Forming a baseline for open health data.

Towards a scorecard based on a preliminary mapping study.

Michael Haider

A dissertation submitted to the University of Dublin, in partial fulfilment of the requirements for the degree of Master of Science in Health Informatics

2017

Declaration

I declare that the work described in this dissertation is, except where otherwise stated, entirely my own work, and has not been submitted as an exercise for a degree at this or any other university. I further declare that this research has been carried out in full compliance with the ethical research requirements of the School of Computer Science and Statistics.

Signed:___________________

Michael Haider

20th June 2017

Permission to lend and/or copy

I agree that the Trinity College Library may lend or copy this dissertation upon request (haiderm@tcd.ie).

Signed: ___________________

Michael Haider

20th June 2017

Acknowledgements

Family. I am blessed to have such a loving, kind and supportive family. Thank you for everything.

Trinity College Dublin. I remember the day that I received the admission-letter very well and feel lucky begin given this opportunity. I want to especially thank the course director and my supervisor for this dissertation, Lucy Hederman. Firstly, for leading and creating the course and, secondly, for all the time and support I received while writing this dissertation. I also want to thank all lecturers for sharing their knowledge and especially Gaye Stephens. Further, I want to thank the Executive Officer for this course Margaret Murray. It was a true pleasure being part of such an intelligent and enjoyable class and I want to thank all my classmates for this experience.

Republic of Ireland and the European Union. Coming to Ireland in 2014, I continue to feel very welcome and moving here could not have been easier. I am grateful for this journey and hope that many more Europeans will continue to seek such opportunities. The tuition of the course was part of the Graduate Skills Conversion Programme and partly funded by the Higher Education Authority, which I am grateful for.

Healthcare professionals. I want to thank all healthcare professionals. They daily do what I can’t and their devotion to serving their patients is a true inspiration.

Table of Contents

List of tables

List of figures

Abbreviations

AHP	Analytic Hierarchy Process
API	Application programming interface
CDSS	Clinical Decision Support Systems
CDSS	Clinical Decision Support Systems
CEM	Clinical Element Model
CT	Computer Tomography
DICOM	Digital Imaging and Communication in Medicine
fMRI	functional Magnetic Resonance Images
HCLS IG	Health Care and Life Sciences Interest Group
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
LiMDAC	Linked Medical Data Access Control Framework
OWL	Web Ontology Language
PDQM	Portal Data Quality Model
RDF	Resource Description Framework
SOA	Service Oriented Architectures
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WHO	World Health Organization
XML	Extensible Markup Language

Terminology

This section describes important and frequently used terminology in the dissertation that should allow the reader an easier start into the dissertation.

Open data can be described as a well-established paradigm and global trend to make data freely available to everyone (Vasa and Tamilselvam, 2014).

Open government data is a subgroup of open data that is publicly available government-related data (Kučera, Chlapek and Nečaský, 2013).

Linked data is a set of best practices for connecting and publishing structured data on the Web (Bizer, Heath and Berners-Lee, 2009). The primary goal of the linked data movement is making the Web useful for sharing and interlinking data (not only documents) at a very detailed level (Samwald *et al.*, 2011).

Health data, consists of ***biomedical research data*** (e.g. genomic sequencing technologies and microarrays); ***clinical data*** (e.g. patient records and clinicians' documentations) ***health business data*** (e.g. costs and utilization) and ***private patient data*** (e.g. sports data and insurance data) (Holzinger, 2016). Alternative definitions will be provided in chapter 3.3. Health data.

Open data initiatives share a noble cause: increase transparency (Janssen, 2011), increase opportunity (Alexopoulos *et al.*, 2014) and increase collaboration. (Sandoval-Almazan and Gil-Garcia, 2014).

1 Introduction

"Data, and in this case, big data will become the third fundamental window into humanity after the microscope and the telescope. It's that important."
- John Nosta (John Nosta, 2016)

1.1 Preface

In the algorithmic age, the use and availability of big data sets in healthcare has become not only essential but increasingly an imperative. Health data, not only complex but also arbitrary, faces unique challenges like no other domain. Transforming health data into knowledge can range from contributing to medical breakthroughs to taming unsustainable health system costs. This dissertation deals with the availability side of open health data, more particularly how to evaluate open health data. The outcome is a scorecard that, naturally, also acts as a guideline on what to consider when publishing open health data. While much research has been done on open data generally, limited research is done on open health data specifically. This work tries to fill this research gap and aims to establish a baseline for open health data. Given the limited amount of research focusing on the health domain, a preliminary mapping study had to be executed initially and is central to this dissertation. The clusters identified were "Open data applications, platforms, portals and initiatives", "Open data enablers" and "Open data evaluations and guidelines". From the three clusters evidence was collected that could be relevant for the scorecard. Based on these insights the author created the open health data scorecard. This scorecard shall not be understood as a complementary (in the sense of another scorecard evaluation already existing points of analysis (such as data quality)) one to open data generally but one that is tailored to the health domain. Since there is very little research that concentrates on evaluating open health data, benchmarking the outcome is only possible in a very limited way. Only further research and the research community's discourse can holistically evaluate the outcome. What is more, weighting the individual components of the scorecard (e.g. is data quality or data openness more important) exceeded the scope of this research and could become the focus of further studies. Also, different health regions could be evaluated with the new scorecard in a future publication and be compared to other evaluations on the same regions.

1.2 Research questions

This section will talk about the research questions and how they will be addressed throughout the dissertation.

Based on the previously stressed importance of available datasets and a, for the author's understanding, continuously increasing demand for open data, the author first wanted to understand the current state of open health data. After an initial literature screen, it turned out that there is little research done on open health data specifically but various research done on open data generally. The research focusing on open health data specifically often deals with a specific area but doesn't equip the reader with a broader view of the whole ecosystem. The author, therefore, set out to capture the status quo of open health data in a way that would allow to create some sort of baseline. While a baseline can have several meanings, the baseline created here shall be understood as a starting point (Merriam-Webster, 2017) to understand the open health data domain and provide an overview of the research field. It is a high level undertaking to understand the current state of open health data or in simpler words "what do we know about open health data and how does it work". In light of this the first research question was crafted:

What is the current baseline for open health data?

Given the epically big scope that needed to be understood, captured and the diverse existing literature, the appropriate research method was found in a preliminary mapping study (more on the methodology in chapter 2 Methodology).

The first research question will, thereafter, be dealt with in chapter 3 Preliminary mapping study.

By capturing the state of the art and establishing a baseline for open health data, the author understood that this would provide a rich resource, able to deliver interesting insights. Through the preliminary mapping study, the author also could not identify an evaluation framework that assesses open health data generally, which represents a research gap. In light of this the second research question was built:

What might be an appropriate scorecard for evaluating open health data?

The research question was formed because the author found that most existing open data scorecards or evaluation frameworks deal with open data generally and not open health data specifically. Finding further research questions after establishing a baseline of a research domain is a common occurrence in systematic mapping studies (Budgen *et al.*, 2008). The dissertations not only undercovers an additional research question but also answers it. The second research question will, thereafter, be dealt with in chapter 4 Developing the scorecard.

Since the involvement of all stakeholders is often stressed in the health informatics domain, the author intended to answer the research questions so it could be understood by a diverse audience and without deep technical knowledge.

1.3 Overview of dissertation

This chapter provides the reader with an overview of the different chapters of the dissertation.

Chapter 1 gives an introduction into the dissertation, describes the two research questions and equip the reader with an overall overview of the dissertation.

Chapter 2 describes the selected methodology for the dissertation and how it ties into the defined research questions.

Chapter 3 deals with the preliminary mapping study that was carried out to answer the first research question.

Chapter 4 extracts the insights gathered from the preliminary mapping study and builds a scorecard to evaluate open health data. By building the scorecard, the chapter answers the second research question.

Chapter 5 concludes the dissertation by summarizing the main findings and opportunities for further research.

2 Methodology

This chapter informs the reader on the methodology that was used to answer the two research questions described in 1.2. Research questions. Given the aspired undertaking to understand the current baseline for open health data, first the author had to find the appropriate methodology to answer the initial research question: “What is the current baseline for open health data?”. This process will be described in section 2.1. In 2.2, the author will describe the process that was carried out to answer the two research questions.

2.1 Literature review for methodology

While crafting the initial research question for this dissertation, it became clear that the author would need to follow a process that supports broad research endeavours. For such undertakings, Kitchenham & Charters suggest a systematic mapping study, which built the foundation for this dissertation:

“A systematic mapping study allows the evidence in a domain to be plotted at a high level of granularity. This allows for the identification of evidence clusters and evidence deserts to direct the focus of future systematic reviews and to identify areas for more primary studies to be conducted.” (Kitchenham and Charters, 2007)

That this is the right approach to tackle the research questions is further confirmed by Petersen et al., who affirms that the main goal of a systematic mapping study is to provide an overview of a research area (Petersen *et al.*, 2008).

A systematic mapping study is a form of secondary research that examines a broader topic and classifies the literature in that specific research area. The main benefit is in providing the research community with a baseline for further research and providing current researchers in the research field with an overview. What is more, clusters identified during the systematic mapping study undercover where there might scope for a complete systematic literature review (Budgen *et al.*, 2008). Also, it is common to use a high level research question (Valença *et al.*, 2013) (Kitchenham, Budgen and Pearl Brereton, 2011). Therefore, a mapping study is the adequate methodology to answer the research question: what is the current baseline for open health data?

While mapping studies do not follow a strict form of analysis (Budgen *et al.*, 2008), Budgen et al describes some common themes:

- searching
- inclusion/exclusion
- bias/validity

These might seem familiar to the reader given that they are closely related to systematic literature reviews, a methodology associated with systematic mapping studies (Budgen *et al.*, 2008).

To further expand on the three mentioned general themes in Budgen et al., the author used a similar systematic mapping study protocol to the one used in Valença et al. (Valença *et al.*, 2013). While mapping studies often deal with software engineering (e.g. (Petersen *et al.*, 2008), Valença et al. conducted a systematic in the field of business process variability. This area seems related to this dissertation considering a similar high degree of technical content but not to the extent of software engineering. Therefore, the process below suits this dissertation well:

1) Research question
2) Search process
3) Selection criteria
4) Data extraction and analysis

Additionally, the process described above is backed up by (Budgen *et al.*, 2008), (Kitchenham, Dyba and Jorgensen, 2004), (Petersen *et al.*, 2008), (Kitchenham, 2010) & (Kitchenham, Budgen and Pearl Brereton, 2011).

The mapping study conducted in this dissertation had to be slightly adapted to fit the unique scope of this dissertation and also to consider its limitations and resource constraints.

The following limitations had to be considered:

- **diverse terminology**

Like with many concepts in health informatics, open health data is not universally defined and touches on several other research areas. For example, open data could also be represented as open linked data in the literature or as open government data, which also relevant for this dissertation. An open data evaluation framework could also be called a classification scheme, an open data scorecard or simply open data evaluation. Further, there are also evaluations conducted in papers focusing on open data quality or in publications looking at open data applications.

- **broad research area**

Due to diversity of open data generally and open health data specifically, the research area is published in different journals dealing with different research domains. There are relevant articles in technical journals, health informatics journals or in public policy journals, to only name a few. Further, relevant articles are part of grey publications such as reports and presentations. Including only a selected number of journals from one domain, would only provide a limited view of the research area. Further, very little research is done on open health data specifically but rather open data in general. If this dissertation would have only looked at open health data publications, there would be gaps in its analysis. For a better understanding, it is worth looking at the search results on Google Scholar for open data and open health data (why Google Scholar was used will be described later). While a search for *allintitle: "open health data"* returns 18 results, a search for *allintitle: "open data"* returns 4,810 results (exact phrase, excluding patents and citations, all years). Not using an exact phrase would even return 8,910 (results with all other settings remaining) and these results could also be relevant.

Considering the timeframe of this dissertation and, inherent, resource constraint, it was not feasible to analyse and process 4,810/ 8,910 papers. Further, the author wanted to go beyond the mapping study and also extract insights that are relevant to evaluate open health data.

To honour the limitation of completeness, Kitchenham (Kitchenham, 2010) suggest a preliminary mapping study that was carried out for this dissertation. Kitchenham 2010 also describes the central difference between a systematic mapping study and a preliminary mapping study, which should make the reader confident that the right methodology was chosen:

"The difference between systematic literature reviews and mapping (scoping) studies is the type of question they answer. A systematic literature review asks a fairly specific question such as ''Which of the Chidamber and Kemerer OO metrics are good predictors of fault-proneness", while mapping studies ask more general questions such as ''What do we know about OO related metrics"."

The process used in Kitchenham (Kitchenham, 2010) follows the following steps:

1) Identifying relevant papers
2) Data extraction
3) Aggregating primary study results

The reader can see that the steps themselves are fairly similar to the systematic mapping study described by Valença et al. (Valença *et al.*, 2013) and mainly differ in how the search process is conducted. Following, the dissertation will provide further insight into the individual steps of the preliminary mapping process. The research processes followed for this dissertation will be described in the following section.

2.2 Methodology process

After describing the literature research on methodologies, the author will now outline the research process that was followed for this dissertation.

0) Research question

The research questions are described in depth in section 1.2 Research questions as well as during the course of this section. As a reminder, the initial research question was: “What is the current baseline for open health data?” and the second research question, which was built after the research gap was identified, is: “What might be an appropriate scorecard for evaluating open health data?”.

1) Identifying relevant papers / search process

The search process followed different steps, which will be individually described below.

1a) Literature discovery/ manual search

The identification of relevant papers started off with a literature discovery process. During the literature discovery process the author did a basic web search for queries like “open data”, “open health data” and “open data framework”. Further, websites form organizations such as the World Health Organization (www.who.int), the Open Knowledge Foundation (okfn.org) or McKinsey Global Institute (www.mckinsey.com/mgi/overview) were scanned for relevant content. While this process did not produce many citations for this dissertation, it helped the author to better grasp the open data domain.

1b) Automatic search with search strings using literature search engines

In the second phase of the search process the author conducted and automated search using literature search engines. The author decided to use Google Scholar for this phase of the search process as it produced a large number of results. Science Direct, for example, showed 500 results (for TITLE(open) and TITLE(data) (all years)), whereas Google Scholar produced 4,810 results (*allintitle: "open data"*, exact phrase, excluding patents and citations, all years) and 8,910 papers not using phrase match. A large pool of publications was vital given the broad undertaking to plot the full open data domain and considering that basically every publication is relevant for the scope of this dissertation. By scanning through the results pages, titles and abstracts, the author identified three reoccurring preliminary themes or clusters:

A) Publications dealing with open data applications, open data portals/platforms or open data initiatives. E.g.: (Vasa and Tamilselvam, 2014)

B) Publications dealing with the technical mechanisms of open data. E.g.: (Kamateri *et al.*, 2014) or (Bizer, Heath and Berners-Lee, 2009)

C) Publications evaluating open data. E.g.: (Vetrò *et al.*, 2016).

Besides the three preliminary clusters, step 1b also provided some references that could be used in the course of the dissertation. Further, these papers also pointed to other papers that were highly useful for this dissertation (such as (Attard *et al.*, 2015) pointing to (Martin, Foulonneau and Turki, 2013)). Concluding, this part of the search process already provided a good amount of relevant papers and the author now wanted to find further relevant publications and fine tune the preliminary clusters.

1c) Refined automatic search

After having mapped each preliminary cluster, the author now redefined the search strings to identify further publications and to confirm and refine the preliminary cluster. Google Scholar was again used as the search engine and the search strings will be individually described below. Given the resource limitations of this dissertation, the author decided to identify 25 publications to support one cluster. If there are fewer than 25 publications, a cluster would be discarded. Further, in case there are more than 25 publications, the author would choose randomly which to include. Random selection is not uncommon in preliminary mapping studies e.g.: (Kitchenham, 2010). The literature list, subdivided for each cluster, can be found in appendix 1.

A) For open data applications a search string for *allintitle: "open data application"* returned only a small number (4 in total, where two were relevant). This is to be expected as an application based on open data would not necessarily name "open data application” in its title. Here, the author relied on the publications that were collected previously in 1b such as examples mentioned in (Manyika *et al.*, 2013). The second part of this cluster, open data platforms or portals, the search *“allintitle: "open data portal"*, *allintitle: "open data platform"* and *allintitle: "open data initiative"* returned 24, 26 and 23 results, which confirmed the existence of this cluster. For example, (Chen *et al.*, 2010) was collected with the search for *“allintitle: "open data portal"*. The author decided to cluster these four into one cluster given that they share an underlying goal: open data deployed to solve a problem or generate value in another form. Therefore, the cluster is called “Open data applications, platforms, portals and initiatives”.

