Needed Skills for Utilizing Open Data in Business

Using Open Data as a Support of Business

HAMK
HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

Bachelor's thesis

Degree Programme in Business Information Technology

Visamäki – Hämeenlinna – Finland

Spring 2018

HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

Degree Programme in Business Information Technology
Visamäki – Hämeenlinna – Finland

Author Aleardo Schöni **Year** 2018

Subject Needed Skills for Utilizing Open Data in Business

Supervisor Lasse Seppänen

ABSTRACT

The purpose of this thesis was to introduce the topic of open data and to present the needed skills for utilizing open data in business. The paper was established for the Smart Service research unit in Häme University of Applied Science respectively for their project OpenHäme (Finnish: AvoinHäme).

The theoretical basis of the thesis consists of the research method, an introduction of open data, followed by a presentation of the data management process and the most common open data technologies and tools.

Based on the collected information the four tools ArcGIS, QGIS, PowerBI and Sisense were tested. Predefined evaluation criteria concerning technical, business and training aspects were used as a framework. Afterwards, a short field report on the experiences of working with the tool was established.

The first outcome of the research work was an open data matrix including open data categories and the data management process steps. This matrix can be used as an overview and initial position to work with open data. Furthermore, the comparing showed that open data can be processed with open source tools just as good as with commercial tools.

The results of the thesis indicate that there are enough open source tools existing to work with open data. By following the instruction and respecting some rules working with open data is possible for everyone.

Keywords open data, open government data, linked open data, data management process, open data business

Pages 55 pages including appendices 1 page

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In 2016 the direct market size of open data of the EU28+ countries was estimated to be 55.3 bn Euro. In the following 4 years, until 2020 an increase of almost 37% to a value of more than 75 bn Euro is expected. (Carrara, Chan, Fischer, & van Steenbergen, 2015)

The United Nation (UN) under-secretary General Sha Zukang stated: "The UN-system has accumulated over the past 60 years an impressive amount of information. […] not only for desk of decision makers and analysts, but also to journalists, to students and to all citizens of the world." (United Nations Statistics Devision, 2018).

Tremendous numbers like these indicate that open data is not anymore a phantom of IT driven businesses but rather a valuable opportunity for everyone.

This thesis is established for the smart service research unit in Häme University of Applied Sciences (HAMK) and takes part in the project OpenHäme[1], which aims to create regional prerequisites for the usage of open data in business (Finnish: AvoinHäme – Edellytykset avoimen datan hyödynämiselle liiketoiminnassa). OpenHämes goals are to create the conditions for exploiting open data in business, increase the participants' knowledge and skills regarding the use of open data, combining active players in Kanta-Häme and testing and piloting various open data applications. (HAMK, 2017a)

To provide information and create knowledge for the project, three research questions are answered throughout the paper: What are key technologies and tools needed to exploit open data? Which skills are needed to handle those technologies and tools? How should companies be taught in or acquire these skills?

To answer these questions a research method is defined and introduced after this section. Then the term of open data itself is explained. Afterwards, the so-called data management process on how to turn raw data into knowledge is presented. Furthermore, a selected set of technologies and tools are introduced. Based on this knowledge four open data tools were tested and the results as well as a comparison are presented. The second last chapter introduces a matrix, which summarises the previous gathered information in one figure. The last chapter provides a conclusion and recommendations for the project team as well as companies interested in open data.

[1] https://xn--avoinhme-5za.fi/

2 RESEARCH METHODE

This thesis takes part in the bachelor's degree programme in business IT at Häme University of Applied Sciences in collaboration with Bern University of Applied Sciences (BFH) in their double degree programme.

The organization of the thesis follows the thesis guide provided by HAMK. The key principals are the workplace-orientation, to provide a concrete and feasible topic for the customer as well as the student and to promote and advance the development of the students' professional competences while working for or with a client. (HAMK, 2017b)

The thesis follows the practice-based method and is a workplace development assignment. The goal is to provide a knowledge base to the customer. To fulfil these goals the implementation consists of a description of the knowledge base, as well as the implementation and results of the thesis, which is then summed up in a conclusion. (HAMK, Thesis Guide, 2017b)

2.1 Purpose

The focus of this thesis lies on the technical aspects of open data. What businesses and companies might do once the data is transformed into knowledge is however not part of this thesis.

Furthermore, the thesis focuses especially on open data that concerns or influences the sections of tourism and events, traffic and logistics as well as bio economy. This includes for example the weather, national statistics about tourism, Finnish maps and other different kinds of maps.

The idea is to list a set of skills needed by businesses to work with open data and how they can learn or acquire these skills. The aim of the thesis is to serve and help the project team as well as in open data interested businesses as an initial position, guideline and background information.

2.2 Knowledge Base

The chapters 3 Open Data, 4 Data Management Process and 5 Technologies & tools present the basic knowledge concerning open data. Based on the content of these chapters the practical implementation is conducted.

2.3 Implementation

Throughout the implementation four different tools were tested. To get reliable results a defined set of requirements served as base for the tests. After finishing these tests, a set of specifications for each tool was

gathered, which allowed to compare commercial and open source tools to each other, so a set of recommendations could be established.

As result the thesis provides an introduction to the topic of open data. Furthermore, a clearly structured process is presented and common tools and technologies to handle open data are introduced. Finally, all information is shown in a matrix, which gives the reader a clear overview.

3 OPEN DATA

According to opendatahandbook.org open data became interesting to the public in 2009. Around this time multiple governments (e.g. USA, UK, Canada and New Zealand) announced projects to publish their public data and therefore make it accessible for anybody. (Open Knowledge International, 2018d)

Opendefinition.org defines the term open knowledge as follows: “Knowledge is open if anyone is free to access, use, modify and share it.”. (Open Knowledge International, 2018f)

But what is open data and what exactly does it include? In this chapter a definition of open data is given and steps to achieve open data are listed. Afterwards the term of open government data is explained. Furthermore, linked open data is introduced and finally a brief list of institutions working in the open data industry are presented.

3.1 Definition

The term open data consists of two words. First the word “open”, which according to Murray-Rust is used for instance in “open source” for software and means free or libre. This means anyone has the right to use the software or the data in case of the open data. (Murray-Rust, 2008)

A definition of open source is something “[…] that can be used, studied, and modified without restriction, and which can be copied and redistributed in modified or unmodified form either without restrictions, or with restrictions only to ensure that further recipients can also do these things. […].” (Murray-Rust, 2008). This definition can analogous be used for open data.

According to opendefintion.org, knowledge that is being transferred is called work. But before something can be called open work it has to satisfy some requirements. First, the work must be under an open licence or denote the absence of copyrights and similar restrictions. Second, the work must be provided as a whole and should be downloadable in the Internet without charges. Third, the work must be provided in a form that a computer can read and understand so that the data as well as individual elements in the data can be easily accessed and modified. Finally, the work must be provided in a format that places no restrictions, monetary or otherwise, upon its use and can be fully used with free/libre/open-source software tools. (Open Knowledge International, 2018f)

Also, the licenses protecting the work need to be compatible with other open licenses and satisfy some restrictions. The license must allow free

use, redistribution, modification, separation, compilation, non-discrimination, no additional legal terms and no charge of the work. Data or work satisfying all the above-mentioned requirements can be called open. (Open Knowledge International, 2018f)

3.1.1 From Data to Open Data

Open Knowledge International recommends in their open data handbook three key rules on how to open data up, respectively how to transform data into open data.

The first rule is to keep it simple. Every start is difficult. It is therefore recommended to start small, simple and fast. In a first step, only one dataset or a smaller part of a big dataset can for instance be provided to the public. It is not advisable to make all the data public in one step. Nevertheless, the goal should always remain to open as many datasets as possible. (Open Knowledge International, 2018c)

The second rule is to engage early and often. It is very important to engage potential users like citizens, businesses or developers. It is also important to re-use data as early as possible. This ensures that the offered services stay up to date. Not all the data will reach the user directly. Rather the data will first pass via info-mediaries, who then transform the data into knowledge. Only afterwards this knowledge is transferred to the user. (Open Knowledge International, 2018c)

The third and last rule is to address common fears and misunderstandings. It is quite common that some question and fears might show up while working with open data in a large institution such as a government. In those moments it is essential to identify the most important issues and to address them as soon as possible. (Open Knowledge International, 2018c)

Keeping those three key rules in mind, data can be opened with the steps presented in the next four subchapters.

3.1.2 Dataset Selection

The title of this chapter already summarizes what has to be done. Nevertheless, it is important to know that the whole process of opening data up is iterative and it is therefore normal to come back to this first step.

If it is already decided which dataset should be opened the process continues with the next step in chapter 3.1.3. However, in most cases especially big institutions have problems with this first step. In this case it might be helpful to identify datasets, which could be interesting for the

company to publish. As it will be explained, not every dataset is suitable to be opened. (Open Knowledge International, 2018a)

Afterwards, this list of possible datasets can be presented to the community. The community consists of the people who would actually work with the data and it is easier for them to identify possibly valuable datasets. Additionally, the cost basis can give further insights. How expensive is it to collect this data? It is likely that the data which is more expensive to collect is more interesting to the public. The ease of release plays another important role. In the beginning small releases might be useful as a catalyst for larger behavioural changes. (Open Knowledge International, 2018a)

Finally, it is always good to be informed what other people do. A list of what exactly the different agencies or departments are doing could give further insights on the importance of the data they produce. (Open Knowledge International, 2018a)

3.1.3 Applying Licenses

All available open data should have a license on it, simply because it helps to clarify upcoming issues even though sometimes it is not needed. Open Knowledge International offers a suitable kind of licence for open data. (Open Knowledge International, 2018a)

3.1.4 Data Availability

The process of making data available includes some simple steps. First of all, the data should be available if possible as a free download in the Internet or at a price no more than the costs of the reproduction. Moreover, open data should not be used as a business and agencies should not undertake any costs. Furthermore, the available data should be complete. Finally, for greater re-use, the data should be available in a machine-readable format. This means raw data is more valuable than for instance statistics in a PDF (Portable Document Format). (Open Knowledge International, 2018a)

3.1.5 Make Data Discoverable

Once the data is openly available it needs to get used. This means the data must be findable for people. It is important to know that the more famous a dataset is, the more likely it is that new and useful tools will be built to analyse the dataset and finally find new business insights. (Open Knowledge International, 2018a)

3.2 **Open Government Data**

Not only is open data getting more and more interesting to the public, but also the data produced by a government gets more attention. This so-called open government data led to a movement, which aims to define regulations to support a set of rights and obligations that promote transparency, accountability, participation and collaboration within the government and the public sector. It is also in the government's own interest to promote this culture of openness and transparency. (Luna-Reyes, Bertot, & Mellouli, 2014)

The publication of open government data leads to more transparency within the government. Citizens are able to see what exactly the government is doing. But transparency also means sharing and reusing data. Interested individuals are able to use open government data to analyse and visualize their governments data and might find useful insights. This automatically leads to an increase in participation and engagement of citizens. Instead of only once in years, citizens are constantly able to get in contact with their government. (Open Knowledge International, 2018e)

Furthermore, open government data can influence ones personal life in a positive way. For example, Tine Müller from Denmark created the website findtoilet.dk. Based on open government data she created a map, which shows all the Danish public toilets. Thanks to this idea people with bladder problems can easily find the next public toilet and therefore increase their quality of life. (Open Knowledge International, 2018e)

Additionally, open government data has a keen influence on the economy. According to different studies open data has an economic value of multiple billions of Euro annually in the EU only (Open Knowledge International, 2018e). Based on open data companies can build new products or institutions are able to improve their services.

