important and popular topic these days, and in order to understand and implement ways of handling them, it is important to have historical knowledge, and be able to put it into a context.
The task was to collect and systematize historical environmental and health information from different sources in Grenland, and create a technical solution with a search-able database for this data.
Then, in order to gain insight from the data, data analysis and big data and machine learning techniques was to be applied in order to find information from the available data. Scrum was to be used as the project planning and management system.
The main goal became then to offer a platform for professional and non-professionals, where as much data as possible was gathered and stored, and where data analysis on the data could be made. Sub-goals were to become a reference source for monitoring of health and environment-related data, to to offer all types of data, to create a user-friendly platform which also provides information and access to other related pages, and to implement a contact list that could help researchers get in touch.

To obtain data from the sources proved to be difficult: Some contacts did not respond to the requests, others have shown little or no interest, and it was pointed out that the data already was stored in another solution, and that they therefore was skeptical of duplicating their data to another system. Some data in the form of reports in pdf format was collected, as well as numeric data in csv and excel-format. The data format would be another challenge regarding the task, and it will discussed later. Luftkvalitet.info, together with nilu.no, has an API for extracting data, and some of this data was downloaded. It is however not always appreciated to scrape entire websites for data without permission, so only some sections were downloaded for analysis. All the collected numeric data was environmental data.
The technical solution is a website linked to a server and database. The architecture for the website was Model-View-Controller (MVC), which had the benefits of easily decoupling and separating the different layers of logic, allowing for efficient code reuse and

parallel development[49]. This was especially beneficial in this project with many types of inputs, as different representations (views) easily could be created based on the data structure, without having to edit or create much code.

The database design for the project was chosen to be NoSQL, which stand for ”Not only SQL”. NoSQL includes several types of databases, including documents and tables, is very scalable and are adressed for large volumes of structured, semi-structured and unstructured data[50]. The main reason for NoSQL was that the system had to be able to store any type of data, which excluded purely relational databases. Further, in order to develop and create a NoSQL database there is no need for a detailed database model: it is object-orientated and the database will update directly from the MVC-model. This meant that there not had to be any work done on the database side. This saves a lot of time during development, and also during operation, especially when the system could be expected to be presented with many different types of inputs from multiple sources, all in a different format. Implementation of the website and database went well, after studies on how to implement the MVC-structure and the MongoDB-format, which is an NoSQL database for documents.

As the project went on and very little data and interest was offered by contacts, the goal of becoming a preferred site both for professionals and non-professionals regarding health and environmental topics in Grenland, became more and more unrealistic. Even the goal of adding a substantial amount of data became unrealistic, which was the basis for the whole site.

With no data, very little analysis and machine learning could be applied to the data, and some of the focus shifted to improving the web site functionality, possibly by using machine learning. Different layouts, presentations and functionalities were tested, including recommendations based on other users behaviour, which now is included, but this would be a goal far less important than the main part: to create a searchable technical solution, with data.

The projections of what this site could be became almost endless, but with no value added in the important part: the data collection. A broad range of good ideas and topics for the web site are listed up in this report, but when so much was unclear about the site, more effort was instead focused on making sure the parts included in the website worked well, instead of including parts that may not provide much value compared to extra time maintaining the site. Focus was also shifted to looking into data analysis on the small data sets that were obtained.

The project description seems clear enough, but as the time went the authors of this report feel that the specifications changed significantly and became increasingly vague: Being told that this system was asked for, it was assumed that contacts was aware of the solution and would provide data. When that was not the case, the assignment shifted over to ”selling” the solution, which is something that should have been done before initiating a complete build of a large project such as this. If the goal from the start had been to present a solution to sell, a dummy solution would have been created, just for presentation. In a dummy solution, a lot of ”fake” functions and applications are implemented, with the purpose of showing the potential customer what the system could to, most often without actually being implemented. This could then include analysis and machine learning techniques as well as functions for a better user experience: comments, better search functions, news, updates from other sites, recommendations, popular items, etc, all time-consuming to implement.

This project, however, worked from the assumption that this system was ”sold”. When a system is sold, the way to develop is to implement the most important things first from the requirements, while receiving feedback on a regular basis. It would be very time-consuming and wasteful to spend large amount of hours programming and editing in things that the customer may not want.