B) Capturing further relevant papers for the preliminary cluster “technical mechanisms of open data” was tricky given that papers would not generally name technical mechanisms of open data in their title but very specific concepts. Through step 1b the author learned about the general technical aspects of open data. Reoccurring themes were linked data principles, metadata and data sharing. This knowledge was used to create the search strings for this cluster: *allintitle: "linked data"* (6740 results) that e.g. pointed to (Bizer, Heath and Berners-Lee, 2009), *allintitle: "open data metadata"* (9 results) that pointed to (Martin, Foulonneau and Turki, 2013) *and allintitle: "open data sharing"* (32 results) to (Poldrack and Gorgolewski, 2015). Further, the author found that such technical mechanisms are better described as enablers since that would include the technical prerequisites but also leave room for further work that would also consider e.g. sociological aspects for this cluster. Therefore, it was decided to call this cluster “Open data enablers”.

C) Publications for the cluster on evaluating open health data were easier to discover through search strings since there are no alternative descriptions but only synonyms. Therefore, also broader search queries were used: *allintitle: open data framework* (175 results), *allintitle: open data evaluation* (97 results). Naturally, evaluations also act as guidelines and therefore, the author decided to name this cluster to “Open data evaluations and guidelines”.

1d) Selection validation

For further check and balance and validation reasons, the author included the following step, which might be novel for mapping studies. The author used Mendeley as a referencing tool for this dissertation, which also provides further reading recommendations based on the user's library. These weekly recommendations were a good tool to check if the newly recommended publications would fit into the defined clusters. For example, recommendation (Tygel *et al.*, 2015) describes semantic metadata layers for open data and, therefore, fits into the "Open data enablers" cluster. Recommendation (Ohemeng and Ofosu-Adarkwa, 2015) describes the Ghana Open Data initiative and therefore fits into the defined cluster "Open data applications, platforms, portals and initiatives". This process did help to confirm the selection of the clusters.

2) Selection criteria

Given that a random number of articles could be scanned for a preliminary mapping study (Kitchenham, 2010), the selection criteria is not as well defined as it would be in a systematic mapping study. Therefore, the following selection criteria were followed:

1) Publications had to be in English
2) Posters, summaries of articles, presentations or tutorials were not included
3) Whitepapers, books and reports were included
4) Studies where abstracts were not accessible, were not included.

While this is not directly relevant to the selection of the publications, the author thinks that the following two paragraphs are best suited in this part of the methodology section. In the following the author describes how the scope was set for the dissertation and which aspects were included:

- Given the high degree of related topics and interconnectedness of topics in this dissertation, it was hard to set the right scope. Data security, privacy and de-identification, for example, were chosen to not be included in this dissertation as they would have exceeded the scope of this dissertation. For this dissertation we assume that personal identifiable information has been stripped from datasets, is appropriately anonymized, data privacy and data security are insured. A case where such mechanism did not work was presented in (Martin, Shah and Birkhead, 2017) and shall only be mentioned as a reminder how important the area is.

- Health data definitions, on the other hand, were included because the author thinks that they are important for the overall readability of the dissertation. This is due to one of the key findings of the dissertation, being that most open data evaluations only consider a specific domain of all available health data such as only considering public health data. Consequently, the scorecard developed in this dissertation considers different health data types. In light of this, it is important to equip the reader with a greater sense of the diversity of health data and, therefore, particular attention has been paid to describe the different health data types in section 3.3 Health data.

3) Extracting and clustering

In the final step, the author created a short description for the three emerged clusters and further sub-clustered each cluster. For example, open data applications dealing with medical images would be described in its own section and not combined with open data in the public health domain.

4) Discovering a research gap

While the preliminary mapping study would stop at step 3 “Extracting and clustering”, the author of this dissertation wanted to use the collected pool of knowledge to extract further insights that could be used to create a scorecard to evaluate open health data. This is not uncommon for mapping studies as Budgen et al. states:

“A mapping study is very time-consuming. Even for a PhD topic, it generally extends well beyond the time normally spent on background reading. In compensation, a thorough mapping study can itself provide additional opportunity for publication.”(Budgen *et al.*, 2008)

The dissertation uses this additional opportunity and tried to fill the exposed research gap, represented in research question two:

<u>What might be an appropriate scorecard for evaluating open health data?</u>

Based on the collection of rich resources from the mapping study, the dissertation is able to deliver a novel approach to scorecard open health data initiatives. This scorecard shall not be understood as a complementary one (in the sense of another scorecard evaluation already existing points of analysis (such as data quality)) but one that is tailored to the health domain. Since there is very little research that concentrates on evaluating open health data, benchmarking the outcome is only possible in a very limited way. Only further research and the research community's discourse can holistically evaluate the outcome. A similar situation is described in Burégio et al. 2010:

"However, with the intention of refining and constructing this work with feedback from the community of researchers involved in workshops, conferences and journals which we have been analyzed, we decided to develop this study in different phases." (Burégio, de Lemos Meira and de Almeida, 2010).

The following section of this chapter summarise the methodology and provides an overview in figure 2.1.

2.3 Summary

Concluding, the author conducted a preliminary mapping study to create a baseline for open health data. The baseline itself tries to answer "what do we know about open health data and how does it work". From this pool of knowledge, the author discovered a research gap and consequently built the second research question: "What might be an appropriate scorecard for evaluating open health data?". By extracting insights from the mapping study the author delivers a scorecard that should support the evaluation of open health data specifically. Figure 2.1 shows the methodology process that was followed for this dissertation and the steps conducted to answer the two research questions.

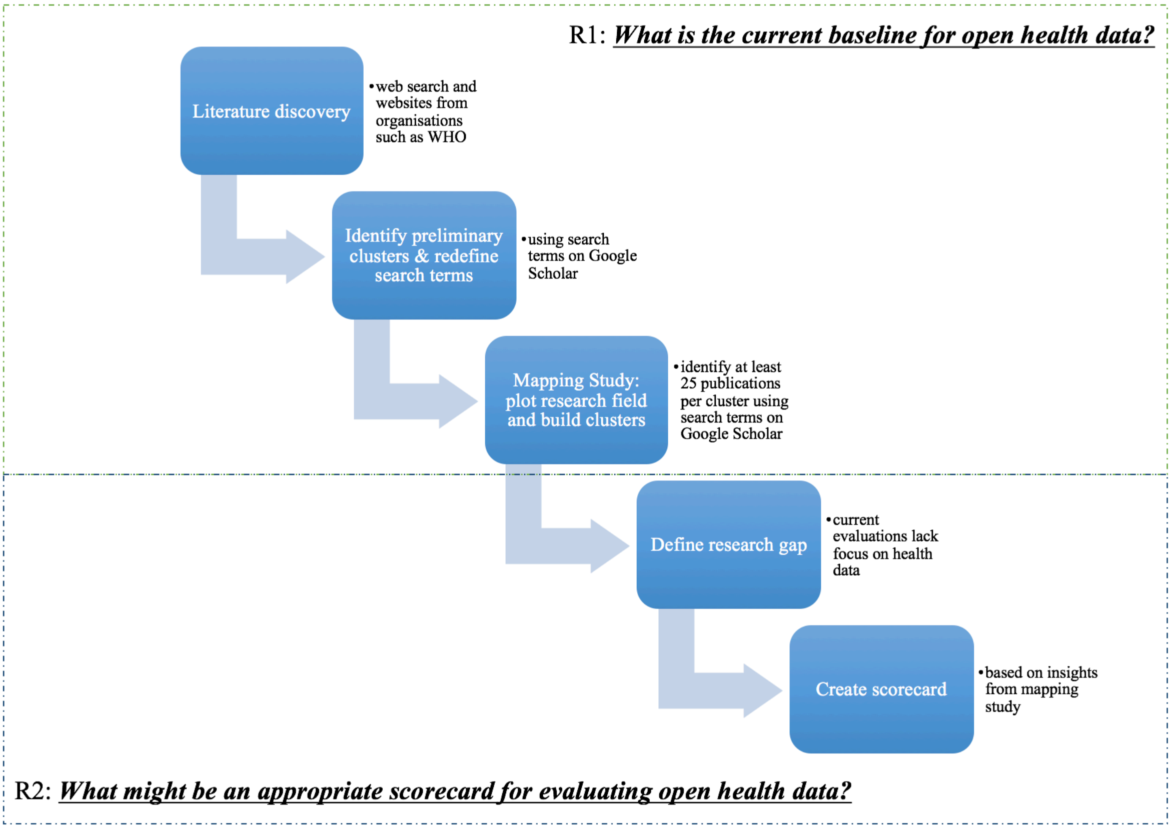

R1: What is the current baseline for open health data?
Literature discovery
•web search and websites from organisations such as WHO
Identify preliminary clusters & redefine search terms
•using search terms on Google Scholar
Mapping Study: plot research field and build clusters
•identify at least 25 publications per cluster using search terms on Google Scholar
Define research gap
•current evaluations lack focus on health data
Create scorecard
•based on insights from mapping study
R2: What might be an appropriate scorecard for evaluating open health data?

Figure 2.1: Describes the methodology process and the steps conducted to answer the research questions.

3 Preliminary mapping study

This chapter answers research question one, "What is the current baseline for open health data?", presented in section 1.2. Further, out of this research the scorecard was crafted and refined, which then answers research question two. The following pages present a rich and wide-ranging pool of knowledge for the domain of open data. Although the large scope and highly dynamic research field, the author hopes that this chapter will give any reader a great understanding of open health data that would be otherwise hard to collect.

As described in the methodology section, to lay the groundwork for evaluating open health data, a preliminary mapping study has been conducted. Out of the process three evidence clusters emerged that will be described in great detail in the following pages. First, the author will describe open data, related terminology and health data.

3.1 Open data definition

Open data can be described as a well-established paradigm and global trend to make data freely available to everyone (Vasa and Tamilselvam, 2014). Open Knowledge International further describe three core features of open data:

- Availability and access
  The data should be available completely and ideally made accessible as a download over the internet
- Re-use and redistribution
  Deals with the licence and terms of the data, which should permit the re-use, redistribution and data fusion with other datasets.
- Universal participation
  Describes the absence of discrimination against any potential data user or for certain purposes (e.g. education only)

(Open Knowledge International, 2017).

There are many related terms that are sometimes used synonymously, or represent concepts that are also relevant to the dissertation, such as: linked government data, linked data, linked open government data, linked open data, open government. The relationship between those terms has been described in Attard et al. 2015 focusing on open government data as shown in figure 3.1.

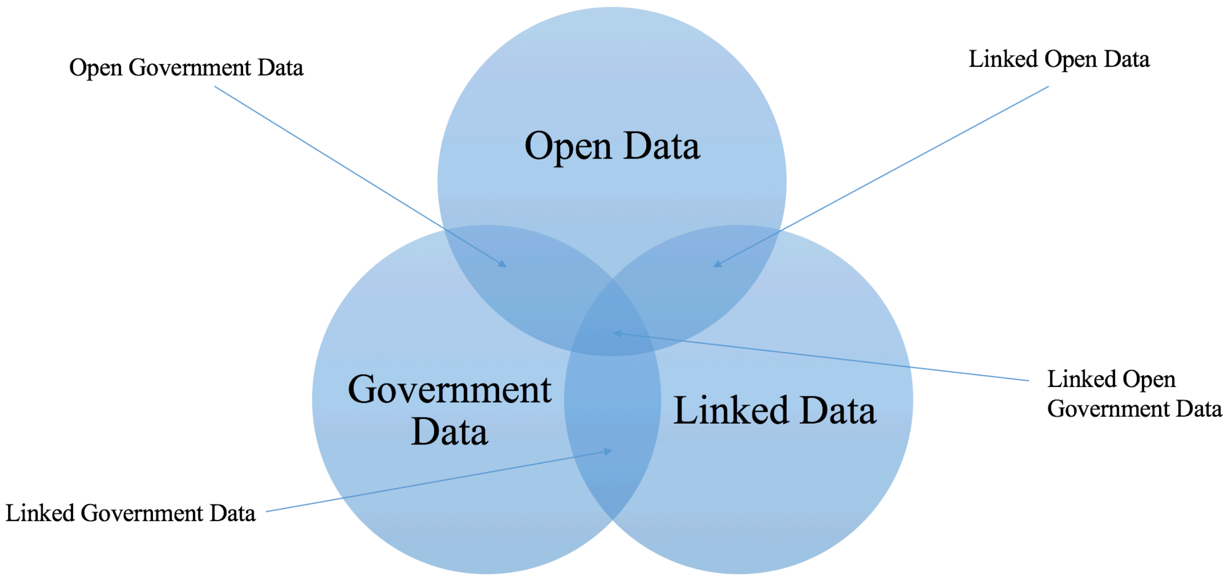


Figure 3.1 Open data definitions from (Attard *et al.*, 2015) describing the interplay between open data, government data and linked data.

Following, the dissertation will look at open data initiatives, which is also a frequent point of analysis in the literature.

3.2 Open data initiatives

Open data initiatives share a noble cause: increase transparency (Janssen, 2011), increase opportunity (Alexopoulos *et al.*, 2014) and increase collaboration. (Sandoval-Almazan and Gil-Garcia, 2014). Transparency in the context of open data, increases the accountability and auditability of leadership decisions. Nonetheless, increased transparency also inherits obstacles as described in “The Tyranny of Transparency”, a highly cited article (Strathern, 2000). Opportunity, on one hand, describes the rise of economic output due to commercial activities that are building services or products on top of open data (Attard *et al.*, 2015). On the other hand, opportunity in the context of health data, also entails the strive to increase the quality of care by discovering new knowledge in data sets (Holzinger, Dehmer and Jurisica, 2014). Collaboration in this context, not only describes the active participation of citizens in the governance process, often described as participatory government (Rojas, Bermúdez and Lovelle, 2014), but also researchers coming together to discover unknown insights in open data sets.

Before discussing open data in greater detail it is important to understand the specifics of **health data**. Health informatics can be defined as the concentration on the use of data in the healthcare process (Parry, 2014). Recent publications also acknowledge that the healthcare process nowadays more broadly spans around the well-being of individuals as well as the collective (Holzinger, 2016). Given this more holistic understanding and developments such as the quantified self, this work will use the term health data (instead of healthcare data) as a way to describe data that is medically relevant to one's well-being.

Like many concepts in health informatics, it is not universally agreed how health data sources are defined and classified. An excursion in the fields of research in the health domain might help answer this question. The dissertation will, therefore, offer two existing definitions of health data.

3.3 Health data

Health data in the context of open data is often understood as the health sector's performance e.g. patients' satisfaction with health or infections rates on a country level (Open Data Barometer, 2015) (Global Open Data Index, 2017). Inherently, the area of analysis in published related work is often public health data. Still, there is promising work done in other areas of health data such as making medical images (Clark *et al.*, 2013) or drug data for pharmaceutical research (Samwald *et al.*, 2011) freely available. This represents a potential research gap since data quality has different aspects in different health disciplines such as biomedical data and neurological image data. Understanding and scoping the data ecosystem is of eminent importance not only for health informatics in general (Holzinger, 2016) but also when trying to evaluate open health data. In light of this, the dissertation will describe two existing definitions of health data.

3.3.1 Holzinger defintion

Holzinger 2016 (Holzinger, 2016) defines four large data pools in the health domain based on how the data is produced:

- Biomedical research data, including, e.g. -omics data from e.g. genomic sequencing technologies or microarrays.
- Clinical data, including, e.g. medical survey, laboratory test, patient records or clinicians' documentations.
- Health business data, including, e.g. management data, costs or utilization.
- Private patient data, including, e.g. wellness data, sports data and insurance data.

3.3.2 Herland et al. definition

Herland et al. (Herland, Khoshgoftaar and Wald, 2014) defines four pools of health informatics data based on four levels:

Molecular – Bioinformatics

Bioinformatics focuses on data that is produced on the molecular level. Common forms are gene expression data streams, which are increasing at a fast rate due to mainstream availability of genomic test (Fulda and Lykens, 2006). Additionally, the variants of such data streams are increasing because of technical innovations that allow the recording and generating of more molecular data per person (Herland, Khoshgoftaar and Wald, 2014).

Tissue – Neuroinformatics

Neuroinformatics deals with tissue level data based on brain images. Such images are commonly (Health Service Executive, 2017) generated through MRIs (Magnetic Resonance Images) and CTs (Computer Tomography). Generally, users of neurological data try to establish a correlation between medical events and brain image data. Besides the disciplines application in health, neurological data also acts as a rich source for insights in other areas. For example, recent research tackles the dual-task interference phenomena, a cognitive limitation preventing the high performance of simultaneous tasks, by using functional magnetic resonance imaging (Jenkins *et al.*, 2016).

Patient – Clinical informatics

Clinical Informatics surface around making use of patient data to enhance the clinical decision making processes. Data is generated and recoded through professional's examinations, medical devices or sensors.

Population – Public health informatics

Public Health Informatics leverages macro level health data produced for example by health facilities (e.g. available hospital beds, infections) or newer mediums such as social media.

After having described important terminology for this dissertation, the author will now explain the three clusters that emerged out of the preliminary mapping study.

