Universiteit Leiden

Master Computer Science

Designing a FAIR (Findable, Accessible, Interoperable and Reusable) Data point for Digital Health in Kazakhstan

Name:	Aliya Aktau
Student ID:	s2027372
Date:	28/04/2020
Specialisation:	Computer Science and Business Studies
1st supervisor:	Prof.dr. M.E.H. van Reisen
2nd supervisor:	Dr. K.J. Wolstencroft

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

ACKNOWLEDGMENT

I would like to express my gratitude and appreciation to my thesis supervisors, Professor Mirjam van Reisen and Katy Wolstencroft for their excellent guidance and inspiring mentorship that helped me enrich my knowledge and pass through this difficult but no less exciting path. Along with them, I would like to express my gratitude to my husband Olzhas, who supported, encouraged and believed in me and my son Esmagambet, who is just there and always gives me joy and motivation ...

Abstract

by Aliya Aktau,

:

This explorative study focuses on innovation by the use of FAIR principles in Kazakhstan. For this, initiatives in the field of digital health were studied in order to understand how digital health is organized in Kazakhstan and what are the regulatory frameworks. Internships were held at the GO FAIR office and the Ministry of Health of Kazakhstan to better understand the real situation and how FAIR technology can solve healthcare problems in Kazakhstan. To deploy FDP, cancer data were obtained from the Ministry of Health of Kazakhstan. Deployability was investigated by using two theoretical lenses were provided by the Theory of Agenda-Setting of Kingdon and the Theory of Planned Behavior of Ajzen to study the relationship between public policy and attitudes, social norms, and FAIR principles. As a result, a mock FDP was developed using data on oncological diseases in Kazakhstan. It is concluded that FAIR Principles can be included in digital healthcare systems in Kazakhstan, to solve the problems at a technical policy level. At a users-level, the study finds that it will require understanding the relevance of FAIR-architecture for public health and scientific research in Kazakhstan, and thus, how FAIR-architecture can improve personalized medicine.

.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENT i

ABSTRACT ii

CHAPTER

1 Introduction 1

1.1 Problem statement 1

1.2 Research gap 3

1.3 Objectives 4

1.4 Kazakhstan 5

1.5 Research Questions 5

1.6 Research relevance 7

1.6.1 Academic relevance 7

1.6.2 Societal relevance 7

1.7 Ethical considerations 8

1.8 Research Outline 8

CHAPTER

2 Theoretical Framework 9

2.1 FAIR Data Principles 9

2.2 The Theory of Planned Behavior 10

2.3 Kingdon's agenda setting model 11

2.3.1 Problem stream 12

2.3.2 Policy stream 13

2.3.3 Political stream 13

2.3.4 Policy entrepreneurs 13

2.4 Conclusion 14

CHAPTER

3 Research methodology . . . 16
3.1 Research design . . . 16
3.2 Literature search . . . 19
3.3 The internships . . . 20
3.4 Data collection . . . 20
3.5 Implementation . . . 21
3.6 Obstacles . . . 22

CHAPTER

4 Digital Health policy of Kazakhstan . . . 23
4.1 Fundamental documents regulating access to data . . . 23
4.2 What is the policy (legislation) and origin of the policy? . . . 24
4.2.1 Regulatory frameworks for digital health in Kazakhstan . . . 26
4.2.2 Policies on cancer . . . 27
4.3 Implementing Digital Health in Kazakhstan . . . 27
4.3.1 Electronic Health Records . . . 28
4.4 The acceptance of FAIR Principles by the Theory of Planned Behavior 29
4.5 Internship outcomes . . . 30
4.6 Challenges of the digitization . . . 30
4.6.1 Data infrastructure issues . . . 30
4.6.2 Policy issues . . . 32
4.6.3 Open Science policies . . . 33
4.7 Conclusion . . . 33

CHAPTER

5 FAIR-ness of digital health data of Kazakhstan . . . 34
5.1 Information systems of Kazakhstan . . . 35
5.2 Analysing information systems . . . 38
5.3 Health related data available on the Open Data Portal of Kazakhstan . 39
5.4 Assessment of the FAIRness of health-related data . . . 41
5.4.1 Provenance of the policy . . . 41
5.5 F - Findability . . . 41
5.5.1 An analysis of F1 - F4 . . . 42
5.5.2 Challenges . . . 44
5.6 A - Accessibility . . . 45
5.6.1 The current state of the art . . . 45
5.6.2 Challenges . . . 46

