Interim Report
EE5500

Name: Wojciech Lesnianski
Student number: 1644612

Electronic and Computer Engineering
School of Engineering and Design

Dr. Ali Mousavi

Thursday, September 28, 2017

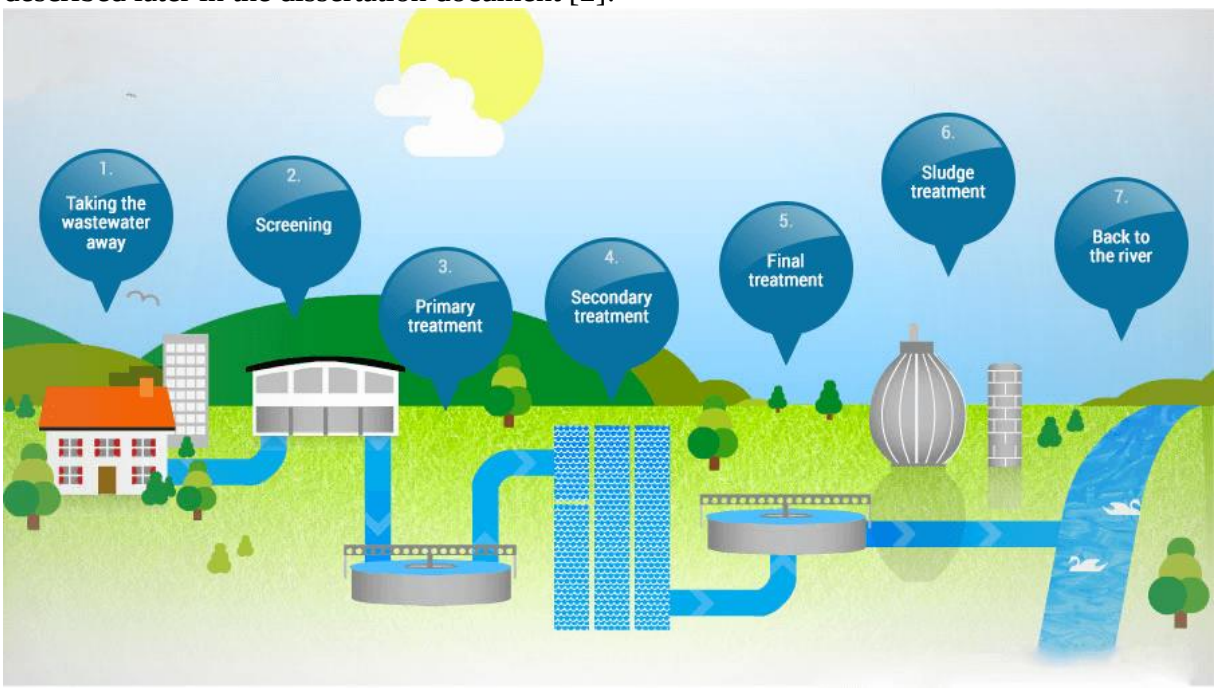
Table of content

Introduction

Acquisition, analysis and modelling of historical and real-time water-plant data. Overcoming some of the existing barriers of interoperability and harmonisation of data and information.

Access to clean water is the most basic and fundamental type of the human infrastructure. The quality of life highly depends on the accessibility to clean water. We require water not only for drinking, but also for cooking, and washing. Additionally, various professions and commercial establishments, like farmers or restaurants, could not exist without certain quality and quantity of water. The quantity of clean water in most cases, depends on collecting water and sewage from rivers and lakes, cleaning it in dedicated water-plants and thus bringing it to a specific quality standard, and then distributing it back into the waters.
A software groundwork for acquisition, analysis and modelling of historical and real-time data of water-plants will be the main topic of this master thesis. The Project will be done in partner work, although the tasks will be strictly separated and the outcome of one part of the project won't affect the outcome of the other part. This dissertation is dealing with the problem of acquiring, harmonizing and providing water-related data and leaves the analysis and presentation to the partner project. [1]

The specific cleaning process in the United Kingdom consists of 7 steps, which will be described later in the dissertation document [2]:



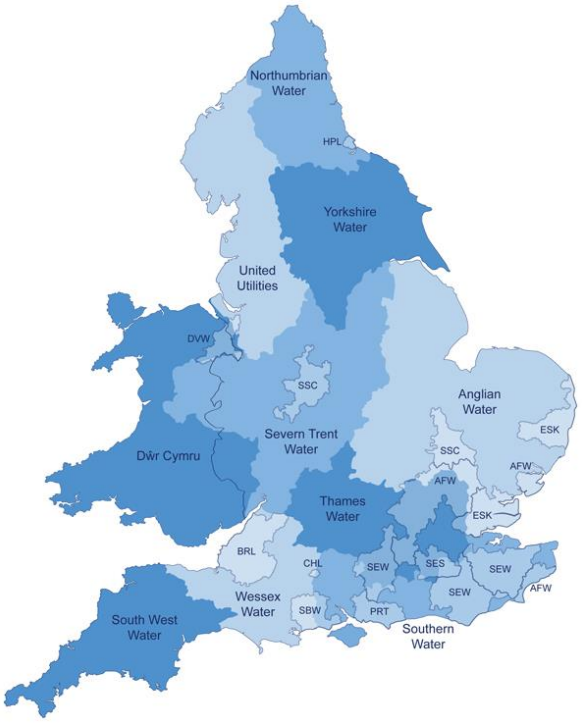
(Source: http://www.water.org.uk/about-water-uk/wastewater 18.09.2017 last accessed: 28.09.2017)

The most individuals will be interested in the outcome of step 7, which also indicates the quality of water available for public usage, nonetheless the incoming and outgoing water of the other steps provides different kind of data which might be interesting for different kind of reports, especially due to the fact, that each step deals with a specific problem, meaning that all possible to gather data will also be gathered harmonized and stored by our system, for further investigation.

Background to the project

On average it costs a UK customer 1 pound a day to drink high quality water. This money goes into the water and wastewater treatment of around 16 billion litres of wastewater, gathered in around 345.000km of sewers, in around 9000 wastewater plants – every day. [3] [4]

The water supply regulations, set by the government, regulate the water treatment process of every water provider whose area is wholly or partially in the United Kingdom. The list of indicator parameters is long and contains minimum, maximum values and ranges within which values are allowed to lie. Only if all regulations apply the water may be called drinking water. With all the regulations and monitoring organisations the quality of UKs water might seem assured – yet the process of doing so is very troublesome and laborious. Twelve big companies, responsible for water and sewerage, cover most of UKs water supply. Additionally, there are some water-only companies providing water for some of the remaining regions. [4] [5]



(Source: http://www.ofwat.gov.uk/households/your-water-company/map/ last accessed: 28.09.2017)

The water quality is regulated UK-wide, yet the way the different companies ensure their quality and monitor their water treatment process is not unified. This makes comparison between companies, as well as getting a global picture difficult. Reacting to lack of quality water in specific regions, or forecasting such a scenario, while still monitoring which of the remaining regions has enough "spare" quality water to help out the company in need would be a lot easier with a common information base. It would simplify the monitoring of local area changes caused by changes in the water and wastewater treatment regulations. To assure better forecasts or more meaningful reports, other information bases, like weather information might be taken into account – but those external systems are not a topic in this part of the (data-gathering) system.

The advantages of a big dataset from various sources are obvious – especially in a case where the geographical location of sources also mattering. Co-operating, comparing, planning,

monitoring and analysing is a lot easier when all the data is stored at seemingly one place in a unified format. Attempts to fulfil this task were started in more than one topic area and will be investigated before the start of an own attempt.

Initial survey

Quality of water in the United Kingdom is ensured by a number of organisations consisting of governmental, regulator and consumer organisations, all of them having their own task including the following [6]:

- Governments
  - Defra
    - Looking after the natural environment
    - Supporting the food and farming industry
    - Sustaining a thriving rural economy
  - Welsh Government
    - Improving the lives of people in wales
- Regulators
  - Drinking Water Inspectorate
    - Providing independent reassurance about UKs water quality
  - Environment Agency
    - Regulating industry waste
    - Regulating water quality and resources in England
    - Managing the risk of flooding from rivers, reservoirs, estuaries and the sea
  - Natural England
    - Helping to protect England's nature and landscapes
  - Natural Resources Wales
    - Ensuring sustainability of resources in England
  - Ofwat (for England & Wales) and WICS (for Schottland)
    - Regulating the water and sewerage sectors
    - Setting price limits for customers
    - Ensuring companies run efficiently
    - Encouraging resilience
- Customer Watchdog
  - CCWater
    - Promoting customers interests to governments, regulators and water companies
    - Providing advice and complaint handling service for customers

Those companies might all have interests in the water specific data. The source of information for possible questions to our system will be their webpages.