3.4 Open data applications, platforms, portals and initiatives

These publications describe open data activities that aim to solve a problem or generate value in another form. They include open data applications, open data platforms, open data portals and open data initiatives.

3.4.1 Other industries

Before depicting health specific cases where open data was applied to create value or solve a problem, the dissertation will start by demonstrating instances form other areas and industries. This was important, since insights can be gathered from other domains and, further, to give the reader a better understanding of the diverse application areas of open data.

Looking at education, we can find an example of how open data was used to build an algorithm that supports parents in finding a school for their children in the city of Boston (Shi, 2015). Historically, this presented a challenge for parents trying to find the right school for their children, balancing quality and location. The system has been called "creative and dynamic" and able to stand the test of time (Dajer, 2012).

In transportation, the publication of open data such as public transport schedules and transit data is embraced by many transport agencies (Kaufman, 2012). In the United Kingdom for example, data published by Transport for London is being used by 8500 developers and powers nearly 500 apps. Further, 42% of London's population uses these apps (tfl.gov.uk, 2017). As the Transport for London's commitment to open data they list that their data is publically owned; creates economic benefit by facilitating the development of new inventions; enhances their reach, ensuring that any person can get travel information and crowdsourcing innovation through their open data initiatives. In addition to transportation data, they also publish related data such as air quality.

An example related to transportation is the OpenStreetMap project. The OpenStreetMap project is a knowledge collective (Haklay and Weber, 2008) that crowdsources map data via a large number of contributors on its platform. Crowdsourcing, in general, is a vital concept in the open data world. The OpenStreetMap project is powered and facilitated by an underlying platform that enables the crowdsourcing process and the capturing of the generated data. Humanitarian operations in disaster response situations often rely on data from Open Street Maps such as during the Haitian Earthquake in 2010. Great parts of Haiti lacked maps coverage on popular map services during the time of the crisis. Contributors used satellite images to identify streets and areas that supported the humanitarian work on the ground in locating people in need (Zook *et al.*, 2010). Within a few weeks 10,000 edits were made by hundreds of people around the world (Keegan, 2010). The OpenStreetMap community also regularly conduct "hackathons" where volunteers meet up to improve the cartographical data (openstreetmap.org, 2017a). The data that is created and on the platform itself becomes open data: *"OpenStreetMap is open data: you are free to use it for any purpose as long as you credit OpenStreetMap and its contributors."*, reads the projects' website (openstreetmap.org, 2017b).

In energy, McKinsey, a management consulting firm, estimates that open data in electricity has the potential to add up to $580 billion value per year across the electric value chain. The biggest opportunity within that amount lies within the lower funnel of the value chain where price transparency and energy efficiency might unlock up to $310 billion. One example would be the investment in new electricity sites. Here, construction planners already rely on publicly available data such as historic weather data to determine how much energy can be generated at different locations. Open data could provide a more complete picture of the total economic and environmental costs. Further, the report identifies five sources of open data: power suppliers, regulators, government agencies, energy users, third party data brokers (Manyika *et al.*, 2013).

Encouraged by the rising inflation on food prices in India, developers created an app, heavily based on open data, that would show the user tailored recipes. These recipes would factor in seasonal prices, proximity to closest store and scan the products for potential allergies. The Rasoi app was built with open source platforms, libraries and web servers such as Java, JQuery and Apache Tomcat (Vasa and Tamilselvam, 2014).

Research by Andrew Whitmore describes a case where open government data could be used to predict wars. The prediction attempt is made by looking for patterns in military spending and extracting insights out of the historic data (Whitmore, 2014).

Before jumping into what might be considered “typical” health related applications (e.g. medical imaging, hospital beds and waiting lists), one more application will be presented, which demonstrates how much data is related and influencing health. The following example of the British Environment Agency (EA) is a good example of areas that are often a good starting point for publishing health related data. This is due to the fact that they often inherent less obstacles and barriers given that there is no personal identifiable information involved.

The British Environment Agency (EA) weekly publishes evaluations of water quality in bathing areas such as ponds (UK Environment Agency, 2017). Each area, like the Serpentine in Hyde Park, is attributed a profile page that shows historical development and ranks the quality with a three-star rating. Since the published data is powered by linked data principles, mashups with other data sets such as air quality or accommodations are feasible and much easier compared to open data sets that are not based on linked data. Additionally, profiles of each location are built based on this liked historical data (Shadbolt and O’Hara, 2013). Linked data presents an important aspect of open data and will be further described in section 3.5 Open data enablers.

Having looked at examples from other industries, the dissertation will now identify applications that are related to the health domain.

3.4.2 Cancer research

“Integrating Open Data on Cancer in Support to Tumor Growth Analysis” describes an interface that links a cancer modelling simulation to additional relevant open data. This process of data integration and fusion is a key research direction in health informatics with the aim to discover new knowledge (Holzinger, Dehmer and Jurisica, 2014). Here, the researchers scanned open data sets for tumour growth related parameters and subsequently fused the data into a simulation tool modelling neoplasms. Such simulations unveil abnormal tissue growth and support a better understanding of the tumour's growth patterns. In total more than 33 databases were identified that were often available as a non-commercial and scientific purpose only download. The used databases are quite diverse including:

- Genomic data sources such as "Catalogue of somatic mutations in cancer" (Forbes *et al.*, 2015) or "Integrative Onco Genomics" (Rubio-Perez *et al.*, 2015)
- Imaging data sources including "The Cancer Imaging Archive" (Clark *et al.*, 2013) and "CancerData.org – Sharing data for cancer research" (Roelofs *et al.*, 2014)
- Incidence data from "WHO Cancer Mortality Database" (WHO, 2016) and "Cancer Incidence in Five Countries" (Bray *et al.*, 2015)
- Disease associations data from "NCI Thesaurus" (Sioutos *et al.*, 2007)
- Literature data from "PubMed Central" (US National Library of Medicine - National Institutes of Health, 2017)

3.4.3 Medical images

The previous mentioned example of the Cancer Imaging Archive (TCIA) is an interesting example of how the complexity of open data (The World Bank, 2017) can be overcome. Here, 23 collections containing 3.3 million images were made available in its first year of operation.

In total more than 400 publications originated out of these data sets (Cancer Imaging Archive, 2017). Given the diverse range of publication, four publications will be depicted at this stage that represent a diverse range of applications:

- "A Comparison of Lung Nodule Segmentation Algorithms: Methods and Results from a Multi-institutional Study" used five collections from the Cancer Imaging Archive consisting of 52 tumours in 41 CT volumes for their study. Algorithms from three academic institutions were tested and showed a considerable difference, especially in a subset of heterogeneous nodules, leading to the papers recommendation that the same software should be used consistently during longitudinal studies (Kalpathy-Cramer *et al.*, 2016).

- "Automated Medical Image Modality Recognition by Fusion of Visual and Text Information" used the Cancer Imaging Archive data to perform a test of their medical image classification framework. The data aided the validation of the framework, which confirmed that it yields state of the art performance (Codella *et al.*, 2014).
- Starting in the 1950ies (Samuel, 1959), machine learning is applied in situations where predictions shall be made based on knowledge extracted from data (Holzinger, 2016). Today machine learning can be described as the fastest growing field in all of computer science (Jordan and Mitchell, 2015) and health informatics is called out as one of the greatest challenges (LeCun, Bengio and Hinton, 2015) (Holzinger, 2016). One of the challenges is that large amount of data is needed to train unsupervised algorithms (Holzinger, 2016) and useable big datasets are rarely available in the health domain (Herland, Khoshgoftaar and Wald, 2014). Hence, the Cancer Imaging Archive's data also shows to be highly valuable for this frontier in computer science. In this space, a publication from Pang et al. used data from the platform to test their medical image classifier to train their deep neural network (Pang, Yu and Orgun, 2017). Another case was presented in 2013 at the Conference on Neural Information Processing Systems, regarded as the flagship meeting on machine learning and neural computation (The MIT Press, 2017) and titles: "Adaptive Multi-Column Deep Neural Networks with Application to Robust Image Denoising". In the publication, the authors present a new technique for removing noise out of corrupted images, where data from the Cancer Imaging Archive could be leveraged (Agostinelli, Anderson and Lee, 2013).

Another interesting aspect of the Cancer Imaging Archive is the platform's different access layers. Data access layers play an important part in the health domain since there are ample cases where it would be beneficial to share health data only with researchers or healthcare professionals (Li *et al.*, 2010). Here the open data definition might reach its limits for health data.

In the following, the author will depict two examples from the Cancer Imaging Archive and their sharing options. The 32.9GB collection "TCGA-BLCA" dealing with bladder endothelial carcinoma, is freely available ("No restrictions; all data available without limitations") on the website without registration or other access checks. On the contrary, the 1.7GB collection "NRG-1308" dealing with non-small cell lung cancer is only available as a limited access data set ("This is a limited access data set. Upon receiving access you may only use it for the purposes outlined in your request to the data provider."). All these data sets are available in the medical imaging standard "Digital Imaging and Communication in Medicine" (DICOM). DICOM has become one of the most popular standards in medicine and enables more services to be integrated in information systems such as RIS (Radiology Information System) and PACS (Picture Archiving and Communication System) (Mildenberger, Eichelberg and Martin, 2002).

3.4.4 Biomedical research

Further datasets and projects related to pharmaceutical research and development were identified in Samwald et al. (Samwald *et al.*, 2011). Samwald et al. also acknowledges that such datasets on the internet, ranging from medicinal chemistry results to the impact of drugs on gene expression, are typically not interconnected. This reduces the ability to extract knowledge and insights. Therefore, the LODD (Linking Open Drug Data) taskforce, a group within the World Wide Web Consortium's (W3C) Health Care and Life Sciences Interest Group (HCLS IG), identified publicly available datasets about drugs and created linked data representation of these datasets (Samwald *et al.*, 2011). The dissertation at hand will describe some of these in greater detail to give the reader a better understanding of the depth of health relevant data that is available. Next the dissertation will depict some of these and provide the full list in Appendix 2.

- The DrugBank, currently available on www.drugbank.ca, is a bioinformatics and cheminformatics resource that combines drug data with drug target information. The database includes 8261 drug entries, 2021 of those are FDA-approved small molecule drugs, 233 FDA-approved biotech drugs (such as proteins), 94 nutraceuticals and more than 6000 experimental drugs. These drug entries are linked to 4338 non-redundant protein sequences and each drug entry card entails more than 200 data fields (Wishart *et al.*, 2006).

- LinkedCT is a project that aims to be the first open semantic web data source for clinical trials. The core dataset is derived from ClinicalTrials.gov, a registry for clinical trials around the world, which was used for the linked data generation process. (Hassanzadeh *et al.*, 2009)

- ChEMBL covers a diverse range of annotated and curated data, which has been extracted from primary medical chemistry literature. The data includes biological readouts such as metabolism and excretion assay measurements. Besides the information extracted from the literature, ChEMBL also integrates deposited screening results from public databases such as PubChemBioasay (Wang *et al.*, 2014). The project has also been called out as transforming the landscape of available medicinal chemistry data (Gaulton *et al.*, 2012) (Bento *et al.*, 2014) (Gaulton and Overington, 2010) (Papadatos and Overington, 2014).

- The TCMGeneDIT project mined research articles for gene-disease-drug associations focusing on Traditional Chinese Medicine (TCM) and consequently published the records as linked data on the internet. Natural language processing tools and rule-based approaches were used to extract possible relationships between TCM effecters and effect (Zhao, 2010) (Fang *et al.*, 2008).

- Chem2Bio2RDF extends the linkage of biological data and drug data to chemogenomic and systems chemical biology information. The projects interlinks aggregated data from various chemogenomics repositories and cross-links them into Bio2RDF (Belleau *et al.*, 2008) and LODD (World Wide Web Consortium, 2012). Further, the paper shows potential use cases such as in the case of adverse drug events and on the identification of multiple pathway inhibitors (Chen *et al.*, 2010).

Besides the above examples from Samwald's 2011 publication more projects and databases can be found such as DBpedia. At its core, DBpedia aims to extract structured information from Wikipedia and link different data sets on the Web to Wikipedia data. Additionally, DBpedia allows the performance of advanced search queries (dbpedia.org, 2017). Research from Negru & Buraga shows a tool 'RDFSpecies' that emphasizes life science data on the platform. It is indented as an educational tool allowing the interactive exploration of sematic web data (Negru and Buraga, 2013). The authors also note that the number of linked biomedical datasets has grown drastically in the last years and that there are several dozen biomedically linked datasets available up to date (Samwald *et al.*, 2011).

3.4.5 Open health data in the public health domain and business health data

There are fewer applications of public health open data and business health data for a number of reasons. The first being that such applications are often displayed in grey publications and reports compared to journals (e.g.: (Stefaan Verhulst *et al.*, 2014)). Secondly, the titles of publications, containing these applications, do often not include the term open data but talk about the broader problem, such as hospital waiting list problems, and include open data as one of the solutions to tackle the problem (e.g.: (British Columbia Medical Association, 2006)). Thirdly, open data applications in the public health domain specifically, are only one example of open health data applications generally. Further, open data applications are only one evidence cluster out of three, which emerged from the preliminary mapping study. Therefore, the chapter does not claim completeness for all open health data applications or application areas but shall give the reader an understanding of the diverse applications of open data in the health domain. This overview also acts as a rich resource for further research. Now that the evidence maps have been defined, further systematic mapping studies could be done for e.g. applications of open public health data specifically. Consequently, systematic literature reviews for specific applications in specific domains, such as waiting lists in the public health domain, can be performed.

The report “Open data: Unlocking innovation and performance with liquid information” from McKinsey & Company, a consulting firm, provide examples of applications in the public health domain. One describes public health agencies collecting data from sources like emergency rooms as a mean to detect disease outbreaks. This data is then used to inform the public so they can take containing and preventing steps. Another example describes providing comparison information to patients so they can choose the most cost-effective care, enabling the patients to become better healthcare consumers and control their costs. Some bodies have already made hospital performance data and data on different types of care available. Such data can be leveraged by e.g.: doctors, insurances and patients to measure and track performance but also to identify the best care available in their communities (Manyika *et al.*, 2013). While the following phenomena is not the focus area of this dissertation it is important to stress the complexity of such data sets, how hard it is to correctly understand patterns and what implications the tracking of such data can have. Jessica Nutik Zitter (Jessica Nutik Zitter, 2017) describes in “Misleading Metrics” a New York Times article “A Surgery Standard Under Fire” form Paula Span (Paula Span, 2015). The authors describe that surgeon’s performance is often measured with a 30-day mortality statistic. In other words, the better the surgeon, the more likely patients will live post 30 days after the surgical procedure. The statistic is described as being widely used by insurance companies, health agencies and often publicly reported. What is more, health consumer often use such data when deciding where to seek health care and treatments (Ketelaar *et al.*, 2011). Jessica Nutik Zitter describes that this poses an incentive to keep patients alive, although their preferences might be different. The concern is further emphasized by a statement from a surgeon conference: “I can’t operate on some people because it’s going to hurt our 30-day mortality statistics.” (Paula Span, 2015). The matter is also described in research called “Beyond 30-day mortality: aligning surgical quality with outcomes that patients value” (Schwarze, Brasel and Mosenthal, 2014) . While, as described earlier, this is not directly the topic of the dissertation, it is certainly an implication that needs to be closely monitored. Further research could undercover similar examples.

Concluding, section 3.4 Open data applications, open data platforms, open data portals and open data initiatives, identified diverse cases of applications that leverage open data to create value or solve a problem. Whether it is the transportation industry or health industry or whether the aim is to increase customer satisfaction or to save lives, the application of open data is diverse and wide ranging.

3.5 Open data enablers

The second cluster "Open data enablers" describes the underlying technical mechanisms that make open data work. The name has been chosen broadly to allow for further research that also includes socioeconomic aspects such as economic or legal parameters. First, the author will describe linked data and semantic web principles as an open data enabler.