Similarly, many good ideas and proposals came up under meetings at USN, but by doing so the goal of this site, and what the final version should be compared to, increased every time, whereas the project actually focused more on the main part and how to solve that: the collecting, storing and analysis of data. The focus seemed to shift away from Environmental and Health-related issues to purely implementing a site that looked and behaved good, taking time from reading and writing on relevant topics, such as machine learning, analysis and Health/Environment-data. It is the authors beliefs that this project is better off being delivered with a well-working website format, although with few user functions, but with a good report with suggestions for future work, than being delivered with many implementations that may not be there in the final version, and a lot of extra code to maintain. This way, the next group can find ideas here or other places, create dummy versions of some, present to the customer and then possibly implement.

Conclusion

The system is created as a web-based solution. Data can be collected and retrieved. Some numeric data analysis is performed, however it is not relevant to all sorts of data, e.g text articles. One of the biggest issues to deal with, is the data collection, especially when it comes to historic data that has not been digitalized.

Machine learning is introduced to the system, and used as a tool to enhance the searching feature of the website by predicting the users behavior. In addition, it is suggested to do data cleansing by machine learning techniques. This is discussed in 5.2

The website will be improved more and more by getting feedback from the customers, therefore Scrum method is used to manage the project.

As future work, using data sources provided in this project could save a lot of time for expanding the project. The database itself has been chosen based on analytic research and the website is available. One may focus on data collection, more detail in data analysis and also consider replacement for Microsoft Azure as server.

References

[1] A. Z. Gjerseth and L. Ang, ‘Development of a database system for environmental and public health information’, University College of Southeast Norway, Tech. Rep., 2016.

[2] A. Chynchenko, ‘Environmental public health information management system’, University College of Southeast Norway, Tech. Rep., 2017.

[3] H. M. L. Kristensen, H. Mølmen, J. Johansson, K. B. Skjelbred and T. Prestvik, ‘Information management system for environmental and public health information’, University College of Southeast Norway, Tech. Rep., 2017.

[4] O. A. Grytten, ‘Environmental public health information management system’, University College of Southeast Norway, Tech. Rep., 2018.

[5] R. Zevenhoven and P. Kilpinen, *Control of pollutants in flue gases and fuel gases*, 3rd. Helsinki University of Technology, 2004.

[6] ‘Directive 2008/50/ec of the european parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for europe’, *Official Journal of the European Union*, May 2008.

[7] [Online]. Available: `http://www.environment.no/goals/4.-pollution/target-4.4/`.

[8] [Online]. Available: `https://en.wikipedia.org/wiki/European_emission_standards`.

[9] G. Husdal, L. Osenbroch, Ö. Özlem and A. Østebrøt, *Cold venting and fugitive emissions from Norwegian offshore oil and gas activities – summary report.* Norwegian Environment Agency, 2016.

[10] D. Lerda, *Polycyclic Aromatic Hydrocarbons (PAHs)*, 4th. 2011.

[11] *Pahs emissions.* [Online]. Available: `https://www.norskeutslipp.no/en/Components/Emission/Polycyclic-aromatic-hydrocarbons/?ComponentType=utslipp&ComponentPageID=232&SectorID=9999`.

[12] *Carbon monoxide emissions.* [Online]. Available: `https://www.norskeutslipp.no/en/Components/Emission/Carbon-mono%20xide/?ComponentType=utslipp&ComponentPageID=77`.

[13] *Reference document on best available techniques in the large volume organic chemical industry*, Feb. 2003. [Online]. Available: `http://eippcb.jrc.ec.europa.eu/reference/BREF/lvo_bref_0203.pdf`.

[14] *Greenhouse gas*, Oct. 2018. [Online]. Available: `https://en.wikipedia.org/wiki/Greenhouse_gas%5C#Greenhouse_gases`.

[15] *Ozone depleting substances*. [Online]. Available: `http://www.mfe.govt.nz/more/hazards/risks-ozone-depleting-substances/what-are-ozone-depleting-substances`.

[16] *Guidance on assessment under the eu air quality directives*. [Online]. Available: `http://ec.europa.eu/environment/air/pdf/guidanceunderairquality.pdf`.