The website husetweb.dk offers its users ways to find out how to improve the energy consumption of their homes. Based on re-using information of maps, government subsidies and the local trade register it is also able to help with the financial planning and finding suitable builders to do the work. Another example is Google Translate, which uses EU documents that are available in different languages to train their algorithms and thus improve their translating service. (Open Knowledge International, 2018e)

Beside citizens and the economy, the third party that benefits from open government data is the government itself. For example, the Dutch Ministry of Education started publishing their education related data online. Since then the amount of questions asked dropped and money as well as time was saved while the governments efficiency increased. The Dutch department of cultural heritage is also releasing their data and

works together with different amateur historical societies and groups. This enables them to execute their tasks more effectively. As a result, the quality of their data is better and employee costs can be saved. (Open Knowledge International, 2018e)

Today it is known that open data creates social and economic value in numerous ways. In the future it will even be possible that new knowledge and insights are found by combining multiple different data sources. It is normal that unexpected insights are unleashed by combining different open government data. But those insights can only be found if open data is available. This means that governments have to publish their data and make it available without restrictions. Only then the economy and private organs can use this data and generate value. (Open Knowledge International, 2018e)

3.3 **Linked Open Data**

As the name indicates linked open data is open data that is linked. The origin of this term lies in the semantic web. In contrary to the web of hypertext the web of data consists of links that machines are able to comprehend. Those links form the relationship between random things described by RDF (Resource Description Framework). (Berners-Lee, 2006)

Tim Berners-Lee defined four simple steps to link data. First, URIs (Uniform Resource Identifier) should be used as names. Second, those URIs should include HTTP (Hypertext Transfer Protocol) so that people can look up those names. Third, when someone looks up a URI, useful information should be provided using standards like RDF and SPARQL. Finally, other links to URIs should be included so that user and machines can access further information. (Berners-Lee, 2006)

According to Tim Berners-Lee a surprisingly big amount of data was not linked in 2006. One of multiple reasons of this problem is that the data is not linked correctly (Berners-Lee, 2006). To encourage people and government data owners to publish their open linked data, Tim Berners-Lee introduced the 5-star principle. Figure 1 shows this scheme to deploy open data. (Hausenblas, 2015)

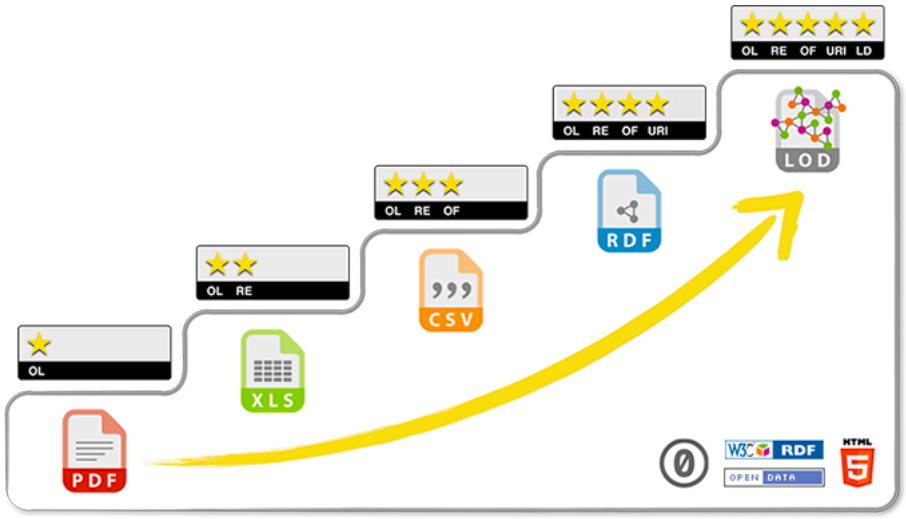

Figure 1 - 5-Star-Developmet-Scheme (Hausenblas, 2015)

The following table serves as explanation to Figure 1. By following those five simple steps open data can easily be transformed into linked open data.

Table 1 - 5 Star-Development-Scheme description

Stars	Description	Example
1	Your data is available on the web with an open license	PDF on website
2	Your data is in a machine-readable format	Excel instead of PDF
3	as 2-star plus non-proprietary format	CSV instead of Excel
4	All previous stars plus usage of and open standard from W3C. Use RDF and SPARQL to identify your data, so others can link to your data	RDF and SPARQL
5	previous stars, plus link your data to other data and provide context	Link to related files & information

In five simple steps data can be turned into linked open data and be re-used for further benefits.

3.4 Institutions

In this chapter a selection of Finnish as well as international institutions and organization is given. All of them play an important role in the open data sector. They provide useful information, offer open data or are a contact point for open data users.

3.4.1 Open Knowledge International

This global non-profit organization helps civil society groups to access and handle open data. They show the value of open data and provide useful tools and skills to organisations. Open Knowledge International with its

international network aims to unlock information and to create as well as share knowledge. (Open Knowledge International, 2018h)

3.4.2 Open Knowledge Finland

OKFI is a non-profit organisation and takes part in the international network of the Open Knowledge Foundation. The goal of the organisation is to promote the opening and usage of open knowledge and to advance the development of the open society based in Finland. It was founded in 2012 and to date it counts more than 400 members including individuals, companies and other organizations. It also offers a wide network to connect people interested in open data. (Open Knowledge International, 2018g)

3.4.3 World Wide Web Consortium

"The World Wide Web Consortium (W3C) is an international community where Member organizations, a full-time staff, and the public work together to develop Web standards. […] W3C's mission is to lead the Web to its full potential." (W3C, 2018).

W3C plays an important role in open data, more specific in linked open data. As mentioned in chapter 3.3, W3C plays an important role for making open data even more powerful. Tim Berners-Lee the inventor of the web and director of W3C is still creating new knowledge from his ideas. (W3C, 2018)

3.4.4 EU Open Data Portal

The EU Open Data Portal was founded in 2012 based on the decision of the European Commission on the reuse of the Commissions documents. It offers a wide range of data from the European Union (EU), their institutions and other bodies. The data can be used freely for commercial or non-commercial purposes. (Publications Office EU, 2018)

The goal of the EU Open Data Portal is to provide free and easy access to the data. They want this data to be reused and turned into innovative benefits and create economical potentials. Additionally, the portal should help to keep and extend the EU institutions transparency and accountability. (Publications Office EU, 2018)

All EU institutions should provide their data free of charge and without any copyright restrictions for use. Already available are for example a standardised catalogue to make the access to EU open data easier, apps and web tools to reuse this data, a SPARQL endpoint query editor, REST API access and tips how to make the best use of the webpage. (Publications Office EU, 2018)

3.4.5 DataBusiness.fi

DataBusiness.fi is a webpage boosting data-driven businesses in Finland. Their main work is to enhance the usability of open data in business operations. Free business and development services concerning open data are offered for companies and communities. Additionally, they offer education and training for any interested individual, company or entrepreneur. (Sibelius, et al., 2018)

3.4.6 Avoindata.fi

Avoindata.fi offers different interoperability tools for Finnish open data. One of their services is to provide standards and guidelines (Avoindata.fi, 2015). In a few simple steps the full service can be used. First, a user account has to be created. Afterwards, different organisation can be joined or a non-public organisation can be created. After those steps data can be published. To support this step, avoindata.fi provides an API as well as an open source license. (Avoindata.fi, n.d.)

3.4.7 The World Bank

The World Bank is an organization that maintains macro, financial and sector databases. Their main goal is it to help governments to develop processes to provide useful information to citizens and show them what they do. For example, they help countries to improve national statistical systems. (The World Bank Group, 2018)

They also want to provide good quality and integrity of data. Based on an international system containing the United Nations (UN), the Organisation for Economic Co-Operation and Development (OECD) and the International Monetary Fund (IMF), they want to improve and develop frameworks, guidance and standards for statistics. (The World Bank Group, 2018)

Additionally, they are building consensus and define agreed indicators and also establish data exchange processes and methods. Finland is a part of this organization since 1948. (The World Bank Group, 2018)

3.4.8 UNdata

The United Nations Statistics Division (UNSD) provides through data.un.org the UN statistical databases. They combined different proprietary databases into an openly available database on the Internet. Consequently, data can be found by browsing or searching with the help of a keyword. (United Nations Statistics Devision, 2018)

In total more than 60 million data points concerning agriculture, crime, education, employment, energy, environment, health, HIV/AIDS, human development, industry, information and communication technology, national accounts, population, refugees, tourism and trade are available. (United Nations Statistics Devision, 2018)

This database thus supports the mission to ensure the development of the statistical system around the globe and advance the dispersion of statistical information. (United Nations Statistics Devision, 2018)

3.4.9 European Data Portal

"The European Data Portal harvests the metadata of Public Sector Information available on public data portals across European countries. Information regarding the provision of data and the benefits of re-using data is also included." (European Data Protal, 2018).

The mission of the European Data Portal is to improve the accessibility and increase the value of open data. This portal offers users to search datasets, provide data, use data and use eLearning modules. (European Data Protal, 2018)

Energy, transport, economy & finance, education, health and science & technology are only a few of the categories of open data, which the portal offers. (European Data Protal, 2018)

4 DATA MANAGEMENT PROCESS

Data is not equal to information or knowledge. Simple letters and numbers do not provide any advantages for businesses. The data first needs to be transformed into information, which afterwards can be turned into knowledge.

This chapter provides definitions of data, information as well as knowledge and also introduces the data management process as well as the data value chain and presents an example of open data usage in tourism.

4.1 **Definitions**

To benefit from open data different stages must be understood. This chapter therefore provides definitions of the different maturity that data can have within the data management process.

4.1.1 Data

According to Hey data has various definitions. Data is for example defined as unprocessed information. Furthermore, in IT (information technology) terms like data stream and packets of data are used. Besides all these different definitions, data can be used for the same purposes. It can be stored, piled-up, recorded, manipulated, captured and retrieved. Usually, data is unstructured and can be described as a bunch of numbers and letters. Therefore, data can get overwhelming quite fast and people tend to get lost. However, with the right skills data can be mined and transformed into information. (Hey, 2004)

4.1.2 Information

Compared to data information is a specific selected amount of data. This means that there are defined boundaries and relationships within the data. Information still contains data but is now structured. For example, the numbers and letters are now structured in a table. (Hey, 2004)

4.1.3 Knowledge

Knowledge in relationship to data and information does not refer to the knowledge that people can possess. The term of knowledge is rather separated in tacit and explicit knowledge. On one hand there is tacit knowledge, which is the one that is in the heads of people. It is hard to extract it and to give it to another person. On the other hand, explicit knowledge is the knowledge that can be encoded. This means the knowledge comes e.g. from a presentation or reports. They can be

duplicated as well as shared and the knowledge can therefore be transferred. (Hey, 2004)

4.2 **Process**

Once open data is available it should be used. The tremendous amount of open data available offers different possibilities to profit from it. But before benefits can be used the data has to be transformed. In separated steps, the raw open data has to be transformed into information. This information can afterward be used to create knowledge.

Figure 2 shows the process of how raw data is turned into knowledge. The first and most important step is to define the knowledge the user is looking for. This means that first of all the goals for the end user needs to be defined so that specific questions can be asked, which will finally answer to the end users' needs. Afterwards, the other proposed steps can be executed. (Santoso & Lamoree, 2000)

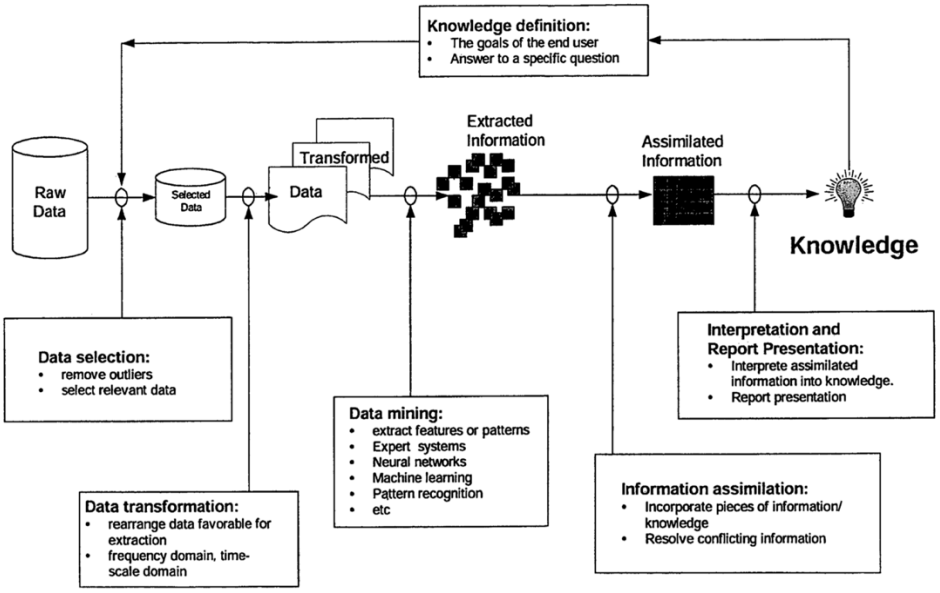

Figure 2 – The process of turning raw data into knowledge (Santoso & Lamoree, 2000)

Once the knowledge the user is looking for is defined he can start with the data selection. This means the relevant part of the data is selected and outliers are removed. This now selected data must be transformed. The transformed data has then to be rearranged for extraction. From this transformed data information can be extracted through data mining. Techniques like extracting features or patterns, using expert systems, neural networks, machine learning or pattern recognition are different ways to find the needed information. From this now so called extracted information, new information has to be assimilated. This includes incorporating pieces of information and knowledge, as well as resolving conflicts in the information. At this point assimilated information is created. This information has now to be interpreted and the results

should be presented in a report. Out of this presentation new knowledge can be gained. (Santoso & Lamoree, 2000)

Like Santoso & Lamoree, the European Commission presented in their paper Creating Value through Open Data a similar process (Figure 3). As already shown in Figure 2, the process includes three different stages: First data, then information and finally knowledge.