3.5.1 Linked data & semantic web principles

The desired underlying technical mechanics of open health data are linked data principles (Tim Berners-Lee, 2006). The borders, similarities, differences and relationships of related terminology, such as linked open government data, open government data or open data in general, will be established and described in section 3.1 Open data. Resource pressures often lead to open data being published in the easiest available format such as PDF, which, to a great degree, doesn't provide the full set of possibilities that inked data does (Shadbolt and O'Hara, 2013). In fact, in a survey of open government data stakeholders, the second most favoured data format is said to be PDF. Further, 52% named HTML, CSV/XLS represented 37%, DOC/RTF 32%, XML 27%, APIs 22% and RDF 18%. The most requested formats for the future are APIs, RDF and XML (Martin and Kaltenbock, 2011). EHealth strategies often include the objective to break through data silos (e.g. (eHealth Ireland, 2017)), which is also one of the potentials of linked data. By using HTTP (Hypertext Transfer Protocol) URIs (Uniform Resource Identifier) and referencing each entity (like a ZIP code) with a single URI creates a consisted linking, which demolishes data silos. It is important to stress the singularity of the URI at this point since multiple identifiers are often used to point to one object (Shadbolt and O'Hara, 2013). Further, the authors (Shadbolt and O'Hara, 2013) argue that the plurality of URIs for the same objects are inevitable in a decentralised world and should at least apply to core reference data. Recent research proposes a linked data approach as a way of establishing interoperability of Clinical Decision Support Systems (CDSS) (Marco-Ruiz *et al.*, 2016). While Service Oriented Architectures (SOA) have been presented in different cases as a way to reuse CDSSs by condensing them in a web service (Kahn and Hederman, 2012) & (Marco-Ruiz *et al.*, 2016), linked services that expose their interface as linked data can overcome the limitations of the syntactic nature of Web service technologies and create a common linked knowledge base that is discoverable through intelligent queries (Marco-Ruiz *et al.*, 2016). Related research has proposed a semantic-web oriented representation of the clinical element model for the secondary use of electronic health records data. The Clinical Element Model (CEM) enables the representation of clinical information, stored in electronic health records, across different organizations (Tao *et al.*, 2013). Since the current representation of CEM did not support formal semantic definitions, the paper introduced a Web Ontology Language (OWL) representation of the CEM specification enabling the sematic web environment with tools like authoring, reasoning and querying (Tao *et al.*, 2013). In the pharmaceutical space Samwald et al. (Samwald *et al.*, 2011) purposes linked data as a foundation for pharmaceutical R&D data sharing. The task force within the World Wide Web Consortium (W3C) Health Care and Life Sciences Interest Group (HCLS IG) identified public drug related data and created a linked data representation of these data sets. Through this linking process new scientific and business questions could be answered and best practices were recorded (Samwald *et al.*, 2011). Research at the Mayo Clinic presented a case study demonstrating publicly available linked open drug data being combined with real patient data from electronic health records for type 2 diabetes patients. The results present a powerful platform for data integration and pooled querying (Pathak, Kiefer and Chute, 2012). While linking and integrating data presents immersive opportunities at the same time the subject faces significant ethical and legal challenges that call for access controls (Kamateri *et al.*, 2014). Kamateri et al. propose the Linked Medical Data Access Control framework (LiMDAC) that leverages linked data principles to create access models across distributed services (Kamateri *et al.*, 2014).

Next, the dissertation will describe sharing as an open data enabler.

3.5.2 Sharing

Central to the open data enablers described here is the principle of sharing. An essential lever to determine the effectiveness of open data lies in the degree of how automated the data can be processed by a machine. This becomes apparent when considering big open data sets that could contain billions of data points. No human could process and make sense of these in a short period of time. Another vivid example from the open data space are freely available archives of legal documents. Although many are publicly available, identifying and locating a specific document or patterns becomes an immense effort vs. a machine able to execute search tasks in a fraction of a second (Walker *et al.*, 2005). Walker et al., therefore, defines four levels of automation when data is shared. Further, Walker et al. differentiates between the amount of human involvement required, the level of standardization and the sophistication of IT. Level 1 describes non-electronic data that uses a minimal degree of information technology to share data, for example: postal mail. Level 2 depicts machine transportable data that uses non-standardised information such as PDF documents. Level 3 shows machine organisable data, which is transmitted through structured messages containing non-standardised data, for example, emails containing free text. The final Level 4 describes machine interpretable data having structured messages, standardised and coded data. What is more, Walker's et al. 2005 publication is already applied to the health domain and, therefore, of high relevance to this dissertation.

Table 3.2: Walker's (Walker *et al.*, 2005) levels of data automation applied to health data, table organization: (Sprivulis *et al.*, 2007)

Level	Data	Definition	Example
1	Non-electronic data	Min. use of IT to share information.	Mail, telephone.
2	Machine transportable data	Transmission of non-standardized information via basic IT. The information within the document cannot be electronically manipulated.	Fax, PDF.
3	Machine organizable data	Transmission of structured messages containing non-standardized data; requires interfaces to translate data from the sending organization's	E-mail of free text, exchange of files in incompatible/ proprietary file formats.

		vocabulary to the receiving organization’s vocabulary.	
4	Machine interpretable data	Transmission of structured messages containing standardized and coded data; systems exchange information using the same formats and vocabularies.	Automated exchange of coded results from external laboratories into an electronic medical record, automated exchange of the patients “active problem” lists between providers.

Further, based on the mentioned examples it becomes clear that machine readability of released data sets is vital to the overall usefulness of the data sets. Since Walker et al. 2005 provides guidance on the different levels of automation for data sharing that has already been applied to the health domain, it will be included in the final scorecard section 4.4 Data Processability. Specifically, with the question: “State of machine readability based on Walker (Walker *et al.*, 2005)?”. At this stage the reader might ask why this table is not included in cluster 3.6 Open data evaluations and guidelines given that the four levels from Walker (Walker *et al.*, 2005) could also be classified as an evaluation framework. The boundary, which frameworks to include in 3.6, has been set to include evaluations and guidelines that specifically assess open data. While the sharing framework from Walker et al. here is related, it is not directly assessing open data aspects. Rather, the dissertation makes use of the framework to introduce processability as an important dimension of open data and its usability. Correspondingly, linked data is described in this section as a technical principle but, an evaluation framework used to evaluate linked data in the context of open data, is introduced in cluster 3.6 Open data evaluations and used in the scorecard.

While these levels describe technical aspects of data sharing, there are other areas of focus such as described in van Panhuis et al. (van Panhuis *et al.*, 2014) about public health data sharing. Van Panhuis et al. classifies six barriers: technical, motivational, economic, political, legal and ethical. As described in the methodology section, this dissertation focuses on the technical aspects of open health data but provides room for further research on such on the barriers described above.

Concluding, the importance of sharing in the health data domain shall be stressed due to the high costs involved, for example when producing medical images (Resnick, 2017).

3.5.3 Metadata

Metadata is defined as "data that provides information about other data" (merriam-webster.com, 2017) and an important contributor to the effectiveness of open health data as the following example will depict. Further, metadata in the context of open data has been described as a critical enabler of data reuse (Martin, Foulonneau and Turki, 2013).

The OpenfMRI database (http://www.openfmri.org) is an open repository for task fMRI data. Task-based fMIRs are of the primary tools of cognitive neuroscientists that provide a way to probe the neural basis of mental functions and representations (Poldrack *et al.*, 2013). In a related publication (Poldrack and Gorgolewski, 2015) the authors highlight that a significant amount of metadata is required to make sense out of the raw task fMRI data. Metadata in this case includes description of task events and their timing. The platform (http://www.openfmri.org) published their own formatting standards for metadata. Contributors have to follow these metadata formatting standards in order to be considered. Afterwards, each data set is validated through a manual curation process that assures that the supplementing metadata are appropriately defined.

This example shows how important proper metadata hygiene is especially in the health domain. As the authors noted (Poldrack and Gorgolewski, 2015) the fMRI data would not even be useful without the proper metadata. Breaking down and defining specific metadata for all health data types (outlined in section 3.3 Health data) would exceed the scope of this dissertation. However, defining health data specific metadata in the context of open health data, represents an interesting question and opportunity for further research. What is more, a publication (Martin, Foulonneau and Turki, 2013) evaluating metadata in the context of open data, based on Tim Berners-Lee's five star rating (Tim Berners-Lee, 2006), was identified and will be further described in section 3.6 Open data evaluations and guidelines.

Kamateri et al. 2014 proposed a linked medical data access control framework called the LiMDAC framework that consists of three linked data models: LiMDAC metadata model, LiMDAC user profile model, and the LiMDAC access policy model. Further, Kamateri et al. offers a suiting definition for metadata:

"Metadata enable data providers to express richer access constraints and data consumers to perform more expressive search queries.".

The metadata set for the described research was included in the RDF data cube vocabulary, which defines metadata related to the cube structure (Kamateri *et al.*, 2014).

Concluding, this section described open data enablers that allow the effective facilitation of open data activities. The described enablers are metadata, sharing, linked data and semantic web principles. Following, the dissertation will describe the third cluster: open data evaluations and guidelines.

3.6 Open data evaluations and guidelines

Publications in this section include evaluations and guidelines to assess aspects of open data and represent the third cluster of the preliminary mapping study.

This section will analyse frameworks that evaluate open data activities. These could be a holistic index like the Global Open Data Index, a journal that deals with a specific branch such as open government data initiatives or a report on the progress of a specific country. The point of analysis of these could be diverse ranging from measuring the value of open data initiatives to measuring socio-cultural and policy issues that derive from open data initiatives. This section shall provide the reader with different angles and strategies of how open data undertakings can be evaluated and scored.

While open data can be analysed from many angles, many publications address and evaluate quality aspects of open data (Kučera, Chlapek and Nečaský, 2013), (Vetrò *et al.*, 2016), (Behkamal *et al.*, 2014), (Kubler *et al.*, 2016), (Calero, Caro and Piattini, 2008), (Umbrich, Neumaier and Polleres, 2015), (Máchová and Lněnička, 2017), (Zuiderwijk-van Eijk and Janssen, 2015)and (Umbrich and Neumaier, 2015).

For example, the Vetrò et al. (Vetrò *et al.*, 2016) approach to building a government data quality framework comes three folded:

1) identify a theoretical data quality model to support the proposed framework
2) derive a subset of quality characteristics from step one
3) map results from the selection of data quality characteristics and metrics from step two with issues emerging from an exploratory survey.

Further Vetro et al.'s research describes the limitations of other frameworks and highlights research gaps they are trying to fill. One of the identified limitations is that other existing studies lack the reference of an underlying theoretical framework with definite definitions of quality characteristics. The evaluation framework proposed by Vetro et al, therefore, builds on top of a quality model with the mentioned characteristics. The other limitation they highlight is that most frameworks do not consider the most granular level of measurement, the cell level (based on a tabular representation). In light of this, Vetro et al.'s work assesses the cell level of open government data instead of assessing a platform as a whole. Where the cell level evaluation is not possible, the dataset level will be assessed instead in Vetro et al.'s publication.

The data quality model chosen by Vetro et al. for step one of their process for creating an evaluation framework is called SPDQM i.e. SQUARE Aligned Portal Data Quality Model (Moraga *et al.*, 2009). Similar research from Behkamal et al. (Behkamal *et al.*, 2014) used "SQuaRE" – ISO 25012 (International Organization for Standardization, 2008) as the data quality model. The SPDQM itself was built upon the Portal Data Quality Model (PDQM) (Calero, Caro and Piattini, 2008) and SQuaRE standard. Among others, SPDQM was chosen because it represented the best set of characteristic that were often also used by other models. Typical characteristic are: accuracy, completeness and timeliness (Scannapieco and Catarci, 2002).

Research by Neumaier et al. (Neumaier, Umbrich and Polleres, 2016) reports on an automated monitoring tool to assess the data quality of open data portals continuously. Further their work describes six quality metrics that focuses on metadata:

- Retrievability – of metadata and resources
- Usage – of available metadata keys to describe a dataset
- Completeness – to which meta data keys are not empty
- Accuracy – extend to which meta data values accurately describe the resources
- Openness – usage of licences and file formats that conform the open data definition
- Contactability – of the data publisher through contact information

Similar metrics have been presented by Reiche et al (Reiche, Höfig and Schieferdecker, 2014).

Although not explicitly labelled as an evaluation framework, the eight open government data principles also provided relevant input to this section. The principles are:

- Complete – all public data should be made available (with regards to privacy, security or privileged information)
- Primary – data are as collected at the source, without modification
- Accessible – data are available to diverse users and for diverse purposes
- Machine processable – reasonable structuring allows automated processing
- Non-discriminatory – data are available to everyone without registration
- Non-proprietary – the data format is not exclusively controlled by one entity
- Licence-free – there is no copyright, patent, trademark or trade secret regulation applied to the data. (with regards to privacy, security or privileged information)

(Open Government Working Group Meeting in Sebastopol, 2007)

Tim Berners-Lee (Tim Berners-Lee, 2009) presents a 5 star framework that provides a roadmap for moving from open to linked data (Shadbolt and O'Hara, 2013):

1) "On the web, open licence". The data is available on the web with an open licence and in any format
2) "Machine-readable data". The data is available as machine-readable structured data, for example as a spreadsheet table instead of a scanned document
3) "Non-proprietary format". The data is available in a machine-readable structured data format and additionally, as a non-proprietary format. For example, CSV.
4) "RDF standards". The data is available in a machine-readable structured data format, in a non-proprietary format and, additionally, uses open standards from W3C. For example, RDF and SPARQL standards.
5) "Linked RDF". The data is available in a machine-readable structured data format, in a non-proprietary format uses open standards from W3C and links data to other data to provide a greater sense of context.

Tim Berners-Lee also notes that linked data does not generally have to be open. Linked data could also be applied within organizations or between groups. So data could get a five star rating without being open. To classify as linked open data, however, data needs to be open and get at least one star (Tim Berners-Lee, 2009). This is important in the health domain considering that there are many instances where data should not be open due to attached personal identifiable information. Still, there are various instances where data is trapped within one department and cannot be shared with the wider health organization. Here, linked data principles applied locally could also vastly support knowledge, information and data sharing internally.

Tim Berners-Lee’s framework has also been applied in research by Martin et al. who assessed the quality of metadata properties of open data sets in Europe as well as evaluates the level of data openness (Martin, Foulonneau and Turki, 2013).

Concluding, the chapter described a diverse selection of open data evaluations. Systematically evaluating all existing open data evaluations offers vast opportunity for further research. No evaluation could be found that focuses on open health data specifically. This might represent a research gap and will be further discussed in the next section.

3.6.1 Research gap

By analysing the frameworks above, the author was able to extract insights and a potential research gap. The research gap and findings informed the scorecard that will be described in chapter 4 Developing a scorecard. The author found that:

Current frameworks do not focus on open health data specifically but open data generally.
The preliminary mapping study found that current evaluations focus on open data in general versus a specific data types such as health data. While some focus specifically on government open data, the ones reviewed do not break down the specific parts of open government data such as open government health data. In fact, no publication could be found that assesses open health data. The dissertation at hand tires to fill this gap with the scorecard presented in chapter 4. The dissertation even takes a step further by differentiating different types of open health data such as biomedical and clinical health data. While elaborating on all the specifics for each health data branch would exceed the scope of this work, it is important to acknowledge and point out that there are differences that have to be taken into account. This presents great opportunity for further research, where the scorecard's sections could be field with specific information for the specific health data type.

The literature that deals with open health data often only considers public health data
Health data in the context of open data is often understood as the health sector's performance e.g. patients' satisfaction with health or infections rates on a country level for example in: (Open Data Barometer, 2015) (Global Open Data Index, 2017). Inherently, the area of analysis in published related work is often public health data. Still, there is promising work done in other areas of health data such as making medical images (Clark *et al.*, 2013) or drug data for pharmaceutical research (Samwald *et al.*, 2011) freely available. This represents a potential research gap since data quality has different aspects in different health disciplines such as biomedical data and neurological image data.

Evaluation frameworks often only look at a specific set of attributes such as data quality. The mapping study and following literature review found that evaluations frameworks often only look at a specific set of attributes, such as data quality or data freshness, when assessing open data. This represents a potential research gap given that the overall effectiveness or usefulness of open data depends on more than one factor. Further, e.g.: data quality can vary between health data types, therefore, it is important to look at the selected attribute such as data quality in the context of the specific health data. For example, a public health dataset is very different from a neurological dataset and needs to be evaluated specifically not generally to assess the overall usefulness of the data set.

4 Developing the Scorecard

This chapter answers the second research question: what might be an appropriate scorecard for evaluating open health data? From the synthesized clusters the authors built the first scorecard consisting of the dimensions: Data Transparency, Data Quality, Data Processability and Data Output. The names of the dimensions were chosen by the author based on what he thinks describes a dimension overall the best and where possible by literature mentioning similar names and concepts.

4.2 Data Transparency

Table 4.1 Content of the Data Transparency dimension at a glance describing the individual questions to assess the fulfilment of the dimension.

Data Transparency

- Is the data collection method and possible limitations described?
- Is the original purpose of the data collection described?
- How timely is the data (e.g.: regularly updated or a one-time release)
- Are the data characteristics, size, encoding, format and structure defined?
- Is the data license/ usage agreement defined?