[17] N. Rodríguez Eugenio, M. McLaughlin and D. Pennock, 'Soil pollution: A hidden reality', 2018.

[18] *Contaminated soil*. [Online]. Available: `http://www.environment.no/topics/hazardous-chemicals/contaminated-soil/`.

[19] E. A. Vik, G. Breedveld and T. Farestveit, *Guidelines for the Risk Assessment of Contaminated Sites*. SFT, 1999.

[20] S. Alriksson, 'Certification for ems iso 14001, emas', *Department of Technology, University of Kalmar*,

[21] J. Beehusen, 'Oslo får felles ikt-plattform', *Teknisk ukeblad*, no. 165, pp. 78–81, Sep. 2018.

[22] *Norwegian institute for air research(NILU)*. [Online]. Available: `https://api.nilu.no/`.

[23] N. H. Eldrup, 'Basics of managing risks in projects', 2012.

[24] *Vishal maini: Machine learning for humans*. [Online]. Available: `https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12`.

[25] *Machine learning: An in-depth guide - overview, goals, learning types, and algorithms*. [Online]. Available: `https://www.innoarchitech.com/machine-learning-an-in-depth-non-technical-guide/?utm_source=kdnuggets&utm_medium=post&utm_content=originallink&utm_campaign=guest/`.

[26] *Maja pantic : Introduction to machine learning & case-based reasoning*. [Online]. Available: `https://ibug.doc.ic.ac.uk/media/uploads/documents/courses/syllabus-CBR.pdf`.

[27] *Artificial intelligence 2e : Case-based reasoning*. [Online]. Available: `https://artint.info/2e/html/ArtInt2e.Ch7.S7.html`.

[28] *Catherine wilcox, wei lee woon, zeyar aung: Applications of machine learning in environmental engineering*. [Online]. Available: `http://www.dnagroup.org/techreps/DNA-2013-03.pdf`.

[29] *4 ways machine learning protects the environment.* [Online]. Available: https://www.ua-magazine.com/4-ways-machine-learning-protects-environment/#.W9_5x7WZ3mE.

[30] *Aleksei shkurin: Water quality analysis using machine learning algorithms.* [Online]. Available: https://www.theseus.fi/bitstream/handle/10024/106320/Shkurin_Aleksei.pdf?sequence=1.

[31] *Microsoft azure: Machine learning.* [Online]. Available: https://azure.microsoft.com/en-us/overview/machine-learning/.

[32] M. Inc, *Mongodb for giant ideas*, 2018. [Online]. Available: https://www.mongodb.com/.

[33] *Blob storage.* [Online]. Available: https://azure.microsoft.com/en-us/services/storage/blobs/.

[34] *Your vision. your cloud.* [Online]. Available: https://azure.microsoft.com/en-us/.

[35] *Mvc.* [Online]. Available: https://www.asp.net/mvc.

[36] M. Otto and J. Thornton, *Bootstrap.* [Online]. Available: https://getbootstrap.com/.

[37] *Run sql server on your favorite platform.* [Online]. Available: https://www.microsoft.com/en-us/sql-server/sql-server-2017.

[38] *Fedora archive.* [Online]. Available: https://duraspace.org/fedora/.

[39] F. Lo, *What is hadoop and nosql?*, 2018. [Online]. Available: https://datajobs.com/what-is-hadoop-and-nosql.

[40] M. Inc, *Hadoop and mongodb*, 2018. [Online]. Available: https://www.mongodb.com/hadoop-and-mongodb.

[41] *Apache hadoop hdfs.* [Online]. Available: https://hortonworks.com/apache/hdfs/.

[42] Y. Li and S. Manoharan, ‘A performance comparison of sql and nosql databases’, Univercity of Auckland, Department of Computer Science, Tech. Rep., 2013.

[43] C. Hadjigeorgiou, ‘Rdbms vs nosql: Performance and scaling comparison’, The University of Edinburgh, Tech. Rep., Aug. 2013.

[44] N. Leavitt, ‘Will nosql databases live up to their promise?’, *Computer*, vol. 43, no. 2, pp. 12–14, Feb. 2010, ISSN: 0018-9162. DOI: 10.1109/MC.2010.58.