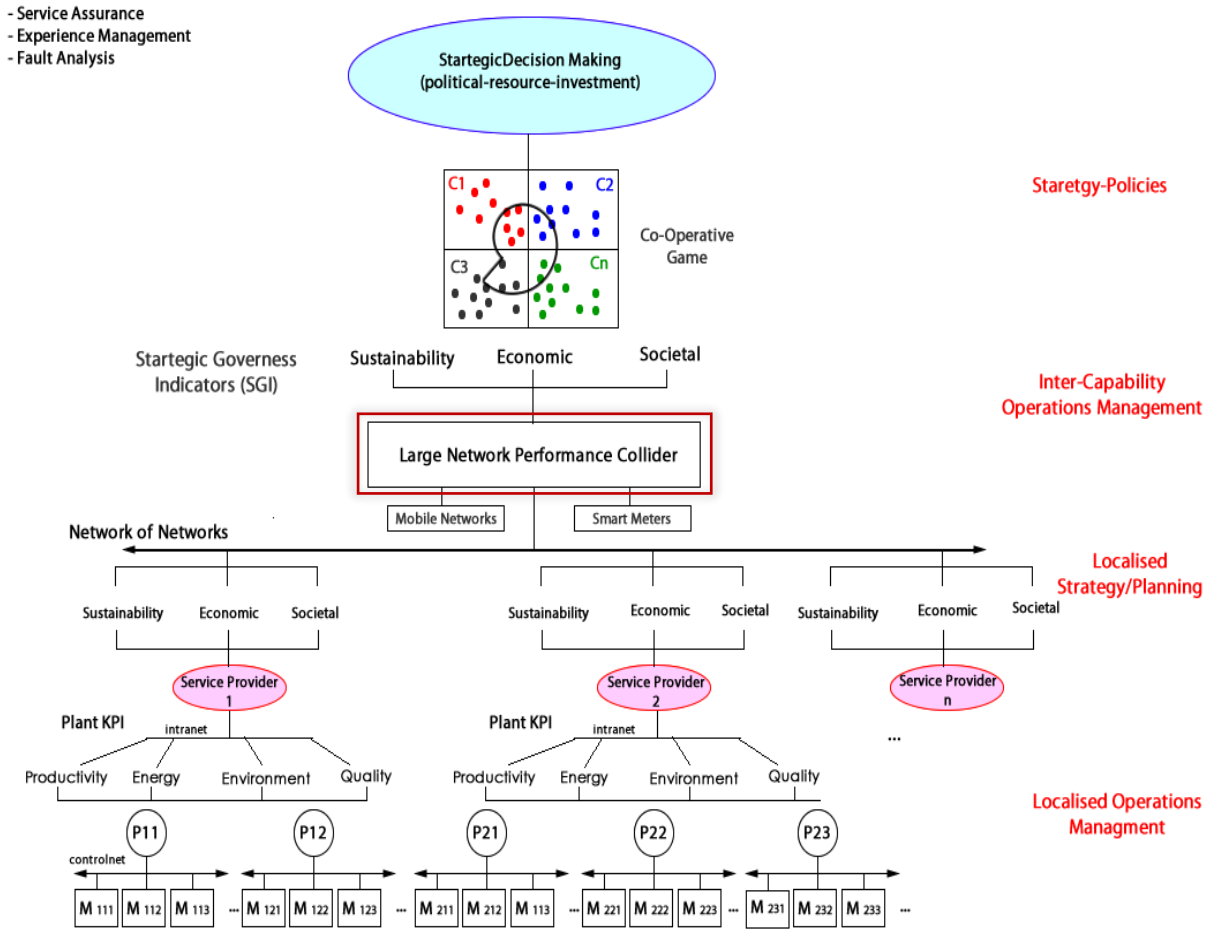
The already mentioned water suppliers will need to be investigated on their provided data and on their interests.

Government website [5] provides all regulations which need to be followed by water suppliers.

The Biobank Standardisation and Harmonisation for Research Excellence in the European Union described an attempt of harmonization of various different sources of data. [7]

Aims and Objectives

The purpose of this project – and this part of the project in specific – is to investigate and design knowledge and data engineering infrastructure for big amounts of water and wastewater treatment process specific data. This includes finding the best way to gather data from water providers in terms of cost, efficiency, effectivity and security, as well as harmonizing and providing the data in the most fitting manner. The below picture shows the idea of how the system should interact with the outside world. In this picture the Large Network Performance Collider is our Data Harmonization Layer and Data Provider on top of which our partner project creates analyses and reports.