Specifically, the Data Transparency dimensions' main concern is the safe use and deployment of the available data sets by the data user. While the data consumer also carries the responsibility of ensuring that the data can be safely used for the intended use, the tools and documentation need to be supplied. This is an important undertaking since open data sets are often created and maintained in an ad-hoc manner (Kienle, 2012). Kienle also mentions possible diverse characteristics of open data sets: size, encoding and structure. To pick encoding as an example, even if data is structured, the encoding could vary. E.g. in CSV files, different conventions can be used to e.g. denote filed separators, record separators, string entities or date and time (Kienle, 2012). The latter, has led to analysed errors in published datasets and acts as a good reminder of the importance of this section. A similar concept of transparency has been presented in Sayogo et al. (Sayogo, Pardo and Cook, 2014)

Generally, this dimension and questions below should help to identify what kind of information might be useful for the data consumer. While this represents an imperative in arguably every industry, the health domain carries special burden given the sensitive nature of all undertakings. A good reminder about the risks if information, such as how the data was collected, is missing can be found in the paper "Gene name errors are widespread in the scientific literature" (Ziemann *et al.*, 2016). Here the authors describe a study on 18 leading genomics journals and their supplementary files published between

2005 and 2015, which are using a spreadsheet software that contains gene names. The 18 papers included 35,175 supplementary spreadsheet files and 7467 gene lists attached to 3597 published papers. The study revealed that of the selected journals there was an error range of approximately 5%-30% spread between the selected journals. On average the error was present in 20% of the selected papers. The errors appear because of the default formatting setting that converts gene names to dates and floating-point numbers. For example, the gene symbol for Membrane Associated Ring Finger "MARCH 1" would automatically be converted to "1-MAR" or Septin 2 "SEPT2" to "2006/09/02". What is more, identifiers such as "231009E13" were being automatically converted to "2.31E+13". This represents a great challenge for the genomics community since supplementary files are a vital resource and frequently reused. Although the issue was already unveiled by Zeeberg et al. in 2004 (Zeeberg *et al.*, 2004) they seemingly remain common. This example is a good reminder how small errors can potentially lead to subsequent errors when the data is reused. Due to the nature of conversions there is always a risk of information loss that cannot be reconstructed. Therefore, detailed documentation about the released data is necessary to enable a safe deployment. Such examples support the insights extracted from the preliminary mapping study.

While dimensions described in this section are sometimes included in data quality dimension, the author made a conscious decision to create a separate dimension to highlight the importance of these aspects. Additionally, considering that data quality aspects can be highly technical, spinning out the aspects described here helps in executing work packages and making the best use of skills. The dissertation will describe the individual sections for the Data Transparency dimension in greater detail that should, among others, support the data user in clarifying if the data is fit for its intended purpose.

Ubaldi (Ubaldi, 2013) names "Primary" as one of eight open government data principles, meaning that data should be collected directly at the source with the highest level of granularity and not in aggregated or modified form. Especially if the stronger, compared to other domains, need for strong de-identification (El Emam *et al.*, 2012) is considered, datasets will appear in modified versions. In light, of this it is important to transparently describe such modifications so that the data user is aware of possible limitations. In total the following five sub-dimension were identified and a summary has been provided in table 4.1.

Is the data collection method and possible limitations described?

The example from Ziemann et al. (Ziemann *et al.*, 2016) mentioned earlier in this chapter acts as a great reminder why the data collection method should be described and attached to released datasets. If data seems corrupted, documentation about how the data was collected can support the reconstruction trail and usability of data.

Is the original purpose of the data collection described?

The original purpose of the data collection should be described to undercover potential biases or unusual trends in datasets. The described research around the thirty day-mortality-rate (Schwarze, Brasel and Mosenthal, 2014) in section 3.4.5 Open health data in the public health domain and business health data is a good example for this sub-dimension. A dataset could show a ten and thirty-day mortality rate statistic. Without further context about why this statistic was collected, the data user might make wrong conclusions.

How timely is the data (e.g.: regularly updated or a one-time release)?

This section entails similarities to what is sometimes called data freshness (Neumaier and Umbrich, 2016) in the available literature. Here, we can consider the mentioned example of transport data application from section 3.4.1 Other industries as a great way to understand the importance of this sub-dimension. For example, a developer could build an application on top of a daily released data set, only to find out that the frequent release got discontinued. While there is obviously no guarantee that the data release can or will stick to the original plan, it acts as a way of guiding continuity. This sub-dimension ties in to the other section in this chapter "*Is the data collection method and possible limitations described?*", since also data collection methods could depreciate. With transparency into these processes, the data user can make an informed decision about the deployment of the data and possible limitations.

Are the data characteristics, size, encoding, format and structure defined?

This question is related to the previous questions and in line with the overall transparency dimension of this chapter. It is vital to further break down the mentioned data characteristics about data size, encoding, format and structure to allow the data user to check e.g. compatibility with existing tools beforehand and solve errors that might occur. While this question also deals with metadata, which are described in section 4.4 Data Processablity, it was a conscious decision to include them in both sections. Firstly, because they are relevant to both sections and secondly because the technical degree of the two dimensions varies. While the metadata aspects described here can be assessed without deep technical expertise, the Data Processability dimension requires deeper domain expertise to assess the degree to which metadata is influencing the ability to process data efficiently and further domain expertise to make sense of health data specific metadata. Finally, an additional differentiator could be established when the individual parts are scored and weighted (see further research opportunities in Chapter 5 Conclusion).

Is the data license/ usage agreement defined?

While open datasets generally strive to be licence free, licences of available datasets can still vary. The specified data licence or usage agreement, is clearly highly important for the data user to understand how, for what purposes, and in which ways he can use the datasets. This is especially relevant for health related data, since there could be medical data sets that should e.g. only be released for researchers and e.g. not for commercial use.

4.3 Data Quality

Table 4.2 Content of the Data Quality dimension at a glance describing the individual questions to assess the fulfilment of the dimension.

Data Quality

- Is the data quality measurement based on a theoretical framework?
- Is the data quality framework specific for the health data type?
- Is the data complete (e.g. no years missing)?
- Is the data accurate?

Opening data without proper data quality controls may jeopardize the overall usability and utilization of open datasets (Vetrò *et al.*, 2016). Based on the number of evaluation frameworks, data quality is arguably the main focus area of open data evaluation frameworks. Generally, data quality is described as the key safeguard and improvement mechanism of data (Herzog, Scheuren and Winkler, 2007). When talking about data quality, famous research from Juran & Godfrey (Juran and Godfrey, 1999) says that data are of high quality if they are "fit for use" for the intended operation, decision making process or other roles. Given the cross-disciplinary nature of data quality, there is no universally agreed definition (Kučera, Chlapek and Nečaský, 2013). Still, three common themes often emerge: accuracy, completeness and timeliness (Scannapieco and Catarci, 2002). Timeliness has already been discussed in the previous section 4.2 Data Transparency. The author thinks that it overall fits better into concepts described there. In light of this, the Data Quality dimension will, amongst others, describe accuracy and completeness.

Since the dissertation established that many evaluation frameworks do not distinguish between different health data types, the following subsections will include examples to highlight such differences. Ideally, there would be a specialized scorecard for each health data type but this would exceed the scope of this work dramatically. Still, this could be an area of analysis for further research. The following sub-dimension describe the Data

Quality dimension in greater detail. In total four sub-dimension were identified and a summary has been provided in table 4.2.

Is the data quality measurement based on a theoretical framework?

There is vast research available around data quality and a number of widely accepted theoretical frameworks. Some of these have been described in the section 3.6 Open data evaluations and guidelines. When describing data quality in the context of open data, researchers often use an underlying theoretical framework as a baseline e.g: (Vetrò *et al.*, 2016). While recommending a theoretical framework for specific health data types would exceed the scope of this work, the author wants to stress the importance of theoretical data quality frameworks at this point. This also offers an opportunity for further research.

Is the data quality framework specific for the health data type?

As established earlier, little research focuses on the specifics types of health data. Given the big differences of health datasets such as hospital performance data vs. genomics data, choosing a quality model or framework that has been tested and previously used for the specific health data type, might prove to be of high effectiveness.

Is the data complete (e.g. no years missing)?

Data completeness is immensely important, particularly in the health domain. For example, considering a data set that shows drug effectiveness over a ten-year timeframe but is missing years in between. Data completeness has also been called "the Achilles heel of drug-target networks" (Mestres *et al.*, 2008). While data completeness is related to concepts described in section 4.2 Data Transparency, it is closer associated with this section about data quality because data completeness is about the data itself and not of describing character.

Is the data accurate?

Data accuracy here is the understanding that the data is free of errors (Vetrò *et al.*, 2016). While the absence of errors in data is vital in any environment, it is especially succinct in the health domain. This is due to the health domain being described as arguably the most technologically intense (Kaushal *et al.*, 2005) and data-richest industry (Smith, 2006).

The data accuracy question is also a great example to understand the interconnectedness of the dimensions of the scorecard presented in this dissertation. With the concepts presented in the Data Transparency dimension, data consumers might be able to undercover errors earlier or fix data to reduce the number of errors.

4.4 Data Processability

Table 4.3 Content of the Data Quality dimension at a glance describing the individual questions to assess the fulfilment of the dimension.

Data Processability

- State of machine readability based on Walker (Walker et al. 2005)?
- How open and linked is the data based on Berners-Lee (Tim Berners-Lee, 2009)?
- Is general metadata and health specific metadata defined for the dataset?
- Is the data user required to use non-standard software?
- Are APIs available?

The Data Processability dimension describes the accessibility aspect of data, accessibility in terms of usability and usability in the sense of how easily can the data be processed or to what degree is the data processing process supported. The Data Processability of open data sets is vital to the definition of "open" and could also conveys concepts of data interoperability as well as automation (Kaiser, Klier and Heinrich, 2007). While this is also a concern of other sections of the presented scorecard, the Data Processability dimension further represents how reusable the available data is. It is closely tied to the Data Quality dimension and points presented here are in some evaluation studies subsumed under data quality. To highlight the importance and further emphasises the various aspects of data processing, the author has decided to attribute its own section to the topic. The importance of data processing in terms of data openness is well described by Attard et al. (Attard *et al.*, 2015) and the example of available archives of legal documents and might not even be digitalized.. The effort to process these manually or without freely available tools would be an immense effort (Attard *et al.*, 2015). On the other hand, if the data includes concepts of Data Processability here, insights can be extracted in a fraction of the time.

The following sub-dimension will describe the individual sections for the Data Processability dimension in greater detail. In total the following five sub-dimension were identified and a summary has been provided in table 4.3.

State of machine readability based on Walker (Walker et al., 2005)?

Walker et al., defines four levels of automation when data is shared and differentiates between the amount of human involvement required, the level of standardization and the sophistication of IT. What is more, Walker's et al. research is already applied to the health domain and, therefore, of high relevance of this dissertation.

Table 3.2: Walker's (Walker *et al.*, 2005) levels of data automation applied to health data, table organization: (Sprivulis *et al.*, 2007)

Level	**Data**	**Definition**	**Example**
1	Non-electronic data	Min. use of IT to share information.	Mail, telephone.
2	Machine transportable data	Transmission of non-standardized information via basic IT. The information within the document cannot be electronically manipulated.	Fax, PDF.
3	Machine organizable data	Transmission of structured messages containing non-standardized data; requires interfaces to translate data from the sending organization's vocabulary to the receiving organization's vocabulary.	E-mail of free text, exchange of files in incompatible/ proprietary file formats.

4	Machine interpretable data	Transmission of structured messages containing standardized and coded data; systems exchange information using the same formats and vocabularies.	Automated exchange of coded results from external laboratories into an electronic medical record, automated exchange of the patients "active problem" lists between providers.

How open and linked is the data based on Berners-Lee (Tim Berners-Lee, 2009)?

Tim Berners-Lee (Tim Berners-Lee, 2009) presents a 5 star framework to assess the openness of data:

1) "On the web, open licence". The data is available on the web with an open licence and in any format
2) "Machine-readable data". The data is available as machine-readable structured data, for example as a spreadsheet table instead of a scanned document
3) "Non-proprietary format". The data is available in a machine-readable structured data format and additionally, as a non-proprietary format. For example, CSV.
4) "RDF standards". The data is available in a machine-readable structured data format, in a non-proprietary format and, additionally, uses open standards from W3C. For example, RDF and SPARQL standards.
5) "Linked RDF". The data is available in a machine-readable structured data format, in a non-proprietary format, uses open standards from W3C and links data to other data to provide a greater sense of context.

Is general metadata and health specific metadata defined for the dataset?

Metadata is defined as "data that provides information about other data" (merriam-webster.com, 2017) and an important contributor to the effectiveness of open health data. In the publication (Poldrack and Gorgolewski, 2015) the authors highlight that a significant amount of metadata is required to make sense out of the raw task fMRI data. Metadata in this case includes description of task events and their timing. This example shows how important proper metadata hygiene is especially in the health domain. As the authors noted (Poldrack and Gorgolewski, 2015) that the fMRI data wouldn't even be useful without the proper metadata. Breaking down and defining specific metadata for all health data types) would exceed the scope of this dissertation. However, defining health data specific metadata in the context of open health data represents an interesting question and opportunity for further research.

Is the data user required to use non-standard software?

How the data was recorded is highly relevant to the data user. Especially in the health domain, where interoperability is an often discussed issue, the software that was used could have a big impact on the processability of the data and the usability of data. In the health domain, an industry with highly specialized machinery, data could have been recorded in a discontinued format due to older medical equipment that is only used in house.

Are APIs available?

Application Programming Interfaces or APIs, among other tasks, describe data structures so data can be passed on from one program to another (Kernighan, 2017). APIs allow a greater sense of automation and, therefore, can vastly contribute to better data processing.

4.5 Data Output

Table 4.4 Content of the Data Quality dimension at a glance describing the individual questions to assess the fulfilment of the dimension.

Data Output

- Are the activities resulting from the open data being tracked (e.g.: research, reports, built apps)?
- Are there initiatives planned to attract new data consumers (e.g.: developers)?
- Are community activities in place (e.g. hackathons)?
- Are there examples about how data can be used?

The Data Output dimension described here deals with the question what kind of impact the released data sets hare having and how they were used. This is not only an important undertaking to estimate the return of investment but also to gather feedback from the data consumers. Such feedback can be highly valuable to reiterate of the data offerings and further expand in other areas to amplify the impact of open data initiatives. Related concepts were described by Martin et al. (Martin, Shah and Birkhead, 2017).

The following subsection describes the individual sections for the data output section in greater detail. In total the following four sub-dimensions were identified and a summary has been provided in table 4.4.

Are the activities resulting from the open data being tracked (e.g.: research, reports, built apps)?

Open data initiatives do not come without resource investments and it is important to assess the impact of the released datasets to evaluate further investments.

A great example is the Cancer Imaging Archive that lists all the publications that resulted from the platform (Clark *et al.*, 2013) (Cancer Imaging Archive, 2017). On the platform, 23 collections containing 3.3 million images were made available in its first year of operation. Out of these collections in total more than 400 publications originated that are listed on the platforms website.

Are there initiatives planned to attract new data consumers (eg.: developers)?

Great open data initiatives can go unnoticed if they are not properly advertised or easily accessible. Further, the data might be useful to a specific type of user that is unaware of this offering. Planning such initiatives shall prevent missed opportunities.

Are community activities in place (e.g. hackathons)?

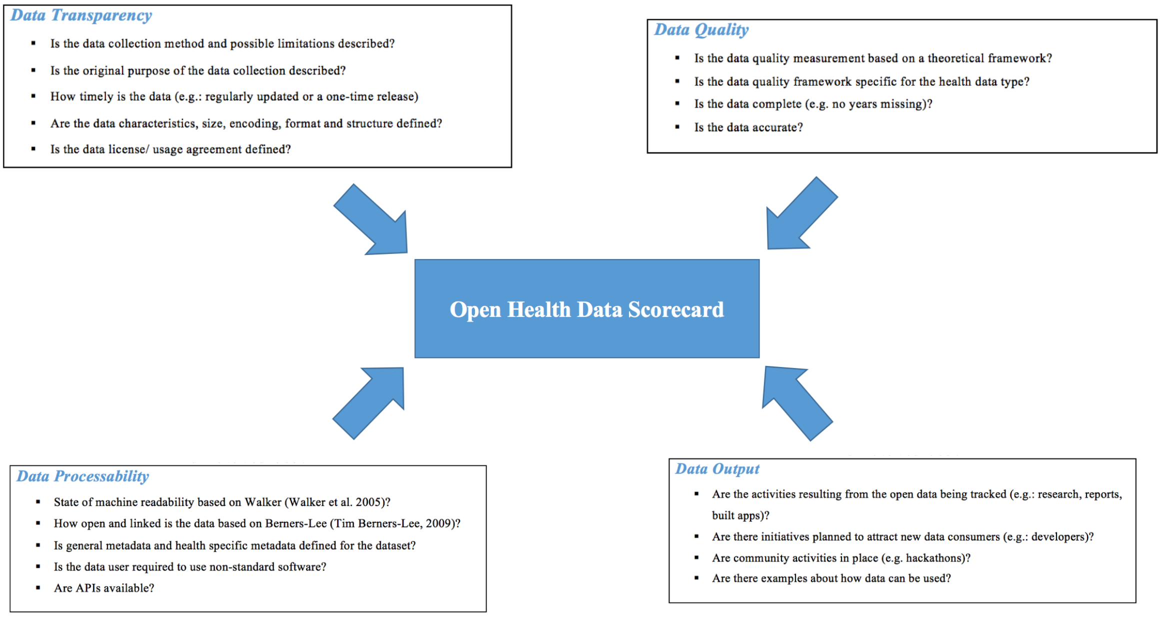
Hackathons or workshops are common in the open data community and can foster collaboration and innovation. They create a greater sense of community and are also a great medium to gather feedback and ideas for improvement of the data offerings.