[45] M. Nebra, *How does an mvc structure work.* [Online]. Available: https://openclassrooms.com/fr/courses/4670706-adoptez-une-architecture-mvc-en-php/4678736-comment-fonctionne-une-architecture-mvc.

[46] *Dioxins and their effects on human health.* [Online]. Available: `http://www.who.int/news-room/fact-sheets/detail/dioxins-and-their-effects-on-human-health`.

[47] *Lov om behandling av personopplysninger (personopplysningsloven)*, 2018. [Online]. Available: `https://lovdata.no/dokument/NL/lov/2018-06-15-38`.

[48] Rpetrusha, *System.reflection.emit namespace.* [Online]. Available: `https://docs.microsoft.com/en-us/dotnet/api/system.reflection.emit?view=netframework-4.7.2`.

[49] *Tutorialspoint: Mvc introduction.* [Online]. Available: `https://www.tutorialspoint.com/mvc_framework/mvc_framework_introduction.htm`.

[50] M. Inc, *Mongodb : Nosql explained*, 2018. [Online]. Available: `nosql_explained`.

Paper A

Task Description

What can we learn from the previous historical environmental issues in Grenland? How can we use historical data on emissions and environmental exposure in Grenland to increase knowledge about environmental and health issues?

The projects should obtain an overview of the emissions and effects of emissions in Grenland for the last 50 years. We should systematize this knowledge in a searchable database to facilitate access to and use of this data for research and development in the future.

The project should uncover what changes in health and environment we have had in Grenland for the past 50 years. We should find out what cleansing knowledge can be conveyed, exported or transferred to today's challenges at local, national and international levels. What can be learned from history and transferred to today's challenges, the green shift and to shape the new industry.

The project will consist of several parts and the students may partly affect the focus and facility in cooperation with the supervisor and Telemark Hospital, Porsgrunn Municipality and the other partners.

Main parts of the project are:

- Collect and systematize historical environmental information from different sources in Grenland.
- Create a technical solution for a searchable database for storing information about historical environmental and health monitoring in Grenland. You should find an appropriate technical platform for the system as well.
- Data analysis and possibly apply big data and machine learning techniques to make it easier to find the proper information in the available data.

- Project planning and management using Agile (Scrum) project methods.
- Write a detailed report of the project.

Paper B

Database structure document

University of
South-Eastern Norway

www.usn.no

Monitoring of environment and health in Grenland

MongoDB document structure

MP-11-18

Faculty of Technology, Natural Sciences and Maritime Sciences

Campus Porsgrunn

Preface

This document describes the document structure of the database used to support website created. This database is MongoDB, running in an HADOOP environment. A description of this is available in the report.

Porsgrunn, 6th November 2018

MP-11-18

Contents

Nomenclature

Symbol	**Explanation**
BSON	Binary JSON
HDFS	Hadoop Distributed File System
JSON	JavaScript Object Notation

1 Introduction

The database is structured by databases of collections.[1] There are some databases containing administrative data, configuration data and such. These are not discussed further.

The chapters describe the database and document structures. The document structure are presented as nested BSON documents.
Some knowledge of MongoDB and object oriented programming is assumed.

2 Database and collections

Main database for the project is named 'main_db'. This database has three collections; 'documents', 'rawdata' and 'users'.
'document' contains reports, articles and other publications.
'rawdata' contains raw data from files such as .csv, .xls, .lvm etc. 'users' contains login information used by the website.

- 'main_db'
  - 'documents'
  - 'data'
  - 'users'

3 Report document structure

This chapter describes the BSON document structure for the collection 'documents'. Table 3.1 shows the document structure. If the report is to be uploaded, GridFS should be used, otherwise a size limit of 16MB is applied. [2]

Table 3.1: Report document structure

Key	**Value comment**
id	Automatic id field
user	User name of the user who inserted data
datatype*	Type of data, environment, health or both
datestored	Date of insertion
name*	Document title
author*	Author
keywords	Some keywords describing the document
date*	Date of publication as datetime
publisher	Publisher
summary	A summary of the report
source^1	Link to the report source
report^1	The report as file