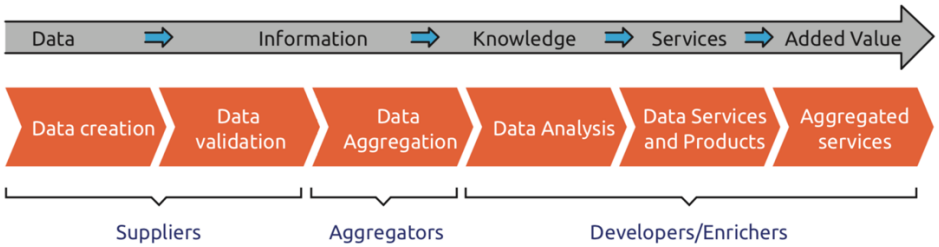

Figure 3 - Data Value Chain Archetypes (Carrara, Chan, Fischer, & van Steenbergen, 2015)

Compared to Santoso & Lamorees version of the process, Carrara et al. added the states of service and added value. Those can be understood as the result of applying the created knowledge. (Carrara, et al., 2015)

Furthermore, three different so-called archetypes play a role. Suppliers create data. Additionally, they validate the data. Aggregators are organizations that collect the aggregated data. Companies and individuals that analyse the data, create a product or services are called developers and enrichers. The differences between them are their functions. Developers on one hand are active in the private sector and design, build and sell web or mobile application to transfer open data to customers. Enrichers on the other hand, use open data to get new and advanced insights for their clients. (Carrara, et al., 2015)

4.3 **Data Value Chain**

There are different types and ways to re-use data. The data value chain is a way to elaborate those types. The goal is to get the biggest commercial and non-commercial value out of the data. (Carrara, et al., 2015)

The orange arrows in Figure 4 presented the main flow of the value chain. The chain starts with the creation of data. This can come from the public sector but also from private organs. Afterwards, the data has to be validated. Once those two steps are done the data is aggregated. At this point the data can be published and made available for anyone so that it can be used. Now people and organizations can analyse this data. Thanks to this step new data services and products can be offered. In the final state there are the aggregated services. During this value chain the data has to be stored and preservated. (Carrara, et al., 2015)

The two arrows on top, public sector and private sector, indicate in which way those two entities participate in the data value chain. Often the

public sector provides the data and the private sector first uses the aggregated services. Carrara et al. even mention with the blue arrows on the bottom, that “data creation” is a public task and the “data services and products” as well as the “aggregated services” are on the (end-) user side. The steps of validation, aggregation and analysing are executed by public content holder as well as re-users. (Carrara, et al., 2015)

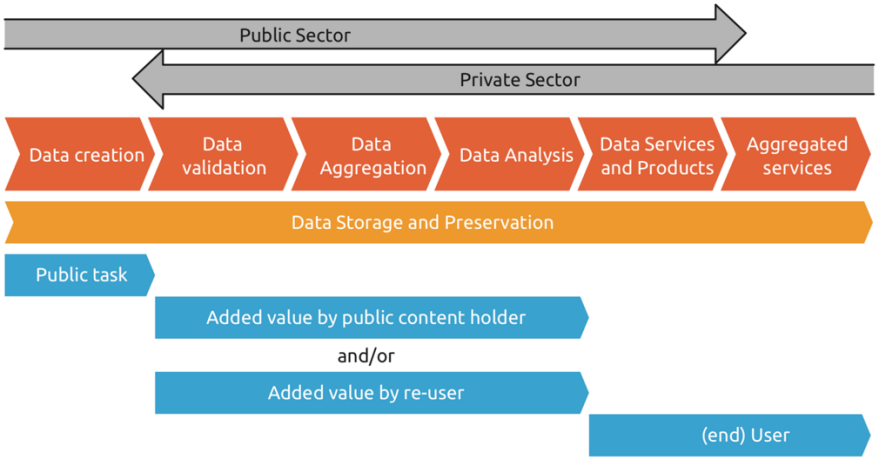

Figure 4 - Data Value Chain (Carrara, Chan, Fischer, & van Steenbergen, 2015)

The needed platforms and technologies are provided by the so-called enablers. Interested individuals and organizations are free to use this service. Enablers play an important role in the data value chain due to the fact that they generate revenue with their work but provide both cost-effective and easy-accessible services for data suppliers and consumers. (Carrara, et al., 2015)

4.4 Example in Tourism

Elenora Pantano, Constantinos-Vasilios Priparas and Nikolaos Stylos presented in their paper “‘You will like it!’ using open data to predict tourists’ response to a tourist attraction” a technique how to collect open data and transform it into predictive information. (Pantano, et al., 2017)

Based on TripAdvisor reviews, two experiments were executed. To simplify the method, only the best (5 stars) and the worst (0 stars) reviews were chosen. Afterwards, the 5 stars were transformed into a 1 and the 0 stars into a 0. (Pantano, et al., 2017)

For the first experiment, 50 classifying functions were build. The goal was it to identify the true value of 1 and 0. Based on the Random Forest method the actual percentage of success of number 1 was 0.66 and the percentage of success of 0 was 0.57. The slight bigger true value of 1 indicated that the chance is higher that the attraction gets a good review instead of a bad one. (Pantano, et al., 2017)

For the second experiment 200 classifying functions were build. As in experiment 1 the goal was to identify the percentage of success for 1 and 0. The best prediction showed the result that 1 is 0.69 and 0 is 0.58. This means in the second experiment the results were slightly better due to the bigger number of data available. (Pantano, et al., 2017)

The authors showed in their paper that it is very well possible to use open data to create business values or even predict the future to a certain extend. (Pantano, et al., 2017)

5 TECHNOLOGIES & TOOLS

Open data cannot be processed without the help of different technologies and tools. In a first part of this chapter the different categories of open data are introduced. Afterwards, a differentiation between static and dynamic data is given. Finally, a selection of open data formats is presented and explained.

5.1 Types of Open Data Sources

There are different sources of open data. The W3C consortium defined eight different categories.

First of all, geo data is used to create different kind of 2D and 3D maps. They include the locations of buildings and roads as well as information on the topography and boundaries (Vathana & Audsin, 2013). Geo data usually contains geographical locations in a format, which is compatible with a geographic information system (GIS). The information can be stored in a database, geodatabase, shapefile, coverage, raster image, dbf table or Microsoft Excel spreadsheet. GIS normally offer analytical functions like data analysing, filtering, visualization, loading, management and publishing. Furthermore, geo data comes in various formats. Normal GIS support most of the commonly used formats. (Esri, 2017a)

Cultural data hold information about cultural works and artefacts of countries. The holders of this kind of data are normally galleries, libraries, archives and museums (GLAM) (Vathana & Audsin, 2013). Another important open data category is the science. The data produced as a result of scientific researches are important open data sources due to their value for further researches. (Vathana & Audsin, 2013)

The financial sector is also a known open data producer. Governmental expenditure and revenue as well as information on the financial markets like stocks, shares and bonds can be useful as open data. A lot of governments also publish a big amount of statistical data. Companies and people interested in this data are offered to use it and get insights out of it. Data concerning the weather is collected daily from different sensors and satellites to predict the weather and climate. There are multiple ways to get this data on the web. (Vathana & Audsin, 2013)

Environmental data contains information related to the natural environment and its topics like pollution, rivers, seas, mountains, volcanos and so on. Finally, governmental data is one of the biggest data pools. Many governments are interested to open their data and therefore provide more transparency concerning their plans and policies to the public. Furthermore, they also hope to get some benefit out of

their data and that interested people get valuable insights out of the governmental data. (Vathana & Audsin, 2013)

Most of the time it does not make sense to just analyse one type of open data. The key to create valuable business insights lies in the combination of different data sources. For example, touristic organizations could combine geo data with weather and environmental data to create a new service for their costumers and offer them more information.

5.1.1 Static and Dynamic Data

The different categories can further be separated into two data types. On one hand there is static data. This means that the data is stored somewhere in a database. This kind of data does not change frequently and rests in its form. Information and knowledge created out of this data is mostly based on a historical data collection. For example, information on the expenses of a country are static data.

On the other hand, there is dynamic data. This kind of data is also known as data stream. This means that the data is collected from some kind of sensor and then transported to a system, which analyses the data in real-time. In other words dynamic data produces information and knowledge out of the actual situation. For example, information gathered by a weather sensor, which is then sent to a website who shows the actual temperature of a region is dynamic data. (Amazon Web Services, 2017)

The analysis of dynamic data makes sense when there is new data created and collected continuously. A lot of companies collect their own data and analyse it in real-time. In the open data scene it is harder to find a data stream that is open. (Amazon Web Services, 2017)

A real live example of an open data stream usage offers the city of Pittsburgh, Pennsylvania. The US American city tracks their snow removal vehicles and show their activity on a map. People interested in this data can then analyse, which roads have been cleaned from the snow and are safe to drive on and those who are not. (Moren, 2015)

There are multiple data tools that already support and integrate dynamic data analysis. Another common way to process data streams is with API and REST API. An API is an application programming interface. They are used in application development and enable the communication between various software components (Wikipedia Foundation Inc., 2018). REST API stands for representational state transfer API and is used for distributed hypermedia systems and also data streaming. (Fielding, 2000)

5.2 **Formats for Open Data**

Regardless of the category or data type, open data comes in different formats. This section presents the most common and most used formats users might face when working with open data.

5.2.1 JSON

JavaScript object notation (JSON) is text based and is used as a syntax for storing and exchanging data. Any programming language can use JSON to send its data to a server. Furthermore, JSON is described as lightweight data-interchange format and is self-describing as well as easy to understand. (Refsnes Data, 2018a)

While exchanging data between a browser and a server, only text can be send. For this kind of data transfer JSON offers a simple way to upload and download open data to a server. (Refsnes Data, 2018a)

5.2.2 XML

XML stands for extensible markup language. Like JSON XML is for storing and transporting data. As a software- and hardware-independent tool XML has no defined function. It just adds information wrapped in tags. Programs can be written to understand this information. XML simplifies data sharing, transport, platform changes and availability. (Refsnes Data, 2018b)

Data in incompatible formats can be exchanged like this between systems. XML stores the data in plain text format and is software- and hardware-independent. So incompatible data will not be lost. Furthermore, XML can be used to upgrade data to new operating systems, new applications or new browsers. (Refsnes Data, 2018b)

5.2.3 Resource Description Framework

A RDF is used to combine data from multiple sources and can be stored with XML and JSON. As identifiers URLs are used to offer interconnection between existing open data. (Open Knowledge International, 2018b)

5.2.4 Spreadsheet

Spreadsheets are for example Microsoft Excel-files. This is a widely known data format. Normally, the data can be used immediately as long as the columns are named and a precise description is given. If there are any further macros or calculation on the spreadsheet it is recommended to document them to make it easier to work with the data. (Open Knowledge International, 2018b)

5.2.5 Comma Separated Files

CSV is a compact data format to transfer large datasets with the same structure. A CSV-file should always be expanded with an accurate documentation file, otherwise it is almost impossible to understand the dataset. The structure of the data is also significant because a single error can impact the interpretation of a whole column. (Open Knowledge International, 2018b)

5.2.6 Text Documents

Text documents can show certain types of data through formats like Word, ODF, OOXML or PDF. However, this kind of format supports no structure and it is hard for machines to understand the data. It is recommended to use a typography markup technology whenever possible. This makes it easier for machines to separate for example between headings and body text. (Open Knowledge International, 2018b)

Plain Text files are normal text (.txt) files. Those are very easy to read for computers. Due to the fact that those files have no structure a parser has to be created to interpret the files. Problems may appear when plain text is switched between different operating systems. (Open Knowledge International, 2018b)

5.2.7 Scanned Images

Scanned images are the least preferable data format. Only in special cases like historical or archival material pictures make sense. Some types at least support the possibility to add a documentation to the picture. (Open Knowledge International, 2018b)

5.2.8 Proprietary Formats

Proprietary formats are formats that are dedicated to a defined system. If data is only used in similar systems, this format works. However, if someone wants to use the data without the system, problems might appear. The data should at least always be linked to the suitable system or better use a non-proprietary system. (Open Knowledge International, 2018b)

5.2.9 HTML

HTML stands for Hyper Text Markup Language. It is mostly used to display content on websites. HTML is a cheap and fast way to display data. If the data is stable and of limited scope it is sufficient. The HTML

format is also the easiest to download and manipulate. (Open Knowledge International, 2018b)

5.2.10 Further formats

Additionally, to the different formats, there are two more. Open files and closed files. Open files on one hand are using formats, which are available for everybody. Closed files on the other hand use proprietary formats. The advantage of open file formats over closed ones are that developers are able to provide useful software to handle the data. This automatically leads to a decrease of obstacles on reusing the data and improved use as well as results. (Open Knowledge International, 2018b)

5.3 **Tools**

Open data often includes a huge amount of data, which cannot be handled by hand anymore. That is why a broad set of different tools exists. Those tools differ in functionalities and purposes. Some tools are just to collect data, others are just to visualize data. By combining different tools, the whole data management process can be covered. This chapter presents a selection of tools to work with open data.