(Source: [8] Page 3)

Defining questions, which the system must/should answer. **Designing a fitting data schema** with appropriate target variables. In the desired solution, the system should **gather data** from different service providers, without their need to adapt their data schema, and **convert** the data to an own, predefined and harmonized data schema as well as **provide** this data to third parties. In the best case, all required data is provided to the public by every service provider in some way – then the only problem is the harmonization of the data. Aside of the real time service provider data, also historical data should be accessible through our system. This data is meant to be used for comparisons and forecasts and needs to be adapted to our data schema as well. Adding new sources of data to the system should not affect the systems performance by a noticeable amount and data requests should be processed in the most efficient way possible. Adding additional variables of interest should be possible.
Aside of the techniques, investigating on what the best technologies are to fulfill the task is also a part of this project. This also includes finding the best Data Host.

Experimental/investigative methods to be adopted

The most important investigative method in this project will be pure research which will show if there is already an existing harmonization system which is close enough to our use-cases to be adapted and applied in order to fulfil this projects purpose. Since there is no such working system in the field of water and wastewater treatment, the investigation will have to be expanded into familiar areas (like electricity supplement) and further into areas, where the field is completely different but data harmonization is also applied (like studies).

The following is a closer look at a project which attempted to harmonize data from very different sources. The article describes this attempt in a highly abstracted manner which could be used as a guidance for our project, since 4 steps out of 5 which the project needed to manage will be a problem in this system as well [7]. A closer look is required at this point in order to get an overview of what will be needed to be taken into account when making a time table and planning the system.

The Biobank Standardisation and Harmonisation for Research Excellence in the European Union aimed at pooling and harmonizing large amounts of population-based studies coming from six different European countries.
The attempt included a lot of communication between the studies resulting in a set of 96 variables describing questions of interest. The harmonization was assessed using:

- The studies questionnaires
- Standard operating procedures
- Data dictionaries

Possibly harmonizable data was processed in an open-source software and transformed from study-specific into the target format. The harmonized data from each centre was then placed on a centralized database for further analysis. The result was a generation of common format variables for 73% of matches considered (96 targeted variables across 8 studies).
The conclusion of this pilot project was: "*New Internet-based networking technologies and database management systems are providing the means to support collaborative, multi-center research in an efficient and secure manner. The results from this pilot project show that, given a strong collaborative relationship between participating studies, it is possible to seamlessly co-analyse internationally harmonized research databases while allowing each study to retain full control over individual-level data. We encourage additional collaborative research networks in epidemiology, public health, and the social sciences to make use of the open source tools presented herein*." [7]

<u>Problems which came up during the project:</u>

- Managing and harmonizing large amounts of data from different sources
- Ethical,
- Legal and
- Consent-related restrictions

<u>Benefits resulting from the harmonization:</u>

- Integrated data allows for lot bigger sample sizes
- Improved generalizability
- Easier ensuring of result validity
- More efficient secondary usage of existing data
- Provides opportunities for collaborative and multi-centre research
- Satisfies Governments, funders and researchers

A closer look into the harmonization process:

1. **Recruiting studies to participate in the project**
   The requirements for a participation were collection of specific data needed for the analysis. They studies were also required to allow remote access to the collected data. Another requirement for the studies was to make study metadata (questionnaires, data codebooks, standard operating procedures) and ethical and legal documents/policies available to the BuiSHaRE coordinating group. A standardized online description form found on the Mica-powered[1] BioSHaRE website. The cataloguing process was helpful for the understanding of the heterogeneity level between the study designs, as well as for the potential sample sizes available for the analyses.

2. **Defining a set of target variables**.
   The purpose of those variables was to answer specific research questions. This set of variables was the **Data Schema[2]** of the project and defined the common format of the different studies. A common schema is needed to work with multiple independent studies. Developing such a schema needs to be done carefully, since it requires a balance between uniformity and acceptance of a level of heterogeneity across the studies, meaning the same questions and data collections are likely to be phrased differently. Those variables were selected within two workshops organised for the studies, each of which targeted one specific *question which needed to be answered by the project*, and defining which variables would be needed to answer it. Each variable was described with 3 properties:
   - Variable(-name)
   - Definition
   - Format

3. **Defining the potential for each of the studies to generate each of the defined target variables**.
   This step included a deeper look into every study and determine their compatibility with the defined Data Schema. This not only included the value being present in a specific study, but also having a meaningful value – e.g. weight not being self-reported by a participant but instead being measured by a doctor. Having this restriction meant, that not all studies would be able to provide all 96 defined variables, but as much as 73% of the sought information. The difficulty in harmonization of data differed between the variables.