Fields labeled * are required. Only one of the fields labeled ^{1} are required

4 Raw data document structure

The raw data is a nested BSON document. Table 4.1 shows the main document. Table 4.2 shows the sub document. The actual data is stored in the sub document. The entire document then contains both meta-data and data. [1]

Table 4.1: Raw data document structure

Key	Value comment
id	Automatic id field
user	User name of the user who inserted data
datatype*	Type of data, environment, health or both
datestored	Date of insertion
name*	Document title
author*	Author
keywords	Some keywords describing the document
date*	Date of publication as datetime
location	Location of collection
data*	A sub document containing the data, see table 4.2

Fields labeled * are required

Table 4.2: Raw data sub document structure

Key	Value comment
column	Column header value
row	Row header value
value	Value of the datapoint

All fields are required

5 User document structure

The user data for the website is stored in this manner. Fields using arrays have an example to show the usage. Password information should be 'hashed' before it's passed to the database.

Table 5.1: User document structure

Key	Value comment
id	Automatic id field
username*	User name for user
password*	Hashed password
name*	Actual name as sub document containing first, middle and surname fields
email*	contact email
phone*	contact phone
privileges*	given privileges as array. ['read', 'write', 'admin']

All fields are required

Bibliography

[1] *Documents*. [Online]. Available: `https://docs.mongodb.com/manual/core/document/`.

[2] *Gridfs*. [Online]. Available: `https://docs.mongodb.com/manual/core/gridfs/`.

Paper C

Class diagram for MEHIG web app

University of
South-Eastern Norway

www.usn.no

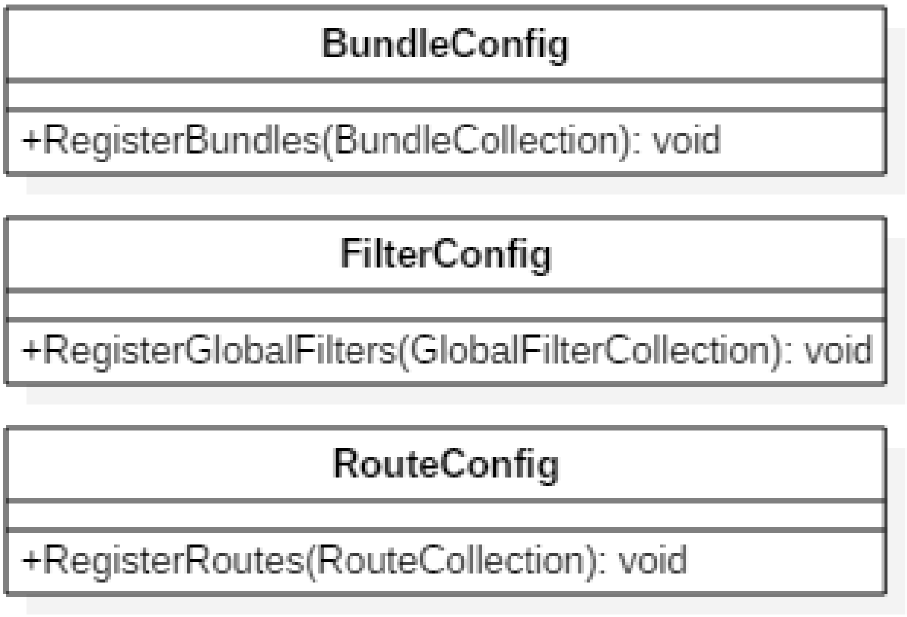
Monitoring of environment and health in Grenland