5.7 Interoperability and Reusability . . . 46
5.7.1 An analysis of I + R . . . 47
5.8 R facet of FAIR . . . 49
5.9 Conclusion . . . 49

CHAPTER

6 Comparing Cancer Data with Global Cancer Data . . . 51
6.1 Data description . . . 51
6.2 Comparison with international cancer database . . . 54
6.3 Conclusion . . . 59

CHAPTER

7 Designing a FAIR Data Point . . . 60
7.1 Objectives . . . 60
7.2 Implementation . . . 61
7.2.1 Implementation choices . . . 61
7.3 FAIRification process . . . 61
7.3.1 OpenRefine . . . 65
7.3.2 FAIR Data Point Components . . . 66
7.3.3 Metadata Specification . . . 67
7.3.4 Metadata of Datasets . . . 69
7.4 Results . . . 70
7.5 Obstacles . . . 72
7.6 Conclusion . . . 74

CHAPTER

8 Discussion and Conclusion . . . 75
8.1 Addressing digital health initiatives using Kingdon's Agenda Setting Model . . . 77
8.1.1 Problem stream . . . 77
8.1.2 Policy stream . . . 78
8.1.3 Political stream . . . 79
8.1.4 Policy entrepreneurs . . . 79
8.2 Obstacles . . . 80
8.2.1 The acceptance of technology . . . 80
8.3 Conclusion . . . 81
8.4 Limitations . . . 82

APPENDIX

.1 Metadata 84

REFERENCES 102

Chapter One

Introduction

1.1 Problem statement

FAIR Data Principles are guiding principles to create reusable research objects. Data should be findable, accessible under well-defined conditions, interoperable without data munging, and therefore optimally reusable to serve for better purposes (Mons et al., 2019). Data being available and useful for the scientific community is the utmost issue at present. The primary goal is to reuse scientific data, and thus, making data accessible both for humans and machines are necessary for that.

A critical moment has now been reached at which the analysis and storage of annotated clinical and genomic information in disconnected bunkers will stop the development of research, in particular, precision cancer treatment (Siu et al., 2016). Data in most electronic health records (EHR) systems is not checked for quality and is not structured so that it can be easily retrieved (Siu et al., 2016). These issues are getting worse when data needs to be compared and used across institutions, and this becomes a significant barrier to cross-border data exchange initiatives (Siu et al., 2016). Although researchers are not entirely able to generate all the necessary data required to bring essential conclusions within one project, they can enhance their research by reusing data from other projects (Grossman et al., 2016). This also considers cancer research. Precision oncology, in particular, its principles and practice might be improved through the exchange of data from hundreds of cancer patients (Grossman et al., 2016). In this regard, bringing all the necessary data for one place making them findable, accessible

under well-defined conditions, interoperable and reusable (FAIR) will help to improve cancer research.

Initially the establishment of the internet of FAIR Data and services (IFDS) started in the domains of life and natural sciences with the perspective of inclusion other domains (van Reisen et al., 2019). These days, the implementation of the IFDS is expanding worldwide and covering more scientific disciplines. The number of articles on this area has increased between 2016 and 2019 and contains 1570 (van Reisen et al., 2019). FAIR obtained acknowledgment of the European Union, the G7, the G20 and US-based big Data to Knowledge and an African Research Cloud in 2016 (van Reisen et al., 2019). However, a vast amount of articles shows that FAIR principles are mostly implemented in the European countries, and lesser in the US (van Reisen et al., 2019). Moreover, large-scale genetic studies of human diseases do not reflect the level of diversity at the global level, since they are mainly based on populations of European origin (Popejoy & Fullerton, 2016). Therefore, the ability to translate genetic research into clinical practice may be inaccurate due to the lack of ethnic diversity in studies of the human genome (Sirugo et al., 2019). This implies that data used in research is mostly biased towards European geographies or European ancestors, causing difficulties in validity and representativity of data (van Reisen et al., 2019). Extending FAIR principles to non-European ancestors may increase the credibility of these studies.