5.3.1 Microsoft Excel

One of the simplest tools to work with open data is Microsoft Excel. This proprietary spreadsheet program enables the user to handle data in form of workbooks and cells or columns as well as rows. (Microsoft, 2018b)

Once data is included charts and analyses can be made. Excel offers different models, which make it easy to handle and produce useful information. Furthermore, Excel offers a set of formulas and functions to make calculations with the data. (Microsoft, 2018b)

One of the disadvantages of Excel is the limited number of rows and columns supported. It is possible that a dataset is too big for Excel to handle. In this case Excel offers the functions of PivotTables and PivotCharts. Those special functions make it possible to store datasets decentralized and use a link to connect the data with the functions. (Microsoft, 2018b)

5.3.2 Power BI

“Power BI is a suit of business analytics tools that deliver insights throughout your organization. Connect to hundreds of data sources, simplify data prep, and drive ad hoc analysis. Produce beautiful reports, then publish them for your organization to consume on the web and across mobile devices. Everyone can create personalized dashboards with

a unique, 360-degree view of their business. And scale across the enterprise, with governance and security built-in." (Microsoft, 2018g).

This short presentation of Power BI summarizes his main functions. Power BI can be used by analysts, business users, IT people and developers. Furthermore, Power BI offers connections to useful data storage services like Google Analytics, Azure SQL Database, Web Pages, MySQL, Oracle and Salesforce just mentioning some. (Microsoft, 2018g)

5.3.3 Social Engagement

Social Engagement is a tool to collect data on social media. It is part of the ERP (Enterprise Resource Planning) tool Microsoft Dynamics (Microsoft, 2018a). The goal is to connect standard ERP processes with data collected from social media. This means important insights of the clientele are faster integrated in the companies' system. Thus, the sales department can use social selling to build credibility. The marketing employees can scale their brands reputation. Basically, anyone in the company can get social insights out of the customers and get a better understanding of their needs. This finally leads to an optimization of the offered services and a higher customers satisfaction. (Microsoft, 2018h)

5.3.4 Yahoo Query Language

YQL is an interface which allows the user to query, filter and combine data from the Internet. The tool makes it possible to transform basic HTML into reusable data and even allows to create an API where non-exist. Furthermore, simple CSV files can be loaded and used. The idea behind Yahoo Query Language is to mix and match data from different sources to extend open data. With just a few code lines, YQL can be integrated in applications for commercial or non-commercial use. (Yahoo Developer Network, 2018)

5.3.5 ArcGIS

"ArcGIS Desktop is the key to realizing the advantage of location awareness. Collect and manage data, create professional maps, preform traditional and advanced spatial analysis, and solve real problems. Make a difference and add tangible value for your organization, your community, and the world." (Esri, 2018i).

ArcGIS is a tool processing geo and location data. This licensed program helps the user to create for example simple web maps but also complex analytical models. Maps and 3D scenes can be created and designed. The tool offers a modelling framework and an analytical toolbox, which empowers the users' analytics. The used data can be controlled and

managed within the tool. ArcGIS offers the user a huge amount of open data already integrated and connected to the tool. (Esri, 2018a)

5.3.6 Google Charts

Google Charts is a JavaScript based tool that enables the visualization of data on a website. It includes everything from simple line charts up to complex hierarchical tree maps. A gallery further provides a set of prepared templates, which make it fast and easy to use. (Google, 2018)

Afterwards implemented Google Chart libraries list and illustrate the data. Those charts can then be customized and via embedded JavaScript displayed on the website. (Google, 2018)

The data can be stored directly on the website, a database or any other Chart Tools Data source protocol supporting the data provider. SQL-like query language, Google Fusion Tables and Salesforce are just some of the supported protocols. (Google, 2018)

5.3.7 Quadrigram

This drag & drop data editor can access any XLS or CSV files directly from Google Drive. It supports all the basic analytical functions. Data can be loaded, analysed and shared. This visual programming environment lets anyone construct and share interactive data visualization projects without any programming skills. The goal of connecting programming and spreadsheets is to increase the efficiency in analytical processes for businesses. (Bestiario_, n.d.)

5.3.8 QGIS

QGIS is an open sources graphical information system solution. The development started in 2002. The idea behind QGIS is to provide a tool so anyone interested in geographical analyses can process and work with geo data. In the beginning QGIS was only meant to be a data viewing tool. Today, QGIS is a user-friendly GIS, providing common functions and features. Furthermore, the tool is frequently used for daily GIS data-viewing needs. (QGIS Development Team, 2018d)

The source code of QGIS is publicly available and is developed with the Qt toolkit and C++. The tool is available on most Unix platforms as well as Windows and macOS. (QGIS Development Team, 2018f)

5.3.9 ExtraHop

ExtraHop is a company that offers data-driven IT products. Their basic functions are cloud-based real-time analytics and machine learning. This

means ExtraHop analyses data as fast as it is created to get the most accurate and timely source of intelligence possible. (ExtraHop Networks, 2018)

5.3.10 Sisense

Sisense released its first commercial tool in 2010. Their goal was to create a tool powerful enough to handle big amounts of data from different sources but also easy enough to be used by anyone. In their eyes, data analytics has to be fluent, easy and fast. (Sisense Inc., 2018i)

With their product they offer different functionalities. Sisense enables the user to mash up and analyse data. Furthermore, the tool supports the user to take action with the gained insights and execute actions to make use of them. Additionally, different embedding functions are available and data can be secured and controlled by the tool. (Sisense Inc., 2018g)

On their website, Sisense presents four differences compared to their competitors. First, they have an open single-stack architecture. This means they offer an end-to-end business intelligence software. Furthermore, they have an open API framework, which can be customized to the needs of the user. Also they claim to have a 10-100 times faster technology compared to others, which speeds up the processes. Machine learning is another difference, which should lead to more insights. Finally, Sisense can be deployed almost everywhere in a company. (Sisense Inc., 2018c)

5.3.11 Apache Hadoop

“The Apache™ Hadoop® project develops open source software for reliable, scalable, distributed computing.” (Apache Software Foundation, 2017).

Hadoop is a framework, which allows the distributed processing of large datasets. With simple programming models it is possible to process across clusters of computers. This means one can work on a single machine or an accumulation of computers. (Apache Software Foundation, 2017)

Hadoop consists of different modules. First, Hadoop Common are utilities that are needed to support the other modules. Then the Hadoop Distributed File System provides high-throughput access to the application data. Additionally, the YARN framework is included for the job scheduling and the cluster resource management. And finally, MapReduce is a YARN-based system for parallel processing of large datasets. Due to those abilities and functionalities Apache Hadoop can be used as open data processing and analytic tool. (Apache Software Foundation, 2017)

5.3.12 Apache Storm

Storm is a distributed real-time computing system from the organization Apache. The main function of Storm is to process streaming data. The tool is not limited to any programming language. (Apache Software Foundation, 2015)

The application area includes real-time analytics, online machine learning, continuous computation, distributed RPC, ETL and others. Storm is known to be fast. Due to its scalability and fault tolerance a million tuples can be processed per second and per node. (Apache Software Foundation, 2015)

Storm adapts to existing queuing and database technologies. This mean no additional integration costs arise. Storm first consumes stream data and then divides this data, so it can be distributed in different stages of computing. (Apache Software Foundation, 2015)

5.3.13 MongoDB

MongoDB is a document database. It adapts to its user needs with its scalability and flexibility. Data is stored in a JSON-like way. This means individual fields can change as well as the whole structure. (MongoDB Inc., 2018a)

To analyse the data ad hoc queries, indexing and real-time aggregation functions are provided. Due to the fact that MongoDB is a distributed database the core aspects are the high availability, horizontal scaling and geographical distribution. (MongoDB Inc., 2018b)

Last but no least MongoDB is free and open source. This means the tool can be accessed by anyone interested and the source code is available on the Internet. (MongoDB Inc., 2018a)

6 TESTING

To get more insights and a closer look on some of the tools they were tested. Based on popularity, availability and functionality four different tools were selected.

It was not possible to create new scenarios from scratch. Therefore, the provided tutorial served as a base and starting point for the testing. Once the tutorial was finished first results could be presented. Afterwards, the tool was tested again based on the gathered knowledge from the tutorials.

The selected tools are ArcGIS Pro, QGIS, Power BI and Sisense. For none of the selected tools, except for Power BI, previous knowledge existed. The tools were learned from scratch. Due to this fact a new user would need the same time and would get similar results as presented throughout the following chapters.

6.1 Evaluation Criteria

To get reliable results a defined set of questions and evaluation criteria are needed. Based on WINDWARDs “Your Complete Checklist for Evaluation Reporting Software” (Winward Studios Inc., 2014) an adapted checklist for tools was created. This checklist served as a framework for the testing.

The evaluation criteria are separated into four different groups. The first group of criteria are the technical aspects. The second group are the business aspects. The last two groups are the training and the experiences from working with the tool.

In the chapters 6.1.1 to 6.1.4 the evaluation criteria of each group will be explained in detail.

6.1.1 Technical Aspects

For the technical aspects the focus lies on the system compatibility, the development process, on sharing and co-creation, on data and database connectivity as well as on the output formats.

The first question about the system compatibility tries to answer for which operating system the tool is available and if there are any other system compatibility restrictions or challenges to implement the tool into existing systems.

In the development process the focus lies on how difficult it is to learn use the tool and to create a new analysis.

In the part on sharing and co-creation the possibility for collaborations with team members is elaborated.

The data and database connectivity questions focus on the data sources. Supported formats and data security are a further concern. Also functions like filtering and calculation are essential and therefore need to be supported.

Finally, the questions concerning the output format are asked. For example, the possibilities to present results and the way they can be transformed or stored.

6.1.2 Business Aspects

For the business aspects monetary and support questions are answered. Monetary topics are e.g. the acquisition and administration costs. The support section includes the vendor communication, user communities and forums as well as the tools documentation.

The acquisition section focuses on the costs that may arise to get licenses or buy versions as well as the costs to implement the system in the company. This also includes additional hardware and software costs. The administration costs further include the on-going maintenance, if needed the IT resources, the licenses and the costs of the learning process.

The support section checks the accessibility of the vendor, what kind of customer service he offers and the references he provides. Afterwards, user communities are inspected and how the vendor handles forums. Finally, the focus lies on the provided documentation.

6.1.3 Training

In the training section the given samples are first analysed and checked on their value for the user. In a second part, the offered training possibilities are reviewed. Then the question will be answered if there is a tutorial and if the vendor guides the user through it or if one has to do it on his own. Finally, it is checked if there are any online seminars or even courses provided.

6.1.4 Experience & Hands-on

For the last evaluation criteria experience & hands-on a review on the acquired experiences during the work with the tool will be given.

6.2 **ArcGIS Pro**

ArcGIS Pro provides a 21-days trial version, which supports all the functions. For this reason, an account was created and a subscription was made. Additionally, HAMK University of Applied Sciences offers a full version of ArcGIS on their servers.