Class diagrams for MEHIG web app

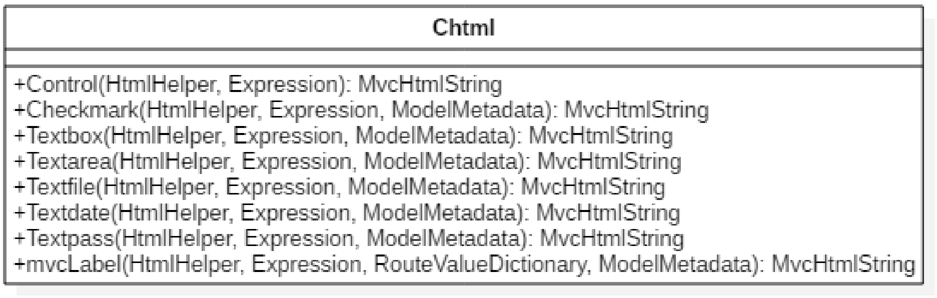
MP-11-18

Faculty of Technology, Natural Sciences and Maritime Sciences

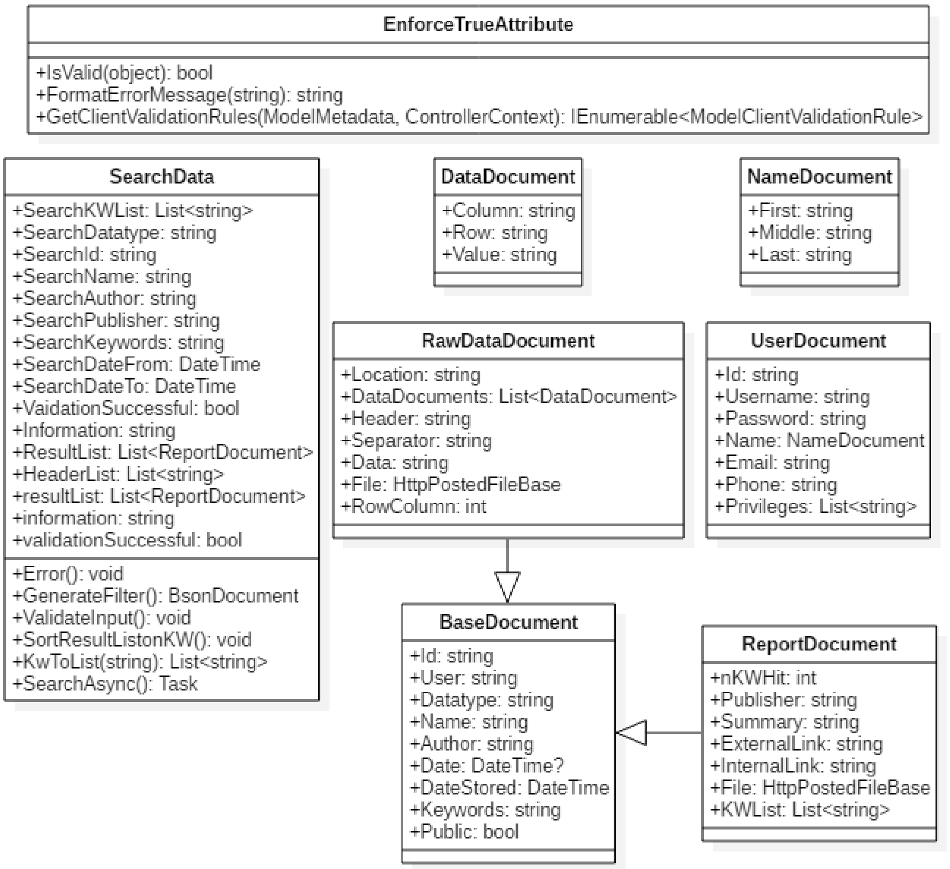
Campus Porsgrunn


BundleConfig
+RegisterBundles(BundleCollection): void
FilterConfig
+RegisterGlobalFilters(GlobalFilterCollection): void
RouteConfig
+RegisterRoutes(RouteCollection): void