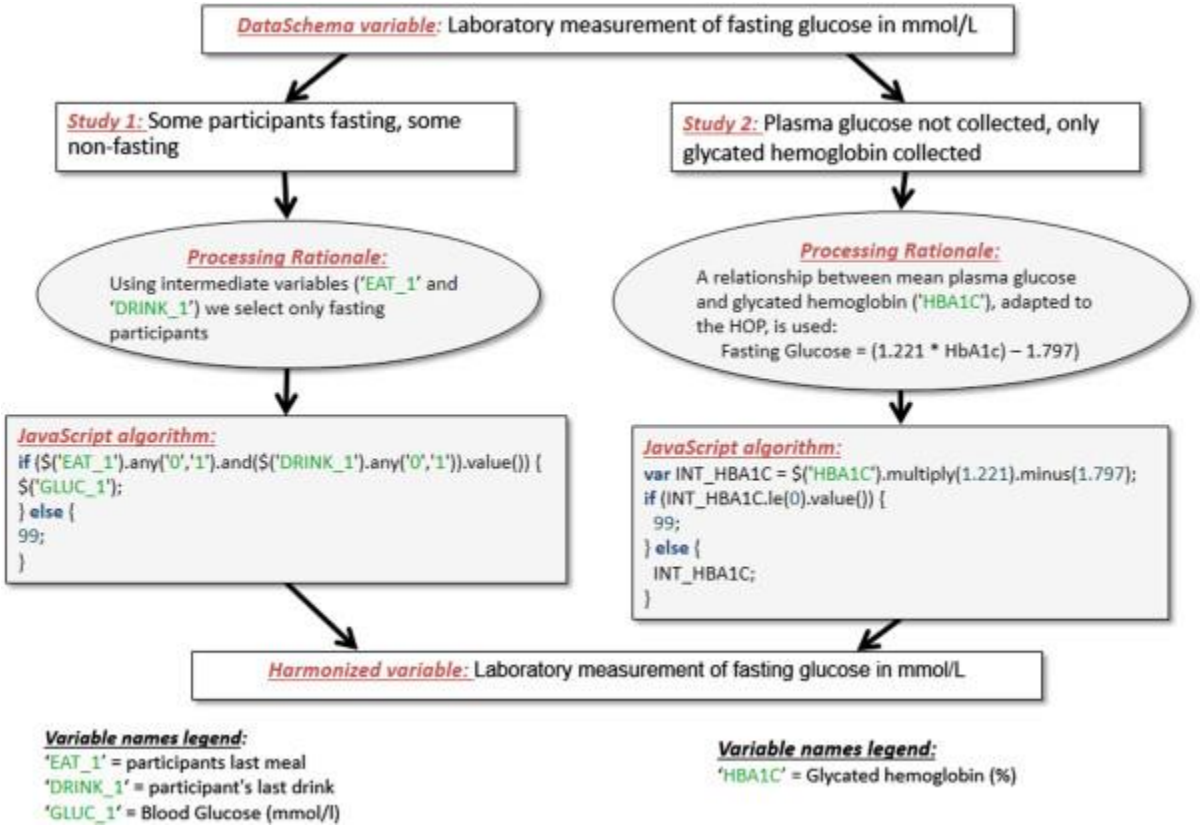
4. **Processing of the acquired data**
   For processing of the variables the, a software called Opal[3] was used. This software required data, needed to calculate values covered by the target variables to be extracted to dedicated Opal servers, as well as a reference to the DataSchema structure. Afterwards, the algorithms, calculating the harmonized variables data were defined for each of the studies. The derivation of each variable was completely independent of the derivations of the same variables in different studies.

[1] Mica [38] is a software application developed to create web portals for individual epidemiological studies or for study consortia. Including number of participants, their characteristics, methods of recruitment and so on...