Are there examples of how data can be used?

Showcasing exemplary applications how the data was used can help to spark new ideas for potential data users. The previously mentioned example of the Cancer Imaging Archive (Clark *et al.*, 2013) (Cancer Imaging Archive, 2017) lists all publications that resulted from the released data sets. This list also acts as a great resource for potential data users to brainstorm new ideas and build up on existing ideas.

4.6 Forming an open health data scorecard

This section provides an overview of the organized scorecard in table 4.6, including all dimensions. More advanced visualizations of the scorecard or building an application out of the scorecard, could be subject of analysis for further research. Figure 4.1 shows the different dimensions that were included in the scorecard.



Data Transparency

- Is the data collection method and possible limitations described?
- Is the original purpose of the data collection described?
- How timely is the data (e.g.: regularly updated or a one-time release)
- Are the data characteristics, size, encoding, format and structure defined?
- Is the data license/ usage agreement defined?

Data Quality

- Is the data quality measurement based on a theoretical framework?
- Is the data quality framework specific for the health data type?
- Is the data complete (e.g. no years missing)?
- Is the data accurate?

Open Health Data Scorecard

Data Processability

- State of machine readability based on Walker (Walker et al. 2005)?
- How open and linked is the data based on Berners-Lee (Tim Berners-Lee, 2009)?
- Is general metadata and health specific metadata defined for the dataset?
- Is the data user required to use non-standard software?
- Are APIs available?

Data Output

- Are the activities resulting from the open data being tracked (e.g.: research, reports, built apps)?
- Are there initiatives planned to attract new data consumers (e.g.: developers)?
- Are community activities in place (e.g. hackathons)?
- Are there examples about how data can be used?

Figure 4.1 Four dimensions of the open health data scorecard that were defined in chapter 4.

Table 4.6 Compiled open health data scorecard showing the four dimensions defined in chapter 4.

Open Health Data Scorecard	**Score** **(e.g.: 1-10)**
Data Transparency	
○ Is the data collection method and possible limitations described?	
○ Is the original purpose of the data collection described?	
○ How timely is the data (e.g.: regularly updated or a one-time release)	
○ Are the data characteristics, size, encoding, format and structure defined?	
○ Is the data license/ usage agreement defined?	
Data Quality	
○ Is the data quality measurement based on a theoretical framework?	
○ Is the data quality framework specific for the health data type?	
○ Is the data complete (e.g. no years missing)?	
○ Is the data accurate?	
Data Processability	
○ State of machine readability based on Walker (Walker et al. 2005)?	
○ How open and linked is the data based on Berners-Lee (Tim Berners-Lee, 2009)?	
○ Is general metadata and health specific metadata defined for the dataset?	
○ Is the data user required to use non-standard software?	
○ Are APIs available?	
Data Output	
○ Are the activities resulting from the open data being tracked (e.g.: research, reports, built apps)?	
○ Are there initiatives planned to attract new data consumers (e.g.: developers)?	
○ Are community activities in place (e.g. hackathons)?	
○ Are there examples about how data can be used?	

4.7 Applying and scoring the open health data scorecard

Applying and scoring the scorecard to real-life cases was not part of the scope of this dissertation but offers various opportunities for further research. The scorecard could evaluate open data initiatives around the world and be further refined with the lessons learned from applying the scorecard. How to appropriately score the individual dimension of the scorecard (e.g. should data quality carry a bigger weight than data processing?) would require more research and following iterations. Therefore, the dissertation will only briefly touch on existing research that deals with scoring evaluations.

4.7.1 Analytic hierarchy process

Research by Kubler et al. (Kubler *et al.*, 2016) applied an analytic hierarchy process (AHP) to compare 146 open data portals. The AHP was originally presented in (Saaty, 2001) and is a useful tool to address multi-criteria decision making problem. The process organizes critical aspects of a problem similar to how the brain would structure knowledge. The method followed a three step process:

1) Pairwise comparison based preference measurement
2) Pairwise comparisons as ratio measurement
3) TOPSIS-based (Technique for Order Preference by Similarity to Ideal Solution) alternative ranking.

These steps will not be described in further detail as it would exceed the scope of this dissertation but should leave the reader with a sense of possible scoring techniques.

The dimensions were exemplary scored and depicted in a radar chart, which will be described in the following paragraph.

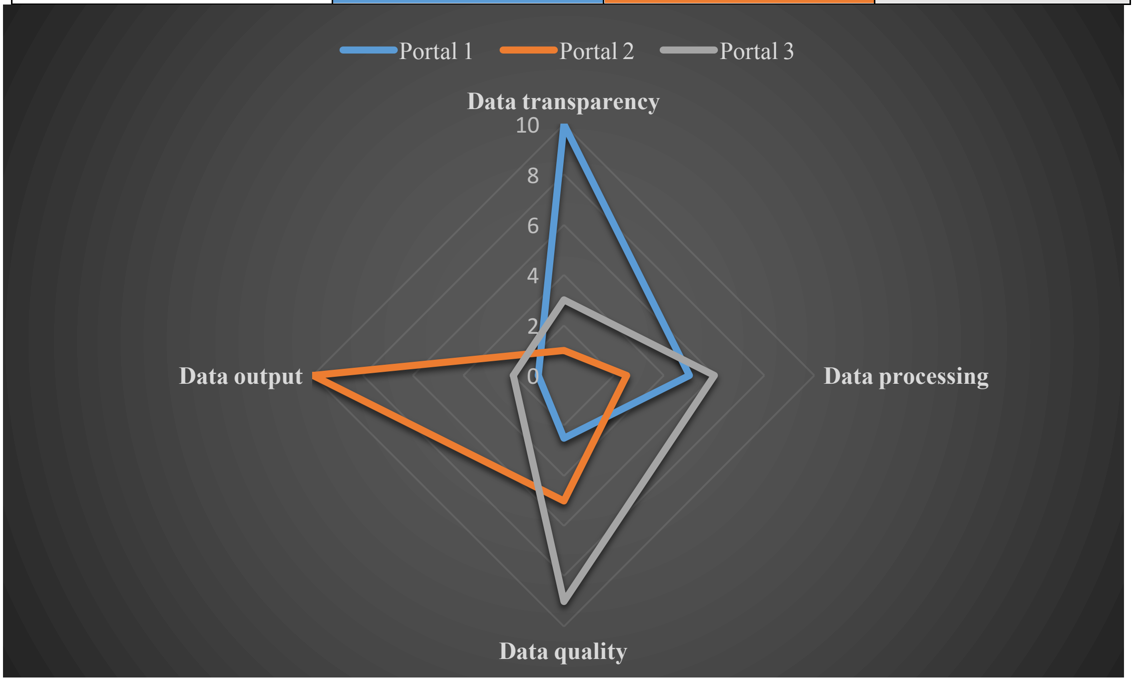
4.7.2 Displaying the results with a radar chart

Given that the research (Kubler et al. 2016) also deals with multiple dimensions to assess open data aspects, the radar chart is well suited for the dimensions presented in this dissertation.

The dissertation assumes and example where three portals for open health data were assessed used the scorecard described in chapter 4. The hypothetical example evaluation could be displayed like the following illustration.

Table 4.7: Shows an example of how the open health data scorecard could be applied.

	Portal 1	**Portal 2**	**Portal 3**
Data transparency	10	1	3
Data processing	5	2.5	6
Data quality	2.5	5	9
Data output	1	10	2



Going into further detail would exceed the scope of this dissertation but additional work on applying and visualizing the scorecard offers vast research opportunities.

5 Conclusion

The author of this dissertation set out to capture the status quo of open health data in a way that would allow to create some sort of baseline. While a baseline can have several meanings, the baseline created here shall be understood as a starting point to understand the open health data domain and provide an overview of the research field. It is a high level undertaking to understand the current state of open health data or in simpler words "what do we know about open health data and how does it work". In order to answer this question, the author conducted a preliminary mapping study.

The preliminary mapping study resulted in three clusters: "Open data applications, platforms, portals and initiatives", "Open data enablers" and "Open data evaluations and guidelines". The first cluster "Open data applications, platforms, portals and initiatives" describes open data activities that aim to solve a problem or generate value. The second cluster "Open data enablers" describes the underlying technical mechanisms that make open data work. The name has been chosen broadly to allow for further research that also includes socioeconomic aspects such as economic or legal parameters. The third cluster "Open data evaluations and guidelines" depicts publications that evaluate open data aspects such as open data quality. These clusters have been described throughout chapter three. They represent a rich resource of knowledge and should allow any reader a quick start into the open data domain.

Through mapping these clusters the dissertation found that there is little research done for open health data specifically, but various research done for open data generally. Besides publications discussing open health data applications, no publications could be identified that look at open health data specific enablers (e.g. open health data specific metadata) or open health data evaluations specifically. Therefore, the dissertation contributes to the research field of open health data and creates vast research opportunity. For example, now that preliminary clusters have been established, further research could be systematically conducted for each of the clusters (E.g. a systematic mapping study on open health data evaluations).

By capturing the state of the art and establishing a baseline for open health data, the author unified a rich collection of resources, which provide interesting insights:

- most research focuses on open data generally rather than open health data specifically
- current frameworks do not focus on open health data specifically but open data generally
- the literature that deals with open health data specifically often only considers public health data
- evaluation frameworks often only look at a specific set of attributes such as data quality

In light of these, the second research question was built: “What might be an appropriate scorecard for evaluating open health data?”. Based on the resources from the preliminary mapping study, the author compiled a scorecard to answer the second research questions. Through this process four dimensions, to evaluate open health data specifically, were crafted:

- **Data transparency**, describing the safe use and deployment of the available data sets by the data user.
- **Data quality**, describing data quality in the context of open health data
- **Data processing**, describing how easy the data can be processed or to what degree the data processing process is supported.
- **Data output**, describing the potential impact of released data sets and how they were used.

Since there is very little research that concentrates on evaluating open health data, benchmarking the outcome is only possible in a very limited way. Only further research and the research community’s discourse can holistically evaluate the outcome. What is more, weighting the individual components of the scorecard (e.g. is data quality or data openness more important) exceeded the scope of this research and could become the focus of further studies. What is more, the dissertation has provided further touchpoints for further research throughout the dissertation and summarized them in Table 5.1 below.

Table 5.1: Summarised opportunities for further research based on this dissertation.

Expanding the scope	Data security, privacy and de-identification, were chosen to not be included in this dissertation as they would have exceeded the scope of this dissertation. They could become subject for further analysis and be included in chapter 3.5 Open data enablers.
Further systematic mapping studies	Now that the three clusters have been mapped in chapter 3 Preliminary mapping study, further research could do a full review of all the publication for one cluster. E.g.: a systematic mapping study on open data evaluations and guidelines.
Creating health data specific scorecards	The scorecards could be adapted to each health data type. E.g.: a scorecard for public health data that highlights specific aspects such as metadata specific to the health data type.
Visualization	More advanced visualizations of the scorecard or building an application out of the scorecard, could be subject of analysis for further research.
Scoring & weighting	How to appropriately score the individual dimension of the scorecard (e.g. should data quality carry a bigger weight than data processing?) would require more research and following iterations.

Applying the scorecard	Applying and scoring the scorecard to real-life cases was not part of the scope of this dissertation but offers various opportunities for further research.

Creating an open health data scorecard cannot be done by a single person. It requires not only open data experts but also experts in the different health data domains to address the specific and often complex requirements of distinctive health data types. Still, the author is confident that this dissertation presents a great starting point for further developments in the open health data space and contributes to the expansion of the research field. The author is excited about the opportunities open data presents in the health domain and hopes that the field will further grow to touch and improve even more lives.

Bibliography

Agostinelli, Anderson and Lee (2013) 'Adaptive multi-column deep neural networks with application to robust image denoising', *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Curran Associates Inc., pp. 1493–1501. Available at: http://dl.acm.org/citation.cfm?id=2999611.2999778 (Accessed: 4 March 2017).

Alexopoulos, Zuiderwijk, Charapabidis, Loukis and Janssen (2014) 'Designing a Second Generation of Open Data Platforms: Integrating Open Data and Social Media', pp. 230–241. doi: 10.1007/978-3-662-44426-9_19.

Attard, Orlandi, Scerri and Auer (2015) 'A systematic review of open government data initiatives', *Government Information Quarterly*, 32(4), pp. 399–418. doi: 10.1016/j.giq.2015.07.006.

Behkamal, Kahani, Bagheri and Jeremic (2014) 'A Metrics-Driven Approach for Quality Assessment of Linked Open Data', *Journal of theoretical and applied electronic commerce research*, 9(2), pp. 11–12. doi: 10.4067/S0718-18762014000200006.

Belleau, Nolin, Tourigny, Rigault and Morissette (2008) 'Bio2RDF: Towards a mashup to build bioinformatics knowledge systems', *Journal of Biomedical Informatics*, 41(5), pp. 706–716. doi: 10.1016/j.jbi.2008.03.004.

Bento, Gaulton, Hersey, Bellis, Chambers, Davies, Krüger, Light, Mak, McGlinchey, Nowotka, Papadatos, Santos and Overington (2014) 'The ChEMBL bioactivity database: an update', *Nucleic Acids Research*, 42(D1), pp. D1083–D1090. doi: 10.1093/nar/gkt1031.

Bizer, Heath and Berners-Lee (2009) 'Linked Data - The Story So Far', *International Journal on Semantic Web and Information Systems*, 5(3), pp. 1–22

Boyce, Horn, Hassanzadeh, Waard, Schneider, Luciano, Rastegar-Mojarad and Liakata (2013) 'Dynamic enhancement of drug product labels to support drug safety, efficacy, and effectiveness.', *Journal of biomedical semantics*, 4(1), p. 5. doi: 10.1186/2041-1480-4-5.

Bray, Ferlay, Laversanne, Brewster, Gombe Mbalawa, Kohler, Piñeros, Steliarova-Foucher, Swaminathan, Antoni, Soerjomataram and Forman (2015)

‘Cancer Incidence in Five Continents: Inclusion criteria, highlights from Volume X and the global status of cancer registration’, *International Journal of Cancer*, 137(9), pp. 2060–2071. doi: 10.1002/ijc.29670.

British Columbia Medical Association (2006) *Waiting Too Long: Reducing and Better Managing Wait Times in BC*, *A Policy Paper by the BCMA’s Council on Health Economics & Policy*

Budgen, Turner, Brereton and Kitchenham (2008) ‘Using Mapping Studies in Software Engineering’, *Proceedings of the 20th annual meeting of the Psychology of Programming Interest Group*, pp. 195–204

Burégio, de Lemos Meira and de Almeida (2010) ‘Characterizing Dynamic Software Product Lines - A Preliminary Mapping Study.’, *SPLC Workshops*, pp. 53–60. Available at: http://dblp.uni-trier.de/db/conf/splc/splc2010w.html#BuregioMA10 (Accessed: 3 June 2017).

Calero, Caro and Piattini (2008) ‘An Applicable Data Quality Model for Web Portal Data Consumers’, *World Wide Web*. Springer US, 11(4), pp. 465–484. doi: 10.1007/s11280-008-0048-y.

Cancer Imaging Archive (2017) *Publications*. Available at: https://wiki.cancerimagingarchive.net/display/Public/Publications.

Chen, Dong, Jiao, Wang, Zhu, Ding and Wild (2010) ‘Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data’, *BMC Bioinformatics*, 11(1), p. 255. doi: 10.1186/1471-2105-11-255.

Clark, Vendt, Smith, Freymann, Kirby, Koppel, Moore, Phillips, Maffitt, Pringle, Tarbox and Prior (2013) ‘The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository.’, *Journal of digital imaging*. Springer, 26(6), pp. 1045–57. doi: 10.1007/s10278-013-9622-7.

Codella, Connell, Pankanti, Merler and Smith (2014) ‘Automated Medical Image Modality Recognition by Fusion of Visual and Text Information’, in, pp. 487–495. doi: 10.1007/978-3-319-10470-6_61.

dailymed.nlm.nih.gov (2017) *DailyMed*. Available at: https://dailymed.nlm.nih.gov/dailymed/index.cfm (Accessed: 31 March 2017).