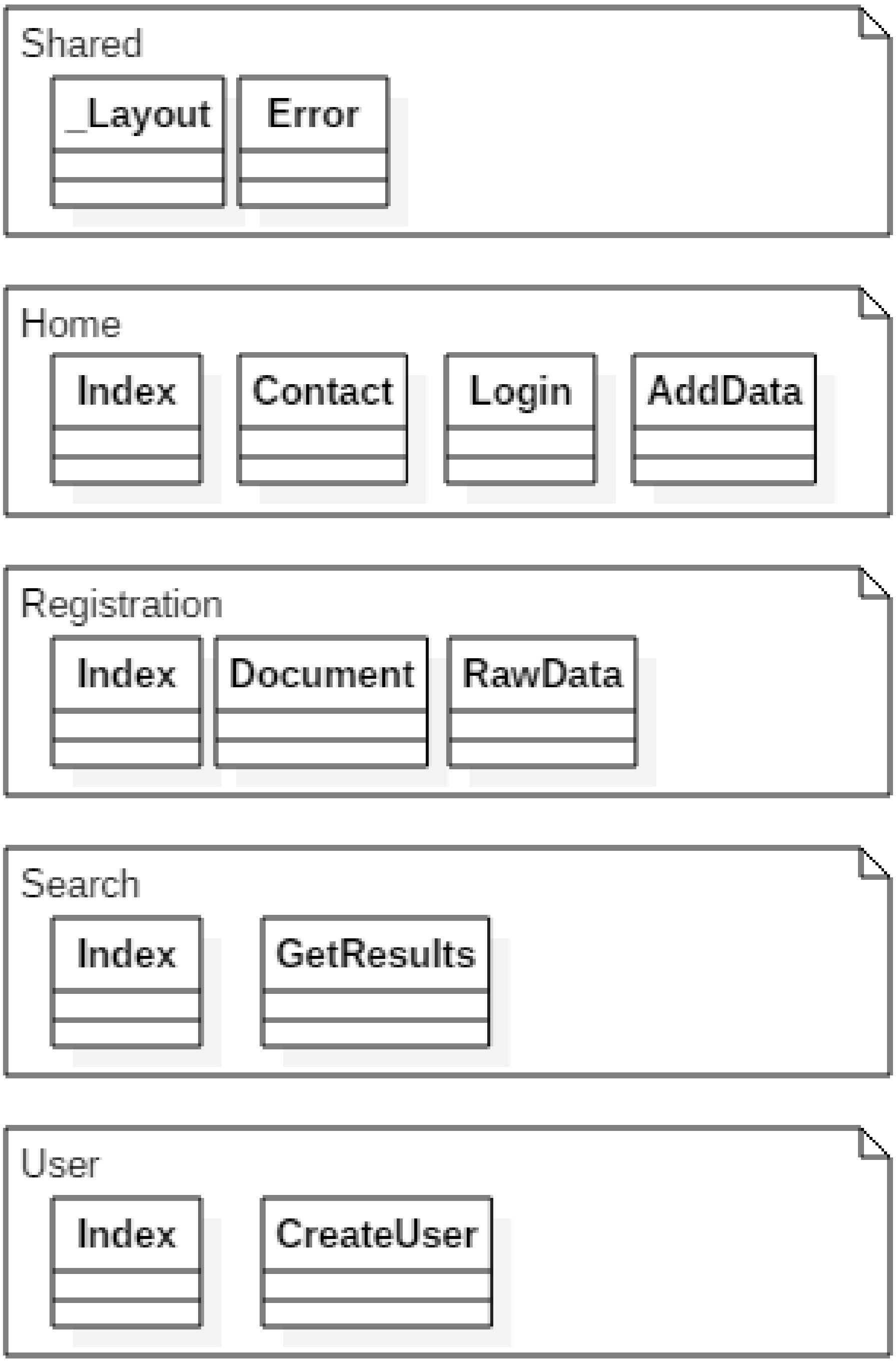
Configuration classes


Chtml
+Control(HtmlHelper, Expression): MvcHtmlString
+Checkmark(HtmlHelper, Expression, ModelMetadata): MvcHtmlString
+Textbox(HtmlHelper, Expression, ModelMetadata): MvcHtmlString
+Textarea(HtmlHelper, Expression, ModelMetadata): MvcHtmlString
+Textfile(HtmlHelper, Expression, ModelMetadata): MvcHtmlString
+Textdate(HtmlHelper, Expression, ModelMetadata): MvcHtmlString
+Textpass(HtmlHelper, Expression, ModelMetadata): MvcHtmlString
+mvcLabel(HtmlHelper, Expression, RouteValueDictionary, ModelMetadata): MvcHtmlString

Class for creating dynamic controllers



Models for the data in the database


Shared
_Layout
Error
Home
Index
Contact
Login
AddData
Registration
Index
Document
RawData
Search
Index
GetResults
User
Index
CreateUser

Views