[2] DataSchema: https://www.bioshare.eu/content/healthy-obese-project-dataschema

[3] "Opal is an software application used to manage study data and includes a software infrastructure enabling data harmonization and data integration across studies. As such, Opal supports the development and implementation of processing algorithms required to transform study-specific data into a common harmonized format. Moreover, when connected to a Mica-web interface, Opal allows users to seamlessly and securely search distributed datasets across several Opal instances." [Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4175511/#B1]

Example of a variable calculation:



(Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4175511/figure/F1/ last accessed: 28.09.2017)

5. Creating analyses using the acquired data

Analyses are a part of the partner project and will not be covered within this document.

Time-plan

	2017					2018		
	August	September	October	November	December	January	February	March
Interim Report	01.08	28.09						
Documentation			01.10					09.03
Define questions to be answered by the system			01.10-07.10					
Investigate service-provider offered data			04.10-14.10					
Define DataSchema			14.10-31.10					
Research for DataHost / Data Integration technique / technology				1.11-14.11				
Research for Data Harmonization technique / technology				1.11-14.11				
Research for Data Providing technique / technology				1.11-14.11				
Specification				14.11-30.11				
Architecture								
Implementation					1.12		28.2	
Testing							14.2	01.03
Integration								01.03-09.03

Interim Report: This document

Documentation: Master thesis including all research questions and all following documents

Define Questions: This step not only includes defining the questions which the system needs to be able to answer, but also finding out all different ways to answer those questions. A question might have different ways to get answers because a variable might have different ways to get calculated.

Investigate service-provider offered data:
Investigate on which service provider offers what kind and which amount of data, what the quality of data is and what the conditions are to be able to access this data.

Define DataSchema: Within this step the data schema will be defined. This is how the harmonized data will look like. It is important to see the target variable coverage by all suppliers, because if a variable can only be calculated by 5% of suppliers it might not be useful to support it.

DataHost Research: This step includes finding the best way to store the data. This contains researching the techniques used system solving similar problems in other fields. Not only does this step include finding the most fitting technology (i.e. Cloud), but also the most fitting technology provider (i.e. Amazon)

DataHarmonization Research:
This includes finding the most convenient way to convert the data from provided DataSchema to our own DataSchema. Again, comparing solution chosen by similar systems is a part of this step.

DataProviding Research:
This step is about finding the best way to provide the collected data to the partner project / 3rd party systems.

Specification: Creating a specification for the system. This document will define the scope of the system as well as the specific technologies and techniques which will be used to implement all system components. This document also includes the specification of external communication with the system.

Architecture: Creating the highest level of the architecture to define, which components of the system will interact with each other and with the outside world.

Implementation: Implementation of the system. (This step will be split in smaller tasks once the system, technologies and techniques are defined.)

Testing: Unit-/ and Black box tests, (Integrationtests)

Integration: Integration of the system and testing with the partner project

Deliverables or specific outcomes

The outcome of this project includes a solid fundamental research on 3 areas: The most fitting way to acquire and store data from public service providers, as well as from any historical data sources in terms of efficiency, pricing and considering any legal issues. This research also takes extendibility into account which is mostly about the adding of new data sources. The second research area is the data harmonization. This includes the most efficient and effective way of determining target variables and the data schema as well as the most appropriate technique and technology for converting the acquired data into the self-defined schema. The last and also least weighted research is on the best way to provide acquired and harmonized data.
Aside of the researches, this project will include at least a prototype of all the 3 mentioned modules, which should be able to pull data from a specified source, transform that data and provide the data for defined 3rd party systems. This prototype will be created the way, so that it also acts as a proof of concept for the research results.

Bibliography

[1] A. S. o. C. Engineers, Failure To Act - the economic impact of current investment trends in water and wastewater treatment infrastructure, 2005.

[2] W. UK, "http://www.water.org.uk," 28 09 2017. [Online]. Available: https://www.water.org.uk/about-water-uk/wastewater.

[3] W. UK, "https://www.water.org.uk," 28 09 2017. [Online]. Available: https://www.water.org.uk/consumers/what-water-companies-do.

[4] F. a. R. A. Department for Environment, "Sewage Treatment in the UK," Crown, 2002.

[5] U. Government. [Online]. Available: http://www.legislation.gov.uk. [Accessed 28 09 2017].

[6] W. UK. [Online]. Available: http://www.water.org.uk/about-water-uk/regulation. [Accessed 28 09 2017].

[7] P. B. Y. M. A. G. B. H. R. W. M. P. R. P. S. L. F. C. M. M. W. R. H. K. K. H. L. H. A.-M. Dany Doiron. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4175511/. [Accessed 28 09 2017].

[8] E. K. &. A. Mousavi, "WWTP-Global-300617," 2017.