6.2.1 Technical Aspects

Due to the fact that ArcGIS Pro is a tool, which heavily uses processing and computing power a detailed list of the supported operating systems is provided. Table 2 presents this list.

Table 2 - ArcGIS Supported Operating Systems (Esri, 2018b)

Supported Operating Systems
Windows 10 Home, Pro, and Enterprise (64 bit): Increased from Creators Update to Fall Creators Update.
Windows 8.1 Pro and Enterprise (64 bit): Increased from the April 2017 update to the Nov 2017 update.
Windows Server 2016 Standard and Datacentre (64 bit): Increased from no update to the Nov 2017 update.
Windows Server 2012 Standard and Datacentre (64 bit): Increased from the April update to the Nov 2017 update.

The development process ArcGIS is similar compared to the data management process. First data has to be imported and then handled. Only then visualization can be generated and analysis can be completed. All those steps are naturally connected to geographical data. So instead of statistical information the user works with locations, raster, 2D and 3D models, layers and topology. (Esri, 2018h)

Sharing and co-creation is integrated in ArcGIS through distributed collaboration. The GIS can be connected and integrated across a network of participants, which enables the organization and sharing of content between individuals, businesses and communities. (Esri, 2018g)

ArcGIS includes a so-called geo-database as a data management tool. The most common data formats are shapefiles, CAD, TIN, grids, imagery and GML. (Esri, 2018d)

6.2.2 Business Aspects

ArcGIS Pro is available in different version. The basic desktop version includes visualizing, building maps, editing data, importing CAD, performing data conversions and generating maps as well as query data. Those functions are available for $1,500 for a single use license, $3,500 for a concurrent or $800 for an annual subscription. (Hodel, 2017)

The desktop standard version allows everything that the desktop basic version includes and further adds additional features for multi-user platforms and for editing enterprise level geo-databases. This version is needed when other products of Esri want to be used. A single use subscription costs $7,000 and an annual one $3,000. (Hodel, 2017)

The third version ArcGIS desktop advanced is meant for heavy lifting geoprocessing. This version is normally only needed by advanced GIS users. The costs are depending on your organization. An annual subscription costs $4,200. (Hodel, 2017)

A commercial subscription includes support from Esri during the defined time. This means that there are no additional administration costs. Further implementation costs may arise but it is also possible to integrate and connect ArcGIS to the companies existing IT systems without the help of any specialists. (Hodel, 2017)

As support ArcGIS provides a help webpage, which includes information to project, maps and scenes, data, analysis and geoprocessing, metadata, layouts, production, workflows and sharing. (Esri, 2017b) Furthermore, additional support is offered on the support webpage, which is divided into community support, technical support and learn ArcGIS as contact point. (Esri, 2018c)

Esri does not give any reference on companies working with ArcGIS but it is probably the most popular geographic information system. ArcGIS is used in almost all branches where there is a possible use for geo data. From the banking sector over insurances to telecommunications and governments ArcGIS is used everywhere. (Esri, 2018j)

6.2.3 Training

The provided tutorial of ArcGIS Pro is divided into five sections. The first lesson is called “create a map” and the user starts with a project. Then he adds data to a map and explores this data. This lesson takes around 15 minutes.

The second lesson is called “symbolize layers and edit features”. As the names already indicates the map will now be symbolized with layers. The previous version was in an unattractive and unclear way. To get further information out of the map key features like landmarks are added. This lesson takes around 30 minutes and gives the user a first introduction into data editing. (Esri, 2018e)

In the third lesson, the created map is converted into a scene. This scene can afterwards be adapted and the view on the city can be customized. By adjusting the rendering settings, the data is displayed in a more

effective way. Now some characteristics of the city can be seen like for example that it is low laying and flat. Due to this fact and its closeness to the water the city is in danger to even small increases of the sea level. In this lesson a first insight can be found due to little changes in about 30 minutes (Figure 5). (Esri, 2018e)

Figure 5 - Venice visualized with ArcGIS Pro

The fourth lesson is about analysing the existing data and takes around 45 minutes. Using the integrated geoprocessing tools, a flood raster can be created and a rise of the sea level can be simulated. Afterwards, the percentage of the city, which would be flooded, can be calculated. This information can then be converted from a raster into a polygon, which shows the user the potential extends of the damage caused by the water rise (see Figure 6). (Esri, 2018e)

Figure 6 - Piazza San Marco visualized with ArcGIS Pro

In the last session of this tutorial, the scene is made ready for presentation. This means additionally to the finished scene, rule packages and multi-patch features are used to give a more realistic appearance. This created analysis combined with an attractive presentation of the map can now be presented to interested people. This final lesson takes

about 30 minutes. Additionally to the tutorial, a broad set of samples is offered on the official website of ArcGIS. (Esri, 2018f)

Udemy.com offers different courses concerning ArcGIS. There are 8 courses for beginners and 6 for advanced users. For the expert level no courses are offered. (Udemy Inc., 2018a)

Table 3 presents a possible way to deepen the knowledge in ArcGIS. Additionally, to the tutorials of ArcGIS, the basic functions are taught again and can thereby be strengthen. Afterwards, additional knowledge can be gathered, which will help creating and analysing geo-data with ArcGIS.

Table 3 - Udemy.com ArcGIS courses (Udemy Inc., 2018a)

Name	**Duration**	**Level**
ArcGIS Training - Become a GIS Analyst	2.5 hours	All levels
Getting Started with ArcGIS Mapping	2.5 hours	Beginner
ArcGIS Desktop For Spatial Analysis: Go From Basic To Pro	5 hours	Beginner
How to Produce Prediction Map in GIS With ArcGIS and Excel?	1.5 hours	All levels
Basics of Python & arcpy, the Python library of ESRI ArcGIS	2.5 hours	Beginner
Advanced Mapping with ArcGIS	2 hours	Advanced
The Ins and Outs of ArcGIS Data Analysis	2.5 hours	Advanced
Publish, Manage, and Consume Services Using ArcGIS Server	1.5 hours	Advanced

Together with the tutorial provided by Esri the in Table 3 presented courses give the user a basic knowledge, which is needed to handle open data with ArcGIS. With those courses around 20 hours are covered. However, it is expected that around 80-130 extra hours of experience are needed to work efficiently with ArcGIS.

6.2.4 Experience & Hands-on

Throughout the over 2.5 hours of tutorial first experiences could be made with ArcGIS Pro. For someone who never worked with a GIS it is important to learn the concept behind working with maps, layers and raster. Sadly, the tutorial did not focus on the data at all. It was just imported from an online portal. It would be interesting to see and handle the data before visualizing it.

After just a few minutes the first map could already be created and the possibilities of a GIS were shown. By importing layer, a first visualization was done pretty quickly. Afterwards, it was comprehensible explained how to handle and work with the map, so the user does not get lost. Furthermore, useful functions like symbolizing landmarks and setting bookmarks can be used in daily work with ArcGIS. Combining different datasets, a 3D map could be created based on the further imported 2D map. All the information and data were provided by ArcGIS.

With geoprocessing and supporting raster data, a scene could be created and made it possible to analyse the degree of damage caused by a raise of the sea level.

While working with ArcGIS only the basic functions were tested. After some hours it was possible to get simple insights out of data. But the tool includes much more difficult and advanced functions to analyse geo data. To get the most value out of ArcGIS a lot of experience is needed. This includes knowledge in geoprocessing and the language Python.

6.3 **QGIS**

Already on their website the QGIS development team lists some reasons why GIS users should select QGIS as tool. One of the biggest advantages is of course that it costs nothing. There are no initial fees or recurring fee. QGIS is completely free to use. (QGIS Development Team, 2018a)

Furthermore, QGIS relies on the input from their community. Users with specific needs can bring them to the development team. Afterwards it is possible that new functionalities may be included. Additionally, user can sponsor the development of the tool or they themselves can participate in the development process. (QGIS Development Team, 2018a)

This means that the tool never stops to improve. In every moment new features are added and existing ones are improved. QGIS was selected as the tool to be tested to have an open source analogy to ArcGIS. (QGIS Development Team, 2018a)

6.3.1 Technical Aspects

QGIS is freely available on Windows, Mac OS X, Linux and Android. The development process follows a traditional GIS and starts by creating or importing data, process and transform this data and afterwards to present it in some kind of reports. (QGIS Development Team, 2018c)

QGIS supports in global two different types of data. On one hand there are vector data which includes over 20 data types. On the other hand, there are the raster data formats also including around 15 data types.

Additionally, QGIS supports broad support of database formats. (QGIS Development Team, 2018d)

Concerning data security, QGIS provides multiple user authentication workflows like http(s)-, database and PKI authentication. The connection to the QGIS Server is secured with SSL. Finally, QGIS provides some basic security considerations and restrictions that can be expanded. (QGIS Development Team, 2018d)

To get the best use of QGIS and its functions some knowledge in Python is required. Using this programming language enables the user to add calculations and functions adapted to his desire. (QGIS Development Team, 2018d)

The knowledge in Python or other programming languages can further be used to extend the tool. QGIS offers in their documentation a part for developers, which includes a Python PyQGIS cookbook for plugins and scripting, a Python Api documentation and a C++ Api documentation. (QGIS Development Team, 2018d)

Throughout the testing it was simple to access older templates. In this testing case it was predefined templates that were downloaded from the Internet and then edited. This also means that editing templates works quickly and easily. (QGIS Development Team, 2018d)

The easiest way to present and share your gained insights is by creating a print composer. Map and atlases can be printed or saved as PDF-file, image or SVG-file. To share results, QGIS integrates UMN MapServer or GeoServer to publish results on webservers. (QGIS Development Team, 2018d)

6.3.2 Business Aspects

One of the biggest advantages of QGIS is that it is an open source tool. This means users receive the full tool without any costs for licenses or subscriptions. Therefore, the acquisition of the tools causes no costs. The integration of the tool into existing systems on the other hand can require some investments. If the company does not conduct an IT department including programmers and software developer external costs may arise due to a lack of knowledge. (QGIS Development Team, 2018a)

Furthermore, there are expected costs in educating and instruction for the staff into the tool. If employees have no experience with GIS systems, it may cost more money and take more time. For advanced GIS user a short time of adaption is expected.

To support users with problems the operator of QGIS offers some solutions. First of all, there are mailing lists. Questions can be asked to different groups like users, developers, community teams, translations, project steering committees and web client 2. Next, QGIS is active on StackExchange, which is a question and answer community that also answers QGIS questions. Furthermore, it is possible to chat and discuss with other people. This way is mostly used for the further development of QGIS. Also, different user groups are a good way to connect with other users. QGIS further offers a support on their website, which contains a lot of information. Additionally, there is a special function to report errors and bugs. It is also possible to get commercial support from third-party companies. Finally, QGIS operates a forum to ask questions. (QGIS Development Team, 2018g)

QGIS as open source tool has a long list of sponsors and donors. Most of them are organization but also individuals participate in the development of QGIS. Furthermore, QGIS is supported by multiple commercial supporters around the globe, which helps interested users to set-up QGIS in their business. (QGIS Development Team, 2018b)

6.3.3 Training

The tutorial of QGIS is document-based and divided in 21 sections. Those different parts are further divided into chapters. Every chapter contains exercises in three different difficulties. The basic chapters introduce essential information to the user. People who are new to GIS systems are recommended to do those chapters. The intermediate difficulty extends the basic course with some further information. The user is asked to work on his own and a basic understanding of GIS tools is asked. The last and most difficult level is the advanced one. Those chapters are for more advanced users that want to gain detailed knowledge in QGIS. (QGIS Development Team, 2018e)

The tutorial starts with a short course introduction followed by a presentation of the interface. Afterwards a first map is implemented. In the next chapter different layers based on vector data are included. This data then receives a dynamic visual appearance. (QGIS Development Team, 2018e)

Chapter 4 focuses on the classification of vector data. In a first part the attributes of the data are presented. Afterwards, the different labels that can be used are introduced. The chapter finishes with a classification of data. The next chapter is all about creating maps. First of all, the included map composer is used. To create a valuable map information like a title and a legend as well as customized items need to be added. (QGIS Development Team, 2018e)

In chapter 6 the already existing data is modified. At the same time new datasets are created. Furthermore, the theme of topology is introduced, explained and implemented into the work. Moreover, the functions of forms are introduced to help the user managing his data in a more user-friendly way as it is per default. The last part of the chapter are actions. Actions are functions that are executed if you click on them. They can be customized and integrated. (QGIS Development Team, 2018e)

Chapter 7 starts with the analysis of vectors. This means a problem arises and is solved with QGIS. Throughout the following chapters the themes of raster are introduced. Afterwards, analyses are completed. In a next chapter, plugins, online resources and the QGIS server are introduced. Moreover, the Geographic Resource Analysis Support System (GRASS) is introduced. Database concepts based on PostgreSQL and Spatial Database Concepts with PostGIS are treated in two separate chapters. To finish the tutorial, a QGIS processing guide is provided and the steps to use spatial databases are given. (QGIS Development Team, 2018e)

Throughout the tutorial the user is asked to work on his own. The results can then be checked on a solving page. Last but no least QGIS asks and offers the user to contribute to this tutorial, so it improves continuously. (QGIS Development Team, 2018e)

Table 4 shows a selection of courses offered on Udemy.com on QGIS. Together with the tutorial provided by QGIS, around 30 hours of courses are offered (Udemy Inc., 2018c). This should give the user a basic introduction on how to handle open data with QGIS. To use the full potential of QGIS, an estimated additional experience of 70 – 120 hours is needed.