Dajer (2012) *Panel gives front-runner status to MIT plan for city schools - The Boston Globe*, *The Boston Globe*. Available at:

https://www.bostonglobe.com/metro/2012/10/27/boston-panel-winnow-options-for-school-assignment-changes-new-mit-proposal-leaps-top/jy8yIsqjLM23w9ECHXpknM/story.html (Accessed: 12 February 2017).

dbpedia.org (2017) *About | DBpedia*. Available at: http://wiki.dbpedia.org/about (Accessed: 1 April 2017).

eHealth Ireland (2017) *Operating Model Guiding Principles - eHealth Ireland, http://www.ehealthireland.ie/*. Available at: http://www.ehealthireland.ie/Knowledge-Information-Plan/Knowledge-Information-Operating-Model/Operating-Model-Guiding-Principles/ (Accessed: 11 March 2017).

El Emam, Arbuckle, Koru, Eze, Gaudette, Neri, Rose, Howard and Gluck (2012) 'De-identification methods for open health data: the case of the Heritage Health Prize claims dataset.', *Journal of medical Internet research*. Journal of Medical Internet Research, 14(1), p. e33. doi: 10.2196/jmir.2001.

Fang, Huang, Chen and Juan (2008) 'TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining', *BMC Complementary and Alternative Medicine*, 8(1), p. 58. doi: 10.1186/1472-6882-8-58.

Forbes, Beare, Gunasekaran, Leung, Bindal, Boutselakis, Ding, Bamford, Cole, Ward, Kok, Jia, De, Teague, Stratton, McDermott and Campbell (2015) 'COSMIC: exploring the world's knowledge of somatic mutations in human cancer', *Nucleic Acids Research*, 43(D1), pp. D805–D811. doi: 10.1093/nar/gku1075.

Fulda and Lykens (2006) 'Ethical issues in predictive genetic testing: a public health perspective.', *Journal of medical ethics*. BMJ Group, 32(3), pp. 143–7. doi: 10.1136/jme.2004.010272.

Gaulton, Bellis, Bento, Chambers, Davies, Hersey, Light, McGlinchey, Michalovich, Al-Lazikani and Overington (2012) 'ChEMBL: a large-scale bioactivity database for drug discovery', *Nucleic Acids Research*, 40(D1), pp. D1100–D1107. doi: 10.1093/nar/gkr777.

Gaulton and Overington (2010) 'Role of open chemical data in aiding drug discovery and design', *Future Medicinal Chemistry*. Future Science Ltd London, UK , 2(6), pp. 903–907. doi: 10.4155/fmc.10.191.

Global Open Data Index (2017) *Methodology | Global Open Data Index by Open Knowledge*. Available at: http://index.okfn.org/methodology/ (Accessed: 26 March

2017).

Goh, Cusick, Valle, Childs, Vidal and Barabási (2007) ‘The human disease network.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 104(21), pp. 8685–90. doi: 10.1073/pnas.0701361104.

Haklay and Weber (2008) ‘OpenStreetMap: User-Generated Street Maps’, *IEEE Pervasive Computing*, 7(4), pp. 12–18. doi: 10.1109/MPRV.2008.80.

Hassanzadeh, Kementsietsidis, Lim, Miller and Wang (2009) ‘LinkedCT: A Linked Data Space for Clinical Trials’. Available at: http://arxiv.org/abs/0908.0567 (Accessed: 31 March 2017).

Health Service Executive (2017) *Stroke - HSE.ie*. Available at: http://www.hse.ie/eng/health/az/S/Stroke/ (Accessed: 4 February 2017).

Herland, Khoshgoftaar and Wald (2014) ‘A review of data mining using big data in health informatics’, *Journal Of Big Data*. Springer International Publishing, 1(1), p. 2. doi: 10.1186/2196-1115-1-2.

Herzog, Scheuren and Winkler (2007) ‘What is Data Quality and Why Should We Care’, in *Data Quality and Record Linkage techniques*. New York, NY: Springer New York, pp. 7–15. doi: 10.1007/0-387-69505-2_2.

Holzinger (2016) ‘Machine Learning for Health Informatics’, in *Machine Learning for Health Informatics*. doi: 10.1007/978-3-319-50478-0_1.

Holzinger, Dehmer and Jurisica (2014) ‘Knowledge Discovery and interactive Data Mining in Bioinformatics--State-of-the-Art, future challenges and research directions.’, *BMC bioinformatics*, 15 Suppl 6, p. I1. doi: 10.1186/1471-2105-15-S6-I1.

International Organization for Standardization (2008) *ISO/IEC 25012:2008 - Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model*, *iso.org*. Available at: https://www.iso.org/standard/35736.html (Accessed: 21 March 2017).

Janssen (2011) ‘The influence of the PSI directive on open government data: An overview of recent developments’, *Government Information Quarterly*, 28(4), pp. 446–456. doi: 10.1016/j.giq.2011.01.004.

Jenkins, Anderson, Vance, Kirwan and Eargle (2016) ‘More Harm Than Good? How Messages That Interrupt Can Make Us Vulnerable’, *Information Systems*

Research, 27(4), pp. 880–896. Available at: http://pubsonline.informs.org (Accessed: 4 February 2017).

Jessica Nutik Zitter (2017) *Misleading Metrics*, *http://healthaffairs.org/blog*. Available at: http://healthaffairs.org/blog/2017/02/21/misleading-metrics/.

John Nosta (2016) *HIMSS16: Thought leaders preview this year's event*. Available at: http://www.philips.com/a-w/innovationmatters/blog/himss16-thought-leaders-preview-this-years-event.html (Accessed: 24 April 2017).

Jordan and Mitchell (2015) 'Machine learning: Trends, perspectives, and prospects', *Science*, 349(6245), pp. 255–260. doi: 10.1126/science.aaa8415.

Juran and Godfrey (1999) 'Quality handbook', *Republished McGraw-Hill*

Kahn and Hederman (2012) 'A Universal Clinical Decision Support System using semantic web services', in *ESWC 2012 Semantic Interoperability in Medical Informatics WSop*. Available at: http://www.tara.tcd.ie/handle/2262/64246.

Kaiser, Klier and Heinrich (2007) 'How to measure data quality?-a metric-based approach', *ICIS 2007 Proceedings*

Kalpathy-Cramer, Zhao, Goldgof, Gu, Wang, Yang, Tan, Gillies and Napel (2016) 'A Comparison of Lung Nodule Segmentation Algorithms: Methods and Results from a Multi-institutional Study', *Journal of Digital Imaging*. Springer International Publishing, 29(4), pp. 476–487. doi: 10.1007/s10278-016-9859-z.

Kamateri, Kalampokis, Tambouris and Tarabanis (2014) 'The linked medical data access control framework', *Journal of Biomedical Informatics*, 50, pp. 213–225. doi: 10.1016/j.jbi.2014.03.002.

Kaufman (2012) *Getting Started with Open Data Getting Started with Open Data*. Available at: http://wagner.nyu.edu/rudincenter/research/ (Accessed: 12 February 2017).

Kaushal, Blumenthal, Poon, Jha, Franz, Middleton, Glaser, Kuperman, Christino, Fernandopulle, Newhouse, Bates and The Cost of National Health Informatoin Network Working Group (2005) 'The Costs of a National Health Information Network', *Annals of Internal Medicine*, 143(3), p. 165. doi: 10.7326/0003-4819-143-3-200508020-00002.

Keegan (2010) *Meet the Wikipedia of the mapping world | Victor Keegan | Technology | The Guardian*, *theguardian.com*. Available at:

https://www.theguardian.com/technology/2010/feb/04/mapping-open-source-victor-keegan (Accessed: 19 February 2017).

Kernighan (2017) *Understanding the digital world : what you need to know about computers, the Internet, privacy, and security*

Ketelaar, Faber, Flottorp, Rygh, Deane and Eccles (2011) 'Public release of performance data in changing the behaviour of healthcare consumers, professionals or organisations', in Ketelaar, N. A. (ed.) *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd, p. CD004538. doi: 10.1002/14651858.CD004538.pub2.

Kienle (2012) 'Open Data: Reverse Engineering and Maintenance Perspective'. Available at: http://arxiv.org/abs/1202.1656 (Accessed: 9 January 2017).

Kitchenham (2010) 'What's up with software metrics?–A preliminary mapping study', *Journal of systems and software*. Available at: http://www.sciencedirect.com/science/article/pii/S0164121209001599 (Accessed: 30 May 2017).

Kitchenham, Budgen and Pearl Brereton (2011) 'Using mapping studies as the basis for further research – A participant-observer case study', *Information and Software Technology*, 53(6), pp. 638–651. doi: 10.1016/j.infsof.2010.12.011.

Kitchenham and Charters (2007) *Guidelines for performing Systematic Literature Reviews in Software Engineering*, *EBSE Technical Report EBSE-2007-01*

Kitchenham, Dyba and Jorgensen (2004) 'Evidence-Based Software Engineering', *Proceedings of the 26th International Conference on Software Engineering*, pp. 273–281. Available at: http://dl.acm.org/citation.cfm?id=998675.999432 (Accessed: 30 May 2017).

Kubler, Robert, Le Traon, Umbrich and Neumaier (2016) 'Open Data Portal Quality Comparison using AHP', in *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research - dg.o '16*. New York, New York, USA: ACM Press, pp. 397–407. doi: 10.1145/2912160.2912167.

Kučera, Chlapek and Nečaský (2013) 'Open Government Data Catalogs: Current Approaches and Quality Perspective', in. Springer Berlin Heidelberg, pp. 152–166. doi: 10.1007/978-3-642-40160-2_13.

Kuhn, Campillos, Letunic, Jensen and Bork (2010) 'A side effect resource to capture phenotypic effects of drugs', *Molecular Systems Biology*, 6, p. 343. doi: 10.1038/msb.2009.98.

Kuhn, von Mering, Campillos, Jensen and Bork (2007) 'STITCH: interaction networks of chemicals and proteins', *Nucleic Acids Research*, 36(Database), pp. D684–D688. doi: 10.1093/nar/gkm795.

LeCun, Bengio and Hinton (2015) 'Deep learning', *Nature*. Nature Research, 521(7553), pp. 436–444. doi: 10.1038/nature14539.

Li, Yu, Ren and Lou (2010) 'Securing Personal Health Records in Cloud Computing: Patient-Centric and Fine-Grained Data Access Control in Multi-owner Settings', in. Springer, Berlin, Heidelberg, pp. 89–106. doi: 10.1007/978-3-642-16161-2_6.

Liu, Wei Ma, Moore, Ganesan and Nelson (2005) 'RxNorm: prescription for electronic drug information exchange', *IT Professional*, 7(5), pp. 17–23. doi: 10.1109/MITP.2005.122.

Máchová and Lněnička (2017) 'Evaluating the Quality of Open Data Portals on the National Level', (1), pp. 21–41. doi: 10.4067/S0718-18762017000100003.

Manyika, Chui, Groves, Farrell, Van Kuiken and Doshi (2013) 'Open Data: Unlocking Innovation and Performance with Liquid Information', *McKinsey Global Institute*, (October), p. 24. Available at: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information (Accessed: 19 February 2017).

Marco-Ruiz, Pedrinaci, Maldonado, Panziera, Chen and Bellika (2016) 'Publication, discovery and interoperability of Clinical Decision Support Systems: A Linked Data approach', *Journal of Biomedical Informatics*, 62, pp. 243–264. doi: 10.1016/j.jbi.2016.07.011.

Martin, Foulonneau and Turki (2013) '1-5 Stars: Metadata on the Openness Level of Open Data Sets in Europe', in *Communications in Computer and Information Science*, pp. 234–245. doi: 10.1007/978-3-319-03437-9_24.

Martin and Kaltenbock (2011) 'Open Knowledge Conference » The Open Government Data Stakeholder Survey', in *CEUR Workshop Proceedings*

Martin, Shah and Birkhead (2017) 'Unlocking the Power of Open Health Data: A Checklist to Improve Value and Promote Use.', *Journal of Public Health Management and Practice*, p. 1. doi: 10.1097/PHH.0000000000000561.

Merriam-Webster (2017) *Baseline | Definition of Baseline by Merriam-Webster*. Available at: https://www.merriam-webster.com/dictionary/baseline (Accessed: 4 June 2017).

merriam-webster.com (2017) *Metadata | Definition of Metadata by Merriam-Webster, merriam-webster.com*. Available at: https://www.merriam-webster.com/dictionary/metadata (Accessed: 30 April 2017).

Mestres, Gregori-Puigjané, Valverde and Solé (2008) 'Data completeness—the Achilles heel of drug-target networks', *Nature Biotechnology*. Nature Publishing Group, 26(9), pp. 983–984. doi: 10.1038/nbt0908-983.

Mildenberger, Eichelberg and Martin (2002) 'Introduction to the DICOM standard', *European Radiology*. Springer-Verlag, 12(4), pp. 920–927. doi: 10.1007/s003300101100.

Moraga, Moraga, Calero and Caro (2009) 'SQuaRE-Aligned Data Quality Model for Web Portals', in *2009 Ninth International Conference on Quality Software*, pp. 117–122. doi: 10.1109/QSIC.2009.23.

Negru and Buraga (2013) 'An Educational Tool for an Interactive Faceted Exploration of DBpedia Life Sciences Data', in. Springer, Berlin, Heidelberg, pp. 503–506. doi: 10.1007/978-3-642-41154-0_39.

Neumaier and Umbrich (2016) 'Measures for Assessing the Data Freshness in Open Data Portals', in *2016 2nd International Conference on Open and Big Data (OBD)*. IEEE, pp. 17–24. doi: 10.1109/OBD.2016.10.

Neumaier, Umbrich and Polleres (2016) 'Automated Quality Assessment of Metadata across Open Data Portals', *Journal of Data and Information Quality*, 8(1), pp. 1–29. doi: 10.1145/2964909.

Ohemeng and Ofosu-Adarkwa (2015) 'One way traffic: The open data initiative project and the need for an effective demand side initiative in Ghana', *Government Information Quarterly*, 32(4), pp. 419–428. doi: 10.1016/j.giq.2015.07.005.

Open Data Barometer (2015) *ODB Methodology - v1.0 28th April 2015*. Available at: http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-Methodology.pdf (Accessed: 26 March 2017).

Open Government Working Group Meeting in Sebastopol (2007) *The Open Government Data Principles Government*. Available at:

https://public.resource.org/open_government_meeting.html (Accessed: 12 May 2017).

Open Knowledge International (2017) *What is Open Data?*, *http://opendatahandbook.org/*. Available at: http://opendatahandbook.org/guide/en/what-is-open-data/ (Accessed: 25 May 2017).

openstreetmap.org (2017a) *Hackathon - OpenStreetMap Wiki*. Available at: http://wiki.openstreetmap.org/wiki/Hackathon (Accessed: 19 February 2017).

openstreetmap.org (2017b) *OpenStreetMap*. Available at: https://www.openstreetmap.org/#map=5/51.500/-0.100 (Accessed: 18 February 2017).

Pang, Yu and Orgun (2017) 'A novel end-to-end classifier using domain transferred deep convolutional neural networks for biomedical images', *Computer Methods and Programs in Biomedicine*, 140, pp. 283–293. doi: 10.1016/j.cmpb.2016.12.019.

van Panhuis, Paul, Emerson, Grefenstette, Wilder, Herbst, Heymann and Burke (2014) 'A systematic review of barriers to data sharing in public health', *BMC Public Health*. BioMed Central, 14(1), p. 1144. doi: 10.1186/1471-2458-14-1144.

Papadatos and Overington (2014) 'The ChEMBL database: a taster for medicinal chemists', *Future Medicinal Chemistry*. Future Science Ltd London, UK, 6(4), pp. 361–364. doi: 10.4155/fmc.14.8.

Parry (2014) 'Health Informatics', in *Springer Handbook of Bio-/Neuroinformatics*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 555–564. doi: 10.1007/978-3-642-30574-0_34.

Pathak, Kiefer and Chute (2012) 'Applying linked data principles to represent patient's electronic health records at Mayo clinic', in *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*. New York, New York, USA: ACM Press, p. 455. doi: 10.1145/2110363.2110415.

Paula Span (2015) *A Surgery Standard Under Fire - The New York Times*, *www.nytimes.com*. Available at: https://www.nytimes.com/2015/03/03/health/a-30-day-surgical-standard-is-under-scrutiny.html?_r=2 (Accessed: 8 May 2017).

Petersen, Feldt, Mujtaba and Mattsson (2008) 'Systematic Mapping Studies in Software Engineering', (Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering)

Poldrack, Barch, Mitchell, Wager, Wagner, Devlin, Cumba, Koyejo and Milham (2013) 'Toward open sharing of task-based fMRI data: the OpenfMRI project', *Frontiers in Neuroinformatics*. Frontiers, 7, p. 12. doi: 10.3389/fninf.2013.00012.

Poldrack and Gorgolewski (2015) ‘OpenfMRI: Open sharing of task fMRI data’, *NeuroImage*, 144, pp. 259–261. doi: 10.1016/j.neuroimage.2015.05.073.