Currently, Kazakhstan lacks FAIR principles in any scientific discipline, including healthcare. The country is willing to use personalized medicine for patients, specifically, within the oncology area (MoH & KazSRIOR, 2018), which in turn requires appropriate data infrastructure. Thus validity and representativity of data is the utmost need, since the more patient data, the better and more accurate the treatment. Thus, the inclusion of Kazakhstani data in the international arena can lead to better results in cancer research both international and in Kazakhstan, and therefore better cancer treatment of patients.

According to Hilbert (2016), in Big Data Analytics, every decision made is based on prior information. With respect to probability, the decision is uncertain unless the structure of prior information is improved. If we improve the structure of prior information on which we base our estimates, then the better the estimate, therefore, the better the decision made (Hilbert, 2016).

Probability of uncertainty in such a case can be reduced based on the improvement of the structure of prior information (Hilbert, 2016). Improvement of the validity and representativity of data as well as data infrastructure, which brings to the improvement of the entire information structure, might help to make better and accurate decisions. Applying FAIR Data Principles might bring Kazakhstan closer to international research, thus improving not only Kazakhstani research but also global research, in particular, cancer research.

1.2 Research gap

Large-scale studies have not captured the level of diversity and mostly based on individuals of European ancestry (Popejoy & Fullerton, 2016). Understanding of the genetic determinants of disease risk came through GWAS of people of European descent, with limited representation from other groups, including from Africa, North and South America, Asia and Oceania (Gurdasani et al., 2019). There were attempts to change the situation in earlier 2010s, and the 1000 Genomes project was launched to diversify genetic data and create a broad understanding of human genetic variation across multiple populations (Clarke et al., 2012). However, as of 2018, most of the genome-wide association studies (GWAS) were still held in European or Asian population groups (where East Asian population group dominates) with the 78 per cent of European ethnicity (Sirugo et al., 2019). This European bias has significant consequences for predicting the risk of disease among the world's population. Integration of more diverse populations for empirical and theoretical reasoning is necessity (van Reisen et al., 2019). Expanding FAIR Data implementation towards other countries can diversify data and help to do the right implications from that. Therefore, the data diversity towards other countries is necessary at this step and requires a new flow of data stream. The diversity among genetic variations can affect the treatment processes as well as disease risk. Studying diverse populations increases the ability to predict disease and understand genetic disease architecture. This will result in precision of medical care (Sirugo et al., 2019).

Recent studies have shown that cancer is so heterogeneous that individual research centres cannot collect enough data to fit prognostic and predictive models with sufficient accuracy

(Vesteghem et al., 2019). Therefore, the exchange of data in oncology is of utmost importance. The principles of data findability, accessibility, interoperability and reuse (FAIR) have been developed to define best practices for exchanging data (Vesteghem et al., 2019). Although the focus on data collection concentrates primarily on interoperability and reuse through common standards and harmonization, there is still need to address the issue of data exchange to make data findable and accessible in practice (Vesteghem et al., 2019). Making cancer data discoverable and accessible by other researchers will help for research, and in particular, oncology field.

1.3 Objectives

Efforts of specific studies and individual researchers, genetic information, detailed clinical data and research/trial datasets are rarely ever connected. Several types of data employed while conducting particular research, such as genomic data, health records and biosamples (e.g. blood or DNA), are currently fragmented in different databases. The number of patients suffering from cancer, including rare types of cancer and scattering of patients worldwide complicate the task of obtaining data for research and clinical trials.

The objective of this research is to understand how digital health is set up in Kazakhstan, including the legislation related to digital health, verification of digital health and interoperability of digital health. Then come up with recommendations on how FAIR data policies can be created and adopted in Kazakhstan's healthcare sector, and how FAIR data principles might address the existing problems within the health sector. This includes the data fragmentation problem within digital health, also the willingness of Kazakhstan to gain knowledge from foreign counties on addressing cancer illnesses throughout exchanging experience. This research is going to build the prototype of FDP using cancer data of Kazakhstan to help Kazakhstan exchange data on global level, and moreover, diversification of data will help to address cancer diseases globally.