Table 4 - Udemy.com QGIS courses (Udemy Inc., 2018c)

Name	**Duration**	**Level**
QGIS 3.0 for GIS Professionals	10 hours	Advanced
Make great maps using QGIS.	1 hours	All levels
Introduction to QGIS Python Programming	3.5 hours	Beginner
[Intermediate] Spatial Data Analysis with R, QGIS & More	4.5 hours	All levels

6.3.4 Experience & Hands-on

After around 12 hours of working with QGIS the following insights were gained. QGIS is a well-designed tool, which supports all the functions that a GIS tool needs. The provided tutorial gives the user a good introduction into the system but there is much more training, knowledge and experience needed to completely master the tool. The more than 460

pages long documentation is the perfect information source to answer any question a user might have.

Throughout the testing time QGIS crashed multiple times. The first time around 1 hour of work was lost. Afterwards, due to stricter precautions in saving only a small amount of work was lost. Every time the program crashed a report information was sent to the development team. It was not possible to recover the lost data and the tool did not give any reason for the crash. This does not mean that the tool itself had a problem. It is also possible that the operating system of the testing environment led to the problems.

Otherwise, QGIS made a good first impression. The connection to geo databases makes it easy to access location open data and to process it.

6.4 **Power BI**

Microsoft's Power BI is not only very interesting due to his various features and functions but also because its possibility to connect to a huge amount of different services. More than 40 different services like Google Analytics, Salesforce and GitHub, just to mention a few, can be connected and data can be integrated. (Microsoft, 2018g)

6.4.1 Technical Aspects

There are exist different versions of Power BI. Power BI Desktop as a desktop application (Windows), Power BI service as online SaaS (Software as a Service) and Power BI apps as mobile applications on Windows phones and tablets, iOS and Android. (Microsoft, 2018g)

The different applications services have different roles in a company. The number crunching only happens on the desktop application. The management interested in reports and result use the web service and the mobile application. The Desktop applications main purpose is to report the creation of information. Web services are for publishing and Power BI Mobile for sharing and for the consumption of the information. (Microsoft, 2018c)

Power BI always follows a defined stream. First, data must be integrated into Power BI Desktop so that a report can be created. Afterwards, new visualization and dashboard can be published to the Power BI Services. Finally, the dashboard can be shared with others, especially people who are on the go. Additionally, it is possible to view and interact with shared dashboards and reports in Power BI Mobile apps. (Microsoft, 2018c)

Power BI consists of the basic building blocks visualizations, datasets, reports, dashboards and tiles. Datasets can be simple excel files or a combination of many different sources. The data can then be filtered and

combined to provide a dataset, which is adjusted to ones needs. So-called connectors connect data from different sources. (Microsoft, 2018c)

Different content packs are integrated to get started and run Power BI quickly. Once your data is connected a refresh schedule can be set so that the data will be updated with the new versions of its source automatically. (Microsoft, 2018c)

Power BI includes over 50 different cloud services such as GitHub and Marketo. Moreover, generic sources are supported through CML, CSV, text and ODBC. A technical speciality of Power BI is the integration of natural language queries Q & A which enables to user to ask question via keyboard input. Furthermore, Power BI integrates the language N and offers a R script editor as well as the DAX Data Analysis Expressions. (Microsoft, 2018c)

6.4.2 Business Aspects

Power BI is available in two different versions. Power BI Desktop is free to use. This version includes the combination of different data sources, the cleaning and preparation of this data, the analysing and creating of reports and the publishing. The other version is called Power BI Pro and costs €8,40 per month per user. This version includes some special features like a real-time view of a company, an automatic update of data and of local sources, collaboration, observation and controlling the access to the data and the usage of application to distribute content. For big companies Microsoft offers special terms concerning the pricing depending on the actual usage. (Microsoft, 2018e)

Microsoft provides multiple ways of support on their Power BI website. First of all, information concerning the service status per region is given and known issues are listed. Also they have four top issues and questions, which are linked to their solution. Finally, they have six support options. First, they recommend see through the guided learning. Then samples of dashboards, reports and desktop files can be reviewed to get some inspiration. Another option is to go through the in-depth articles of the documentation. Questions can always be asked to the community and ideas to improve Power BI can be submitted to the development team. There exists also a possibility to report an issue in case a user finds a bug in the tool. If the user still has a problem he can always create a support ticket, which will be reviewed and answered by professionals. (Microsoft, 2018f)

Additionally, to those support options Microsoft provides a detailed and all time actual documentation. It contains more than 2000 pages and is frequently updated. Famous users of Power BI are Mercedes-Benz, Dell and Meijer. (Microsoft, 2016)

6.4.3 Training

The training Microsoft offers for Power BI is a mixture of videos, literature and visualisations. They indicate the time needed to go through the tutorial, but those numbers are not realistic. It takes almost double the time because the user has to explore Power BI on his own which takes more time than a guided tutorial. (Microsoft, 2018d)

The tutorial introduced all the functions that Power BI Desktop offers. However, the user has to use the function on his own and with his own data. The online learning is separated into 8 parts. The first chapter, getting started, introduces Power BI and shows how it should be used. Then a first look of the block system is offered and the different services are presented. Microsoft says this first chapter takes around 30 minutes. (Microsoft, 2018d)

The second chapter explains in 45 minutes how to work with the data. It shows the data management process and how to connect, clean and transform data. The next lesson introduced for almost 1 hour the theme of modelling. In this chapter all the different methods and tools to model the data are presented. (Microsoft, 2018d)

After this modelling part, visualizations are treated for 2 hours. The tutorial goes through almost all the visualization functions Power BI includes. This gives the user a good overview and shows him what the abilities of the tool are. (Microsoft, 2018d)

The next chapter is called exploring data. For 1 hour the different ways to get insights out of your visualization are introduced. At this moment the dashboards are already finished and the goal is to find answers to question the user might have about his data. Furthermore, the connection between Power BI and Excel are treated within 20 minutes. This chapter also includes how to import Power Pivot and Power View files as well as how to connect to OneDrive. (Microsoft, 2018d)

The second last chapter is all about publishing and sharing your created information. The report publishing and dashboard exporting process is treated. Furthermore, data refreshing, group creating and content packs are introduced. Microsoft says that this chapter should take a little less than 1 hour. (Microsoft, 2018d)

The final chapter introduces within 100 minutes the topic of DAX. Calculation types, functions variables expressions, table relationship and filtering are topics treated concerning the Data Analysis Expression language. (Microsoft, 2018d)

On udemy.com multiple courses on Power BI are offered. Around 200 are for beginners, more than 100 for advanced users and almost 40 for experts. (Udemy Inc., 2018b)

The courses presented in Table 5 give the user a good introduction into Power BI Desktop. Together with the tutorial from Power BI itself, the user gets courses for about 35 hours. The same number of hours or more are expected to be needed to get a deepened understanding of Power BI and to process open data with it. (Udemy Inc., 2018b)

Table 5 - Udemy.com Power BI courses (Udemy Inc., 2018b)

Name	**Duration**	**Level**
Microsoft Power BI - A Complete Introduction	10 hours	Beginner
Power BI Masterclass - Data Analysis Deep Dive	3.5 hours	Beginners
Powerful Reports and Dashboards with Microsoft Power BI	5.5 hours	Advanced
Power BI Desktop Query Editor - Master data transformation	6 hours	All levels
Mastering Intermediate DAX - Power BI, Power Pivot & SSAS	2 hours	Advanced

6.4.4 Experience & Hands-on

Working with Power BI feels comfortable for anyone who already worked with Excel. Power BI includes most Excel functions but extends those with more analytical and business intelligence functions.

The tutorial offers a great way to start working with Power BI. It introduced all the important steps to analyse open data. The workflow of Power BI, first to handle data with Power BI Desktop, then to publish the report with Power BI service and finally inspect it with Power BI Mobile helps the user to keep focused on his objectives.

Microsoft introduces Power BI during more than 8 hours. This time might be enough to read the given information and watch the provided YouTube videos, but if the user wants to get his hand on the tool and to do every step shown in the tutorial it will take much more time.

Once started in Power BI with one dataset it is hard to stop again. First it is enough to get some basic visualization but after some hours of working, more detailed questions are asked and new ways to answer them have to be found. In the end it is possible to create more complex visualizations by combining different data and including calculations.

6.5 **Sisense**

Sisense is the fourth and last tested tool. It offers an ideal way to analyse data without the need of understanding it. The focus lies on presentation and analyses via web browser rather than on data crunching. (Sisense Inc., 2018m)

6.5.1 Technical Aspects

The Sisense software runs on Windows 7 and higher and Windows Server 2008 R2 through Windows 2016. It is recommended to work on a Windows Server rather than Windows 7. Supported HTML 5 web browsers are the Internet Explorer 10 and higher, Google Chrome, Firefox and Safari version 7 and higher. Furthermore, the web application also works on mobile phones and tablet browsers supporting HTML 5. (Sisense Inc., 2018g)

The architecture of Sisense is separated into four parts. The first part, data sources, includes relational databases, spreadsheets & files, web applications and Hadoop big data. The second and third parts are together. On one hand there are the business intelligence server with ElastiCube Server and Sisense Web Server and on the other hand the front-end applications ElastiCube Manager and the Sisense web application. The last layer includes then the devices on which the tools runs. (Sisense Inc., 2018c)

Sisense does not only want to sell their product but also wants to assure a value to its customers. There are different ways existing how to get data. It is possible to use applications like Google Analytics, Salesforce or Microsoft Dynamics. Otherwise, it is possible to use big data tools like Google BigQuery or Hadoop via Hive. Moreover, data warehouses and databases can be connected. Of course, it is also possible to load simple files like CSV, excel sheets or to connect to a generic API. Once the data is loaded into the system visualizations can be created. Those visualizations can then be shown and shared on a dashboard. (Sisense Inc., 2018d)

Additionally, to its products, different add-ons are supported. The program can for example be extended so that it is possible to create aggregated tables, to include visualize widget queries (JAWLine) or be able to do forecasting. Also, dashboards can be exported as dash file or directly as a PDF. (Sisense Inc., 2018f)

6.5.2 Business Aspects

The price of the tool depends on its functionalities. This includes the number of users, the data volume and timeline. Normally licenses are sold on an annual base.

Sisense enables with their tool to do business analytics in a very simple way. Once the data is cleaned and ready to use, dashboards and other visualization can be created with just a few clicks. This means it makes it possible for non-technical users to solve analytical problems. It is also possible to work on dashboards on a mobile device. Furthermore, Sisense offers different cloud solutions for different business sections. (Sisense Inc., 2018b)

Once the data is connected to the system the results can be access immediately. This means that for example it is possible to elaborate a return on investment in weeks instead of years.

The team around Sisense claim to have found a faster way to process data. This means costs in other sections could be saved. Companies working with Sisense are for example Wix, Airbus and Orion just to mention a few. (Sisense Inc., 2018h)

Sisense offers different kind of support. Users with a problem have to possibility to submit a ticket that will be reviewed by the support team. Moreover, solved tickets can be inspected. If a user wants to solve the problem on his own, a documentation website with different sections is linked (Sisense Inc., 2018e). It is also possible to watch the video tutorials or to read through published case studies. Finally, it is possible to discuss either with other users from the community or with developers. Both ways are realized with the help of forums. (Sisense Inc., 2018j)

6.5.3 Training

The tutorial of Sisense is completely video based. The lengths of the videos are between 1 and 6 minutes and each treats one specific topic. Overall, three different parts exist. The first section concerns data preparation in Elasticube. This part includes everything concerning data handling. (Sisense Inc., 2018k)

The second part of the videos is about visualizing the data. This includes different functions as well as filtering and dashboard creating. Furthermore, this section includes the sharing process of the created information (Sisense Inc., 2018k).