Reiche, Höfig and Schieferdecker (2014) ‘Assessment and visualization of metadata quality for open government data’, in *CeDEM14 Conference for E-Democracy and Open Government*

Resnick (2017) ‘Treating depression is guesswork. Psychiatrists are beginning to crack the code. Brain scans and machine learning programs are paving the way for a breakthrough.’, *http://www.vox.com/*. Available at: http://www.vox.com/science-and-health/2017/4/4/15073652/precision-psychiatry-depression.

Roelofs, Dekker, Meldolesi, van Stiphout, Valentini and Lambin (2014) ‘International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining’, *Radiotherapy and Oncology*, 110(2), pp. 370–374. doi: 10.1016/j.radonc.2013.11.001.

Rojas, Bermúdez and Lovelle (2014) ‘Open Data and Big Data: A Perspective from Colombia’, in, pp. 35–41. doi: 10.1007/978-3-319-08618-7_4.

Rubio-Perez, Tamborero, Schroeder, Antolín, Deu-Pons, Perez-Llamas, Mestres, Gonzalez-Perez and Lopez-Bigas (2015) ‘In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities’, *Cancer Cell*, 27(3), pp. 382–396. doi: 10.1016/j.ccell.2015.02.007.

Saaty (2001) *Decision Making with Dependence and feedback. The Analytic Network process. The organization and priotitization of complexity.*

Samuel (1959) ‘Some Studies in Machine Learning Using the Game of Checkers’, *IBM Journal of Research and Development*, 3(3), pp. 210–229. doi: 10.1147/rd.33.0210.

Samwald, Jentzsch, Bouton, Kallesøe, Willighagen, Hajagos, Scott Marshall, M., Prud’hommeaux, Hassanzadeh, Pichler and Stephens (2011) ‘Linked Open drug data for pharmaceutical research and development’, *Journal of Cheminformatics*, 3(5), p. 19. doi: 10.1186/1758-2946-3-19.

Sandoval-Almazan and Gil-Garcia (2014) ‘Towards an Evaluation Model for Open Government: A Preliminary Proposal’, in, pp. 47–58. doi: 10.1007/978-3-662-44426-9_4.

Sayogo, Pardo and Cook (2014) ‘A Framework for Benchmarking Open Government

Data Efforts', *2014 47th Hawaii International Conference on System Sciences*, pp. 1896–1905. doi: 10.1109/HICSS.2014.240.

Scannapieco and Catarci (2002) 'Data Quality under the Computer Science perspective', *Archivi & Computer*

Schwarze, Brasel and Mosenthal (2014) 'Beyond 30-day mortality: aligning surgical quality with outcomes that patients value.', *JAMA surgery*. NIH Public Access, 149(7), pp. 631–2. doi: 10.1001/jamasurg.2013.5143.

Shadbolt and O'Hara (2013) 'Linked data in government', *IEEE Internet Computing*, 17(4), pp. 72–77. doi: 10.1109/MIC.2013.72.

Shi (2015) 'Guiding School-Choice Reform through Novel Applications of Operations Research', *Interfaces*, 45(2), pp. 117–132. doi: 10.1287/inte.2014.0781.

Sioutos, Coronado, Haber, Hartel, Shaiu and Wright (2007) 'NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information', *Journal of Biomedical Informatics*, 40(1), pp. 30–43. doi: 10.1016/j.jbi.2006.02.013.

Smith (2006) 'Microsoft purchases patient-data tracker', *The Washinton Times*. Available at: http://www.washingtontimes.com/news/2006/jul/26/20060726-101329-8737r/.

Sprivulis, Walker, Johnston, Pan, Adler-Milstein, Middleton and Bates (2007) 'The economic benefits of health information exchange interoperability for Australia', *Australian Health Review*. CSIRO PUBLISHING, 31(4), p. 531. doi: 10.1071/AH070531.

Stefaan Verhulst, Beth Simone Noveck, Robyn Caplan, Kristy Brown and Claudia Paz (2014) *The Open Data Era in Health and Social Care - The Governance Lab @ NYU*. Available at: http://thegovlab.org/nhs/ (Accessed: 7 May 2017).

Strathern (2000) 'The Tyranny of Transparency', *British Educational Research Journal*. Taylor & Francis Group , 26(3), pp. 309–321. doi: 10.1080/713651562.

Tao, Jiang, Oniki, Freimuth, Zhu, Sharma, Pathak, Huff and Chute (2013) 'A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data.', *Journal of the American Medical Informatics Association : JAMIA*, 20(3), pp. 554–62. doi: 10.1136/amiajnl-2012-001326.

tfl.gov.uk (2017) *Open data users - Transport for London*. Available at: https://tfl.gov.uk/info-for/open-data-users/ (Accessed: 12 February 2017).

The MIT Press (2017) *Neural Information Processing Systems | The MIT Press*. Available at: https://mitpress.mit.edu/category/discipline/computer-science-and-intelligent-systems/neural-information-processing-systems (Accessed: 4 March 2017).

The World Bank (2017) *Open Data in 60 Seconds | Data*. Available at: http://opendatatoolkit.worldbank.org/en/open-data-in-60-seconds.html (Accessed: 28 February 2017).

Tim Berners-Lee (2006) *Linked Data - Design Issues*, *www.w3.org*. Available at: https://www.w3.org/DesignIssues/LinkedData.html (Accessed: 11 March 2017).

Tim Berners-Lee (2009) *Is your Linked Open Data 5 Star?*, *www.w3.org*. Available at: https://www.w3.org/DesignIssues/LinkedData.html (Accessed: 14 May 2017).

Tygel, Auer, Debattista, Orlandi and Campos (2015) 'Towards Cleaning-up Open Data Portals: A Metadata Reconciliation Approach'. Available at: http://arxiv.org/abs/1510.04501 (Accessed: 8 June 2017).

Ubaldi (2013) 'Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives', *OECD Working Papers on Public Governance*, NO.22(22), p. 61. doi: 10.1787/5k46bj4f03s7-en.

UK Environment Agency (2017) *Bathing water quality*. Available at: https://environment.data.gov.uk/bwq/profiles/ (Accessed: 15 March 2017).

Umbrich and Neumaier (2015) 'Quality Assessment and Evolution of Open Data Portals', *Future Internet of Things*. Available at: http://ieeexplore.ieee.org/abstract/document/7300846/ (Accessed: 20 March 2017).

Umbrich, Neumaier and Polleres (2015) 'Towards assessing the quality evolution of Open Data portals.', in *In ODQ2015: Open Data Quality: from Theory to Practice Workshop, Hrsg. Network of Excellence in Internet Science, Munich, Germany.* Network of Excellence in Internet Science, pp. 1–5. doi: 10.1109/FiCloud.2015.82.

US National Library of Medicine - National Institutes of Health (2017) *Home - PMC - NCBI*, *www.ncbi.nlm.nih.gov*. Available at: https://www.ncbi.nlm.nih.gov/pmc/ (Accessed: 5 May 2017).

Valença, Alves, Alves and Niu (2013) 'A SYSTEMATIC MAPPING STUDY ON BUSINESS PROCESS VARIABILITY', *International Journal of Computer Science & Information Technology (IJCSIT)*, 5(1). doi: 10.5121/ijcsit.2013.5101.

Vasa and Tamilselvam (2014) 'Building apps with open data in india: an experience', *Proceedings of the 1st International Workshop on Inclusive Web Programming - Programming on the Web with Open Data for Societal Applications - IWP 2014*. doi: 10.1145/2593761.2593763.

Vetrò, Canova, Torchiano, Minotas, Iemma and Morando (2016) 'Open data quality measurement framework: Definition and application to Open Government Data', *Government Information Quarterly*, 33(2), pp. 325–337. doi: 10.1016/j.giq.2016.02.001.

Walker, Pan, Johnston, Adler-Milstein, Bates and Middleton (2005) 'The Value Of Health Care Information Exchange And Interoperability', *Health Aff*, p. hlthaff.w5.10. doi: 10.1377/hlthaff.w5.10.

Wang, Suzek, Zhang, Wang, He, Cheng, Shoemaker, Gindulyte and Bryant (2014) 'PubChem BioAssay: 2014 update.', *Nucleic acids research*. Oxford University Press, 42(Database issue), pp. D1075-82. doi: 10.1093/nar/gkt978.

Whitmore (2014) 'Using open government data to predict war: A case study of data and systems challenges', *Government Information Quarterly*, 31(4), pp. 622–630. doi: 10.1016/j.giq.2014.04.003.

WHO (2016) *WHO Cancer Mortality Database*, *http://www-dep.iarc.fr/WHOdb/WHOdb.htm*

Wishart, Knox, Guo, Shrivastava, Hassanali, Stothard, Chang and Woolsey (2006) 'DrugBank: a comprehensive resource for in silico drug discovery and exploration', *Nucleic Acids Research*, 34(90001), pp. D668–D672. doi: 10.1093/nar/gkj067.

World Wide Web Consortium (2012) *HCLSIG/LODD/Data*, *www.w3.org*. Available at: https://www.w3.org/wiki/HCLSIG/LODD/Data (Accessed: 31 March 2017).

Zeeberg, Riss, Kane, Bussey, Uchio, Linehan, Barrett and Weinstein (2004) 'Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics.', *BMC Bioinformatics*, 5(1), p. 80. doi: 10.1186/1471-2105-5-80.

Zhao (2010) 'Publishing Chinese medicine knowledge as Linked Data on the Web', *Chinese Medicine*, 5(1), p. 27. doi: 10.1186/1749-8546-5-27.

Ziemann, Eren, El-Osta, Zeeberg, Riss, Kane, Bussey, Uchio, Linehan, Smedley, Haider, Durinck, Pandini, Provero, Allen, Barrett, Wilhite, Ledoux, Evangelista, Kim and Tomashevsky (2016) 'Gene name errors are widespread in the scientific literature', *Genome Biology*. BioMed Central, 17(1), p. 177. doi: 10.1186/s13059-016-1044-7.

Zook, Graham, Shelton and Gorman (2010) 'Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake', *World Medical & Health Policy*. Blackwell Publishing Ltd, 2(2), pp. 6–32. doi: 10.2202/1948-4682.1069.

Zuiderwijk-van Eijk and Janssen (2015) 'Participation and Data Quality in Open Data use: Open Data Infrastructures Evaluated', *Proceedings of the15th European Conference on e-Government, Portsmouth, UK, 18-19 June 2015; Authors version*. ACPI. Available at: https://repository.tudelft.nl/islandora/object/uuid:c3e2530d-eaa2-409b-a700-b7107db7e159?collection=research (Accessed: 19 June 2017).

Appendix

Appendix 1 – literature list form preliminary mapping study

Open data applications, platforms, portals and initiatives
(Shi 2015)
(Dajer 2012)
(Sarah M. 2012)
(tfl.gov.uk 2017)
(Haklay & Weber 2008)
(Zook et al. 2010)
(Keegan 2010)
(Manyika et al. 2013)
(Vasa & Tamilselvam 2014)
(Whitmore 2014)
(UK Environment Agency 2017)
(Shadbolt & O'Hara 2013)
(Forbes et al. 2015)
(Rubio-Perez et al. 2015)
(Clark et al. 2013)
(Roelofs et al. 2014)
(Bray et al. 2015)
(Sioutos et al. 2007)
(US National Library of Medicine - National Institutes of Health 2017)
(Cancer Imaging Archive 2017)
(Kalpathy-Cramer et al. 2016
(Codella et al. 2014)
(Samuel 1959)
(Jordan & Mitchell 2015)
(LeCun et al. 2015)
(Herland et al. 2014)
(Pang et al. 2017)
(The MIT Press 2017)
(Agostinelli et al. 2013)
(Li et al. 2010)
(Mildenberger et al. 2002)
(Samwald et al. 2011)
(Wishart et al. 2006)
(Hassanzadeh et al. 2009)

(dailymed.nlm.nih.gov 2017)
(Boyce et al. 2013)
(Wang et al. 2014)
(Gaulton et al. 2012)
(Bento et al. 2014)
(Gaulton & Overington 2010)
(Papadatos & Overington 2014)
(Goh et al. 2007)
(Fang et al. 2008)
(Zhao 2010)
(Kuhn et al. 2010)
(Kuhn et al. 2007)
(Liu et al. 2005)
(Belleau et al. 2008)
(Chen et al. 2010)
(dbpedia.org 2017)
(Negru & Buraga 2013)
(Stefaan Verhulst et al. 2014)
(British Columbia Medical Association 2006)
(Manyika et al. 2013)
(Jessica Nutik Zitter 2017)
(Paula Span 2015)
(Ketelaar et al. 2011)
(Schwarze et al. 2014)
Open data enablers
(Demchenko et al. 2014)
(Tim Berners-Lee 2006)
(Shadbolt & O'Hara 2013)
(Martin et al. 2011)
(eHealth Ireland 2017)
(Kahn & Hederman 2012)
(Marco-Ruiz et al. 2016),
(Tao et al. 2013)
(Samwald et al. 2011)
(Pathak et al. 2012)
(Kamateri et al. 2014)
(Walker et al. 2005
Sprivulis et al. 2007
(van Panhuis et al. 2014)

(Martin et al. 2013)
(Poldrack et al. 2013)
(Poldrack & Gorgolewski 2015)
(Tim Berners-Lee 2009)
(Kienle 2012)
(El Emam et al. 2012)
(Kaiser et al. 2007)
(Attard et al. 2015)
(Kernighan 2017)
(Zeeberg et al. 2004)
Open data evaluations and guidelines
(Kučera et al. 2013)
(Vetrò et al. 2016)
(Behkamal et al. 2014),
(Kubler et al. 2016)
(Calero et al. 2008)
(Umbrich et al. 2015)
(Máchová & Lněnička 2017)
(Zuiderwijk & Janssen 2015)
(Umbrich & Neumaier 2015)
(Behkamal et al. 2014)
(Moraga et al. 2009)
(Scannapieco & Catarci 2002)
(Reiche & Höfig 2014)
(Open Government Working Group Meeting in Sebastopol 2007)
(Martin et al. 2013)
(Open Data Barometer 2015)
(Global Open Data Index 2017)
(Ubaldi 2013)
(Neumaier & Umbrich 2016)
(Herzog et al. 2007)
(Juran & Godfrey 1999)
(Kučera et al. 2013)
(Mestres et al. 2008)
(Martin et al. 2017)

Appendix 2 – further biomedical datasets

DailyMed provides officially labelled information about marketed drugs in the United States. Additionally, DailyMed is the official provider of Food and Drug Administration (FDA) label information, which is included in drug package inserts. Such information is available on the platform as a download and the labels have been reformatted to make them easier to read. The National Library of Medicine (NLM) provides the platform as a public service and the content is based on the most recent information submitted to the FDA (dailymed.nlm.nih.gov, 2017). Among others, the data has been used by researchers to dynamically enhance product labels to support drug safety, efficacy and effectiveness (Boyce *et al.*, 2013).

Diseasome includes characteristics of disorders and disease genes linked by known disease-gene associations. The data is generated from the Online Mendelian Inheritance in Man (OMIM) and includes more than 2500 genes (Goh *et al.*, 2007) (Samwald, *et al.*, 2011) (World Wide Web Consortium, 2012).

While adverse drug reactions are an important source of human phenotypic information, the lack of sufficient and accessible data is hindering research efforts. Project SIDER tries to tackle this issue be creating a public, computer-readable side effect source that connects 888 drugs to 1450 side effect terms (Kuhn *et al.*, 2010).

STITCH is a "search tool for interactions of chemicals" and includes a database of interaction information for over 68000 different chemicals and 2200 drugs. Further, it connects them to other data sources containing 1.5 million genes across 373 genomes and their interactions (Kuhn *et al.*, 2007).

Bio2RDF is a mashup of bioinformatics databases spread across the internet. The aim of the project is to build a bioinformatics knowledge system based on RDF. The datasets can be made available in RDF format through a unique URL in the style of http://bio2rdf.org/namespace:id (Belleau *et al.*, 2008).

The RxNorm is a nomenclature that facilitates the electronic exchange for drug information. To allow a smooth transition from different organizations or health information system, a standard nomenclature is needed and crucial to patient safety. Therefore, the National Library of Medicine created the standardized RxNorm nomenclature for clinical drugs. The clinical drug is represented in a semantic normal form that includes active ingredients, strength and form of the drug that is being administered. Further, the RxNorm includes a name for every dose and strength for combinations of clinically significant ingredients. (Liu *et al.*, 2005). Six out of twelve drug vocabularies were made available as part of the LODD could, including: Medical Subject Headings, Metathesaurus FDA National Drug Code Directory, Metathesaurus FDA Structured Product Labels, National Drug Files, RxNorm Vocabulary and Veterans Health Administration National Drug File. Also, the RxNorm is interlinked with the drug bank and Unified Medical Language System (UMLS) (World Wide Web Consortium, 2012).