1.4 Kazakhstan

Kazakhstan is a Central Asian country with the population 18,7 million as of March 31, 2020 [1]. It is the ninth-largest country in the world, with an area 2,7 million square kilometres and population density 7 people per square kilometre[2]. It is a middle-income country with emerging economies mostly dependent on mineral resources, thus extremely sensitive to external changes (Obermann et al., 2016). Over the past decade, Kazakhstan has been actively working on ICT and digital healthcare development infrastructures to make healthcare better accessible to citizens. The market for e-health solutions in Kazakhstan has begun the development and implementation of international standards, and building an integrated health data architecture, primarily with the formation of the Electronic Health Record (EHR) of citizens. Smart medicine, remote diagnostics and e-health are presented as a solution for such a geographically large country (Nazarbayev, 2012) since medical facilities and access to healthcare can be a problem for remote areas.

1.5 Research Questions

Originating from the problem statement and supported by the findings on FAIR Data Principles and digital health in Kazakhstan, the following research questions are formulated in Table 1.1. which will be addressed throughout this study:

[1] https://www.worldometers.info/world-population/kazakhstan-population/

[2] https://www.worldometers.info/world-population/kazakhstan-population/

Objectives:	Sub questions:	Method:	Findings:
To highlight the current situation of digital health in Kazakhstan and to understand how digital health is set up in Kazakhstan	**Sub question 1:** How digital health is set up in Kazakhstan? Policies on digital health and implementations **Sub question 2:** Challenges of digitization	Feasibility study	Chapter 4
Find out what kind of digital health data is available in Kazakhstan and assess the FAIRness of digital data	**Sub question 3:** How can digital health data be accessible from outside? **Sub question 4:** An analysis of digital health data on FAIRness	Case study	Chapter 5
Check health data from the Ministry of Health of Kazakhstan for their alignment with global cancer data.	**Sub question 5:** Compare Cancer Data Parameters with Global Cancer Data	Case study	Chapter 6
Develop a FAIR data-based model that illustrates how health data in Kazakhstan can be linked to and benefit from global Open Science	**Sub question 6:** Implementation, Challenges and Results	Design research	Chapter 7

Table 1.1 Research Questions

1.6 Research relevance

This section will discuss the contribution of this study to society and science.

1.6.1 Academic relevance

This research is conducted to investigate the impact FAIR can have on digital health system in Kazakhstan. Although analysis of various scientific articles related to FAIR has shown that many initiatives in the field of digital health have been carried out in Europe and America based on FAIR, in Kazakhstan digital health initiatives based on FAIR has not received much attention. These articles on FAIR were classified according to their topic, their geographical position, and whether they were articles on FAIR or which referred to FAIR. Ongoing research provides such an investigation, combining it with setting a public policy agenda; discuss how to develop a prototype using FAIR data principles; which data and data parameters are the best to add value to open science around the world and the barriers in adopting the FAIR Data Principles for Kazakhstani Digital Health.

1.6.2 Societal relevance

Dissemination of FAIR data in Kazakhstan can lead to data diversity and weaken the bias in life sciences towards Europe and America. For Kazakhstan the establishment of FAIR Data will provide access to global scientific data, while simultaneously improving science and bringing local scientists to the global level. FAIR compliant (meta)data can lead to cost reduction in many scientific projects which requires much effort and time in the data munging process (Mons et al., 2017). Thus, Kazakhstan may significantly benefit from deploying FAIR for their digital health data. Furthermore, there is a need for global analytics related to health data, in particular, for the analysis of certain diseases / outbreaks. The availability of medical data from various resources, which contribute to revealing important patterns and lead to more accurate decision-making, will help to cope with various diseases around the world as well as in Kazakhstan.

1.7 Ethical considerations

This research will be using data on cancer of anonymous patients from the electronic registry of cancer patients (EROB) from 2013 to 2018 with breast cancer, pancreatic cancer and Sarcoma. Data from the Ministry of Health of Kazakhstan was obtained on the basis of a non-dissemination agreement and that the data will be used for this research. The publication of this study will require additional ethical approval from the Kazakh Research Institute of Oncology and Radiology, as they are the responsible organization for the publication of any cancer research data in the country.