The last part of the tutorial is called management and advanced settings. Treated topics are for example groups, data security and REST API customizing. (Sisense Inc., 2018k)

In total there are videos for 45 minutes concerning the preparation of data in Elasticube. The dashboard creating part and visualizations take around 90 minutes and the introduction of management and advanced settings are about 25 minutes. Additional to this tutorial, Sisense offers a demo where own or sample data can be visualized. (Sisense Inc., 2018k)

Furthermore, Sisense offers a set of web seminars, which users can book and then participate in (Sisense Inc., 2018l). It is recommended to take part in web seminars additionally to the tutorial videos to deepen the knowledge in handling open data with Sisense. In total 80 – 100 hours of experience are expected to be enough to handle open data successfully.

6.5.4 Experience & Hands-on

Going through the tutorial of Sisense gave a good first impression of the possibilities of the tool. The data preparation feels clean and there are just few possibilities where the user could get lost or destroy his data. The separation of the data into schemes in ElastiCube makes sense and helps to understand the data and their relationships. Uploading new data or adding data to an existing scheme is self-explanatory and well supported. Relationships between datasets can be created by connecting them with the mouse.

Once the data is ready for use it can be simply imported to the web application. The following visualization part is well-structured and easy to use. Everything is prepared and with the help of widgets charts can be created by dragging and dropping the chosen values or data.

With just a view clicks, informative visualization can be created and shown on a dashboard. The finished dashboard can then be easily filtered to show exactly the desired information. The arrangement of a dashboard can simply be done by dragging and dropping different widgets.

Globally, Sisense made a really good impression. The widget-based technique to build dashboard and visualize data makes it fast and easy to handle. Results can be simply exported as PDF and then distributed. After over 2 hours of tutorial and 2 more hours of testing it was on one hand possible to handle the tool, to make first analysis and find insights. On the other hand, the software offers way more functionalities that were not tested and to handle these, further experience or lectures would be needed.

7 MATRIX & COMPARISON

This section introduces a matrix which summarises the information provided in previous chapters. Afterwards, a comparison between data types and tools is given.

7.1 Matrix

As presented in the previous chapters, there are different ways along the data management process and the data value chain. Depending on the initial position and the available data the tools and technologies, which should be used are pre-defined.

Figure 7 shows an overview on how to turn open data into business values. As defined in chapter 4 there are three major steps in the data management process. First data has to be found or selected. Afterwards, this data has to be transformed into information, which finally can be converted into valuable knowledge for a company.

Category	Formats	Sources	Tools	Output/End-result
Location	CAD Imagery and raster Shapefile TIN	Opengeoportal.org	Geographical Information System (ArcGIS, QGIS)	Map Image Report
Cultural	Webpage PDF Image Media XML	openglam.org	GLAMwiki Toolset	Documentation
Science	Fileset Paper Dataset Figure	Figshare.com	Figshare API	Fileset Paper Dataset Figure
Finance	CSV GeoJSON JSON RDF-SML XML	Worldbank.org Google Finance	Socrata Open Data API	Analysis Visualisation Report
Statistics	CSV XML EXCEL	stat.fi	IBM SPSS MATLAB	Analysis Visualisation Report
Weather	CSV XML EXCEL	openweathermap.org ilmatieteenlaitos.fi	OpenWeatherMap API NASA Panoply	Weather map Report Visualisation
Environment	CSV XML EXCEL	data.worldbank.org data.oecd.org	Power BI	Analysis Visualisation Report
Government	CSV XML EXCEL	data.oecd.org avoindata.fi	Power BI	Analysis Visualisation Report

Figure 7 - Open data matrix

The presented open data matrix begins with the category of open data in the first column. Next it presents some of the supported and most common used data formats. Afterwards, it includes a data sources where open data can be accessed. The next column names an example of a tool to handle the data. Finally, the last column gives an example of the output of the process respectively the end-result.

Around the matrix the individual steps of the data management process are indicated. It is easy to see which position in the matrix takes part in which process step. The matrix finishes with the information step. What happens with the information and how it is transformed into knowledge is not part of this paper because it depends on how the company uses the results.

A company interested in using open data has two possibilities to start. On one hand they go through the process with the flow and start with the data. In this case they have to define which data they want to use. Afterwards a suitable tool, depending on the data and its purpose has to be selected. Then the data can be integrated, cleaned and analysed. Once enough insights are found, they can be visualized and thus present new information. This information can then be used to change for example business strategies and thereby become knowledge.

On the other hand, it is also possible to start with asking the question: “What results do we want to achieve with the help of open data?”. Maybe the company wants to improve its customer satisfaction by speeding up a process. Maybe a new service should be created. For both cases a different result is expected. Based on these expectations it is possible to define, what kind of information is needed to provide the knowledge for that purpose. Knowing which information is needed means knowing which data is needed. Once the company is sure of what they want, they can then start with the described process.

Both solutions have their pros and cons. For example, it is better to know what kind of result a company expects. In this case the people working with open data have a clear goal in their head and try to reach this goal as good as possible. But this also means, that there will only be the result that a company wants and nothing more. Businesses freely analysing open data may find insights and patterns that give them answers to questions they never thought of.

7.2 **Comparison**

Throughout this paper the topic of static and dynamic data was introduced. Data of almost every category of open data can be either static or dynamic. However, static data is far more common than dynamic. This is due to the fact that a big amount of static data is freely

available on the Internet and can easily be downloaded and integrated into analysing tools.

Dynamic data on contrary is harder to handle. Even though there are tools that support dynamic and stream data handling like ExtraHop along the whole process, it is much more time-consuming to work with this kind of data. Processing real-time data is a quiet new technique. Tools like Apache Storm support real-time data analysing and are used by innovative companies like Yahoo!, twitter or Spotify. This technology is essential for some of their services.

Furthermore, it is also important to know that dynamic and static data have different goals and requirements. They differ in scope, size performance and analyses. Table 6 gives the explanation to the mentioned differences.

Table 6 - Static vs. dynamic data (Amazon Web Services, 2017)

	Static data	**Dynamic data**
Data scope	Processing over all or most of the data in a dataset	Processing over data within a time window or most recent records.
Data size	Large data sets	Individual streams and record
Performance	Minutes and hours	Seconds and milliseconds
Analyses	Complex analytics	Simple responses and monitoring

It is easy to see, that static and dynamic data is used in different ways and serve different goals. In the open data case static data is more relevant due to its bigger popularity and availability on the open data market. But this does not mean that dynamic data can be left out working with open data.

Tools like QGIS and ArcGIS also support to a limited amount real-time processing. For example, ArcGIS offers with their ArcGIS GeoEvent Server the possibility to act as a real-time GIS, which continuously updates its data. The same real-time processing is possible in QGIS.

Power BI Desktop and Sisense both support possibilities to integrate a data stream to process data almost instantly. This means that for all types of data there are tools existing that support and provide functions to handle every step in the data value chain.

The testing showed that open source tools support most of the functions that licensed tools provide. There might be little differences in specific functions for analysing and processing data or in the support. For

example, the provided tutorials of the commercial tools are generally better guided and illustrated than the open source ones. But this also leads to the conclusion, that the user has to work more on his own for with the open source tools which leads to a better understanding of the tool itself.

However, it can be stated that to handle open data, open source tools are just as valuable as licenced tools and therefore more interesting due to their free availability. It is also important to notice that tools like ArcGIS and the licenced version of Power BI are close to be an ERP system, which can handle open data but also has a lot of other functions that are not as important or are even useless in handling open data.

8 CONCLUSION & RECOMMENDATIONS

Today, open data is an essential source of business insights for almost every company throughout all industries. As presented in the paper, open data can be separated into different categories. Governments are normally the most active players in the open data world, but there are multiple non-governmental organization who have the goal to advance the popularity of open data. Independent of these categories is the process that turns raw data into information.

Open data in an unprocessed state has no use for businesses. It first has to be treated and converted into information before it can be used and therefore be transformed into knowledge. All the tested tools are built around this basic process. It is not possible to create knowledge directly out of raw data.

Open data comes in different formats but for every single format there exist tools to handle the data and produce the best possible results. The market offers different technologies and tools to handle open data. Most of them are for free. Throughout the testing, two open sources and two commercial tools were tested and compared. It can be stated, that the open source tools were as good as the commercial ones and are thereby recommended for use.

Companies newly working with open data should start with static data, depending on their industrial sector and needs. There are tons of open data sources that can be accessed. To get a first look at the possibilities free tools like Power BI Desktop or QGIS are perfectly fine. The matrix introduced in chapter 7 can be used as an initial position and shows which way a user has to go based on what he desires.

It is understandable that companies are sceptical about opening their own data up. It is easier to see the possible threats and disadvantages in this process. However, there exist the chance that interested people find valuable insights or even create new knowledge, which then provides essential business opportunities. Companies interested in open data can follow this thesis to understand key technologies, tools and skills needed to work with open data.

REFERENCES

Amazon Web Services (2017). *Amazon*. Retrieved March 4, 2018, from What is Streaming Data?: https://aws.amazon.com/streaming-data/?nc1=h_ls

Apache Software Foundation (2015). *Apache Storm*. Retrieved March 4, 2018, from Apache Storm: http://storm.apache.org/

Apache Software Foundation (2017, December 18). *Apache Hadoop*. Retrieved March 4, 2018, from Welcome to Apache Hadoop: http://hadoop.apache.org/

Avoindata.fi (2015, August 10). *Avoindata.fi - About*. Retrieved March 4, 2018, from https://www.avoindata.fi/en/about

Avoindata.fi (n.d.). *Avoindata.fi*. Retrieved March 4, 2018, from https://www.avoindata.fi/en

Berners-Lee, T. (2006, July 27). *Linked Data*. Retrieved March 4, 2018, from https://www.w3.org/DesignIssues/LinkedData.html

Bestiario_ (n.d.). *What is Quadrigram?* Retrieved March 4, 2018, from http://www.quadrigram.com/about-us/

Carrara, W., Chan, W., Fischer, S., & van Steenbergen, E. (2015). *Creating Value through Open Data.* European Comission.

Definition, O. (n.d.). *Open Definition*. Retrieved March 4, 2018, from http://opendefinition.org/od/2.1/en/

Esri (2017a). *ArcGIS Pro*. Retrieved March 4, 2018, from Data types: http://pro.arcgis.com/en/pro-app/help/data/annotation/annotation.htm

Esri (2017b). *ArcGIS Pro*. Retrieved March 4, 2018, from Help: http://pro.arcgis.com/en/pro-app/help/main/welcome-to-the-arcgis-pro-app-help.htm

Esri (2018a). *ArcGIS Desktop*. Retrieved March 4, 2018, from http://desktop.arcgis.com/en/

Esri (2018b). *ArcGIS Desktop*. Retrieved March 4, 2018, from System Requirements: https://pro.arcgis.com/en/pro-app/get-started/arcgis-pro-system-requirements.htm

Esri (2018c). *ArcGIS Desktop*. Retrieved March 4, 2018, from Support: http://desktop.arcgis.com/en/support/

Esri (2018d, March 4). *ArcGIS Pro*. Retrieved from Databases and ArcGIS Pro: https://pro.arcgis.com/en/pro-app/help/data/databases/databases-and-arcgis-pro.htm

Esri (2018e). *ArcGIS Pro*. Retrieved March 4, 2018, from Overview: https://learn.arcgis.com/en/projects/get-started-with-arcgis-pro/

Esri (2018f). *ArcGIS Pro*. Retrieved March 4, 2018, from Tools Reference: http://pro.arcgis.com/en/pro-app/tool-reference/data-management/compact.htm

Esri (2018g). *ArcGIS Pro*. Retrieved March 4, 2018, from Share with ArcGIS Pro: https://pro.arcgis.com/en/pro-app/help/sharing/overview/share-with-arcgis-pro.htm

Esri (2018h). *ArcGIS Pro*. Retrieved March 4, 2018, from Data Types: https://pro.arcgis.com/en/pro-app/help/data/annotation/annotation.htm

Esri (2018i). *ArcGIS*. Retrieved March 4, 2018, from Homepage: https://desktop.arcgis.com/en/

Esri (2018j). *Esri*. Retrieved March 4, 2018, from Industries: https://www.esri.com/en-us/industries/index

EU Open Data Portal (2018). *EU Open Data Portal*. Retrieved March 4, 2018, from http://data.europa.eu

European Data Protal (2018). *European Data Portal*. Retrieved March 4, 2018, from https://www.europeandataportal.eu/

ExtraHop Networks (2018). *ExtraHop*. Retrieved March 4, 2018, from https://www.extrahop.com/

Fielding, R. T. (2000). *Representation State Transfer (Rest)*. Retrieved March 4, 2018, from Chapter 5: http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

Google (2018). *Google Developers*. Retrieved March 4, 2018, from Charts: https://developers.google.com/chart/?csw=1

HAMK (2017a). *AvoinHäme*. Retrieved March 4, 2018, from Project Information: https://xn--avoinhme-5za.fi/avoinhame-hanke-kuvaus/

HAMK (2017b, January 1). *Thesis Guide.* Retrieved March 4, 2018, from https://hameenamk.sharepoint.com/yhteiset-sisallot/laatukasikirja/koulutus/amk/Thesis/Thesis-guide.pdf

Hausenblas, M. (2015). *5 star data*. Retrieved March 4, 2018, from http://5stardata.info/en/

Hey, J. (2004). *The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link.* Intergovernmental Oceanographic Commission 18.

Hodel, J. (2017, June 1). *Cloudpointgeo*. (Cloudpoint Geographics, Inc.) Retrieved March 4, 2018, from A Summary Of ArcGIS Platform And Products: http://www.cloudpointgeo.com/blog/blog/2014/5/9/a-summary-of-arcgis-platform-and-products

Luna-Reyes, L. F., Bertot, C. J., & Mellouli, S. (2014). Open Government, Open Data and Digital Government. *Government Information Quarterly, 31*(1), 4-5.

Microsoft (2016, October 3). *Customer Stories*. Retrieved March 4, 2018, from https://customers.microsoft.com/en-us/story/top-supermarket-chain-gains-insight-and-boosts-profitability-with-microsoft-power-bi

Microsoft (2018a). *Dynamics*. Retrieved March 4, 2018, from https://dynamics.microsoft.com/en-us/

Microsoft (2018b). *Excel 2016*. Retrieved March 4, 2018, from https://products.office.com/en-us/excel

Microsoft (2018c). *Power BI Desktop*. Retrieved March 4, 2018, from https://docs.microsoft.com/en-us/power-bi/desktop-getting-started

Microsoft (2018d). *Power BI Guided Learning*. Retrieved March 4, 2018, from https://docs.microsoft.com/en-us/power-bi/guided-learning/

Microsoft (2018e). *Power BI Pricing*. Retrieved March 4, 2018, from https://powerbi.microsoft.com/en-us/pricing/

Microsoft (2018f). *Power BI Support*. Retrieved March 4, 2018, from https://powerbi.microsoft.com/en-us/support/

Microsoft (2018g). *Power BI*. Retrieved March 4, 2018, from https://powerbi.microsoft.com/en-us/

Microsoft (2018h). *Social Engagement*. Retrieved March 4, 2018, from https://docs.microsoft.com/en-us/dynamics365/customer-engagement/social-engagement/overview

MongoDB Inc. (2018a). *What is MongoDB?* Retrieved March 4, 2018, from https://www.mongodb.com/what-is-mongodb

MongoDB Inc. (2018b). *Real-Time Analytics*. Retrieved March 4, 2018, from https://www.mongodb.com/use-cases/real-time-analytics

Moren, D. (2015). *Stuck At Home? Watch Snowplows In Real Time*. (Bonnier Corporation) Retrieved March 4, 2018, from https://www.popsci.com/dig-live-plow-tracking#page-2

Murray-Rust, P. (2008). Open Data in Science. *Serials Review, 34*(1), 52-64.

Open Knowledge International (2018a, March 4). *Open Data Handbook*. Retrieved from How to Open up Data: http://opendatahandbook.org/guide/en/how-to-open-up-data/

Open Knowledge International (2018b). *Open Data Handbook*. Retrieved March 4, 2018, from File Formats: http://opendatahandbook.org/guide/en/appendices/file-formats/

Open Knowledge International (2018c). *Open Data Handbook*. Retrieved March 5, 2018, from Homepage: http://opendatahandbook.org/

Open Knowledge International (2018d). *Open Data Handbook*. Retrieved March 4, 2018, from Introduction: http://opendatahandbook.org/guide/en/introduction/

Open Knowledge International (2018e). *Open Data Handbook*. Retrieved March 4, 2018, from Why Open Data?: http://opendatahandbook.org/guide/en/why-open-data/

Open Knowledge International (2018f). *Open Definition*. Retrieved March 4, 2018, from Definition: http://opendefinition.org/od/2.1/en/

Open Knowledge International (2018g). *Open Knowledge Finland*. Retrieved March 4, 2018, from https://fi.okfn.org/about/

Open Knowledge International (2018h). *Open Knowledge International*. Retrieved March 4, 2018, from About: https://okfn.org/about/

Pantano, E., Priporas, C.-V., & Stylos, N. (2017). 'You will like it!' using open data to predict tourists' response to a tourist attraction. *Elsevier*(60), 430-438.

Publications Office EU (2018). *EU Open Data Portal - About*. Retrieved March 4, 2018, from http://data.europa.eu/euodp/en/about

QGIS Development Team (2018a). *About*. Retrieved March 4, 2018, from https://qgis.org/en/site/about/index.html

QGIS Development Team (2018b). *Commercial Support*. Retrieved March 4, 2018, from https://www.qgis.org/en/site/forusers/commercial_support.html

QGIS Development Team (2018c). *Download QGIS for your platform*. Retrieved March 4, 2018, from https://qgis.org/en/site/forusers/download.html

QGIS Development Team (2018d, Februar 11). *QGIS Documentation.* Retrieved March 4, 2018, from https://docs.qgis.org/2.18/pdf/en/QGIS-2.18-QGISTrainingManual-en.pdf

QGIS Development Team (2018e). *QGIS Training Manual*. Retrieved March 4, 2018, from https://docs.qgis.org/2.18/en/docs/training_manual/index.html

QGIS Development Team (2018f). *QGIS*. (QGIS Georaphic Information System Open Source Geospatial Foundation Project) Retrieved March 4, 2018, from https://qgis.org/en/site/

QGIS Development Team (2018g). *Support*. Retrieved March 4, 2018, from https://qgis.org/en/site/forusers/support.html

Refsnes Data (2018a). *JSON - Introduction*. Retrieved March 4, 2018, from https://www.w3schools.com/js/js_json_intro.asp

Refsnes Data (2018b). *XML - Tutorial*. Retrieved March 4, 2018, from https://www.w3schools.com/xml/default.asp

Santoso, S., & Lamoree, J. D. (2000). *Power Quality Data Analysis: From raw data to knowledge using knowledge discovery.* Knoxville: Electrotek Concepts, Inc.

Sibelius, K., Suonsaari, J., Aho, L., Laitinen, A.-M., Luhtunen, K., & Koskiniemi, P. (2018). *Databusiness.fi - Our Services*. Retrieved March 4, 2018, from https://www.databusiness.fi/en/info/our-services/

Sisense Inc. (2018a). *BA Software*. Retrieved March 4, 2018, from https://www.sisense.com/business-analytics-software/

Sisense Inc. (2018b). *Cloud BI*. Retrieved March 4, 2018, from https://www.sisense.com/product/cloud-bi/

Sisense Inc. (2018c). *Company*. Retrieved March 4, 2018, from https://www.sisense.com/company/

Sisense Inc. (2018d). *Data Connectors*. Retrieved March 4, 2018, from https://www.sisense.com/data-connectors/

Sisense Inc. (2018e). *Documentation*. Retrieved March 4, 2018, from https://documentation.sisense.com/

Sisense Inc. (2018f). *Platform, Analze, Add ons*. Retrieved March 4, 2018, from https://www.sisense.com/product/analyze/add-ons/

Sisense Inc. (2018g). *Product*. Retrieved March 4, 2018, from https://www.sisense.com/product/

Sisense Inc. (2018h). *Sisense Compared To The Alternatives*. Retrieved March 4, 2018, from https://www.sisense.com/why-sisense/sisense-compared-to-alternatives/

Sisense Inc. (2018i). *Sisense*. Retrieved March 4, 2018, from https://www.sisense.com/

Sisense Inc. (2018j). *Support*. Retrieved March 4, 2018, from https://www.sisense.com/support/

Sisense Inc. (2018k). *Tutorials*. Retrieved March 4, 2018, from https://www.sisense.com/training/tutorials/

Sisense Inc. (2018l). *Webinars*. Retrieved March 4, 2018, from https://www.sisense.com/webinars/

Sisense Inc. (2018m). *Why Sisense*. Retrieved March 4, 2018, from https://www.sisense.com/why-sisense/

The World Bank Group (2018). *The World Bank - Data - About us*. Retrieved March 4, 2018, from https://data.worldbank.org/about

Udemy Inc. (2018a). *Udemy ArcGIS*. Retrieved March 4, 2018, from https://www.udemy.com/courses/search/?ref=home&src=ukw&q=arcgis

Udemy Inc. (2018b). *Udemy Power BI*. Retrieved March 4, 2018, from https://www.udemy.com/courses/search/?q=power%20bi&src=sac&kw=power%20bi&lang=en

Udemy Inc. (2018c). *Udemy QGIS*. Retrieved March 4, 2018, from https://www.udemy.com/courses/search/?q=qgis&src=ukw&lang=en

United Nations Statistics Devision (2018). *UNdata - About*. Retrieved March 4, 2018, from http://data.un.org/Host.aspx?Content=About

W3C (2018). *About W3C*. Retrieved March 4, 2018, from https://www.w3.org/Consortium/

Vathana, A., & Audsin, D. P. (2013). *An Open Analysis on Open Data.* W3C Consortium.

Wikipedia Foundation Inc. (2018). *Application programming interface*. Retrieved March 4, 2018, from https://en.wikipedia.org/wiki/Application_programming_interface

Winward Studios Inc. (2014). *The Complete Checklist for Reporting Software.* Retrieved March 4, 2018, from http://go.windward.net/Reporting-Checklist.html#.Wp1ot5NuZ24

Yahoo Developer Network (2018). *Yahoo Query Language (YQL)*. Retrieved March 4, 2018, from https://developer.yahoo.com/yql/

APPENDIX HEADING

Matrix with links

Category	Formats	Sources	Tools	Output/End-result
Location	CAD Imagery and raster Shapefile TIN	Opengeoportal.org	Geographical Information System (ArcGIS, QGIS)	ArcGIS Example[2] QGIS Example[3]
Cultural	Webpage PDF Image Media XML	openglam.org	GLAMwiki Toolset	Example[4]
Science	Fileset Paper Dataset Figure	Figshare.com	Figshare API	Figshare Example[5]
Finance	CSV GeoJSON JSON RDF-SML XML	Worldbank.org Google Finance	Socrata Open Data API	Example[6]
Statistics	CSV XML EXCEL	stat.fi	IBM SPSS MATLAB	IBM SPASS Ex.[7] MATLAB Example[8]
Weather	CSV XML EXCEL	openweathermap.org ilmatieteenlaitos.fi	OpenWeatherMap API NASA Panoply	NASA Panoply Ex.[9]
Environment	CSV XML EXCEL	data.worldbank.org data.oecd.org	Power BI	Example[10]
Government	CSV XML EXCEL	data.oecd.org avoindata.fi	Power BI	Example[11]

[2] http://www.esri.com/products/maps-we-love
[3] https://www.qgis.org/en/site/about/screenshots.html
[4] https://commons.wikimedia.org/wiki/Commons:GLAMwiki_Toolset
[5] https://figshare.com/
[6] http://financesapp.worldbank.org/en/summaries/ibrd-ida/
[7] https://www.ibm.com/uk-en/marketplace/spss-statistics
[8] https://se.mathworks.com/help/matlab/examples.html?requestedDomain=true&nocookie=true
[9] https://www.nas.nasa.gov/SC12/assets/images/content/Tamkin_G_panoply_1_big.jpg
[10] https://community.powerbi.com/t5/Power-BI-Showcase/Air-Quality-Report/td-p/20140
[11] http://community.powerbi.com/t5/Data-Stories-Gallery/US-Government-Budget-Analysis/m-p/309333