1.8 Research Outline

In the first chapter we discussed the topic, defined the problem statement along with the goals that are addressed by the main questions and auxiliary questions of the study, and, finally, the significance of the study. Theoretical framework is given on the second chapter. The third chapter consists of the research methodology and which designs are used for this study. The fourth chapter presents findings on digital healthcare initiatives in Kazakhstan, data structure developments, the challenges Kazakhstan has encountered in deploying an integration platform for digital healthcare, and existing challenges in this area. The fifth chapter presents an assessment made for Kazakhstan's digital health data using the FAIR Evaluation Services. The sixth chapter describes datasets used for this study and the value they bring to global science. The seventh chapter presents the process behind designing a FAIR data-based prototype. Last chapter is the conclusion and discussion chapter.

Chapter Two

Theoretical Framework

2.1 FAIR Data Principles

Data-driven technologies are transforming industries, our daily lives and routines, and also how we perform research. More and more data are being produced in our daily lives, including the healthcare ecosystem which gathers data from various sources such as healthcare providers, biomedical institutions and by citizens themselves. There is potential knowledge that needs to be discovered beyond the existing data that might transform health care delivery and life sciences. Adapting FAIR Data Principles could likely power the data collected from numerous sources to improve prevention, diagnosis and treatment of diseases, as well as supporting individuals and societies to maintain their health and well-being. The FAIR stands for Findable, Accessible, Interoperable, and Reusable for both machines and humans (Wilkinson et al., 2016). It was first discussed in 2014 in Leiden, the Netherlands, with the goal of ensuring accurate search, citation and reuse of digital objects over time (Wilkinson et al., 2016). The objectives of these principles are to make data discoverable for machines and humans, easily accessible by machine and humans, interoperable, for data to be able to integrate with other data, and reusable by other parties with the permission of the original source or data owner (Wilkinson et al., 2016). Because of the sensitive nature of health data the Personal Health Train (PHT) have been created, which enables data visiting through establishing distributed data analytics infrastructure (van Soest; Oliver Kohlbacher; Lukas Zimmermann; Holger Stenzhorn; Md. Rezaul Karim; Michel Dumontier; Stefan Decker; Luiz Olavo Bonino da Silva Santos;

Andre Dekker, 2020). For this, the work of the Personal Health Train has demonstrated that it is possible (van Soest; Oliver Kohlbacher; Lukas Zimmermann; Holger Stenzhorn; Md. Rezaul Karim; Michel Dumontier; Stefan Decker; Luiz Olavo Bonino da Silva Santos; Andre Dekker, 2020), but the regulatory and governing frameworks need to be further developed before widespread implementation can take place.

FAIR Data principles are the way of facilitating knowledge discovery from any data. FAIR is introducing itself not as a standard, but as the set of principles in order to alleviate the process of re-use of data (Mons, 2018). These principles provide guidance for scientific data management and stewardship and are relevant to all stakeholders in the current digital ecosystem. They directly address data producers and data publishers to promote the maximum use of research data. Research libraries can use the FAIR Data Principles as a framework for fostering and extending research data services.

For this study, the Theory of Planned Behavior will be used to consider how the adoption of FAIR Data Principles can be met in certain contexts. Besides that, Kingdon's agenda setting model will also be used to identify the existing digital health problems in Kazakhstan, including digital health policies, and possible solutions to the problems will be given as part of this study.

2.2 The Theory of Planned Behavior

The Theory of Planned Behavior will be used to predict and explain human behavior in certain contexts. The theory addresses the attitudes toward the behavior, subjective norms, and perceived behavioral control, as shown in Fig.2.1 (Ajzen, 1991). Attitude is positive or negative feelings of a person about performing a given behavior (Fishbein & Ajzen, 1975). In this case, the attitude of users/medical workers towards technology will be discussed. Subjective norm is "a person's perception that most people who are important to him think that he should or should not perform this behavior" (Fishbein & Ajzen, 1975). As subjective norms, policies on digital health and digital health data will be checked for feasibility of applying FAIR Data Principles for healthcare in Kazakhstan. Perceived behavioral control refers to people's perceptions of their ability to perform the target behavior (Fishbein & Ajzen, 1975). The technical level of

healthcare sector, specifically, health-related data available for the establishment of FAIR-Data will be used as variables of the perceived behavioral control. These three independent variables will help to understand the feasibility of applying FAIR Guiding Principles for the healthcare of Kazakhstan.

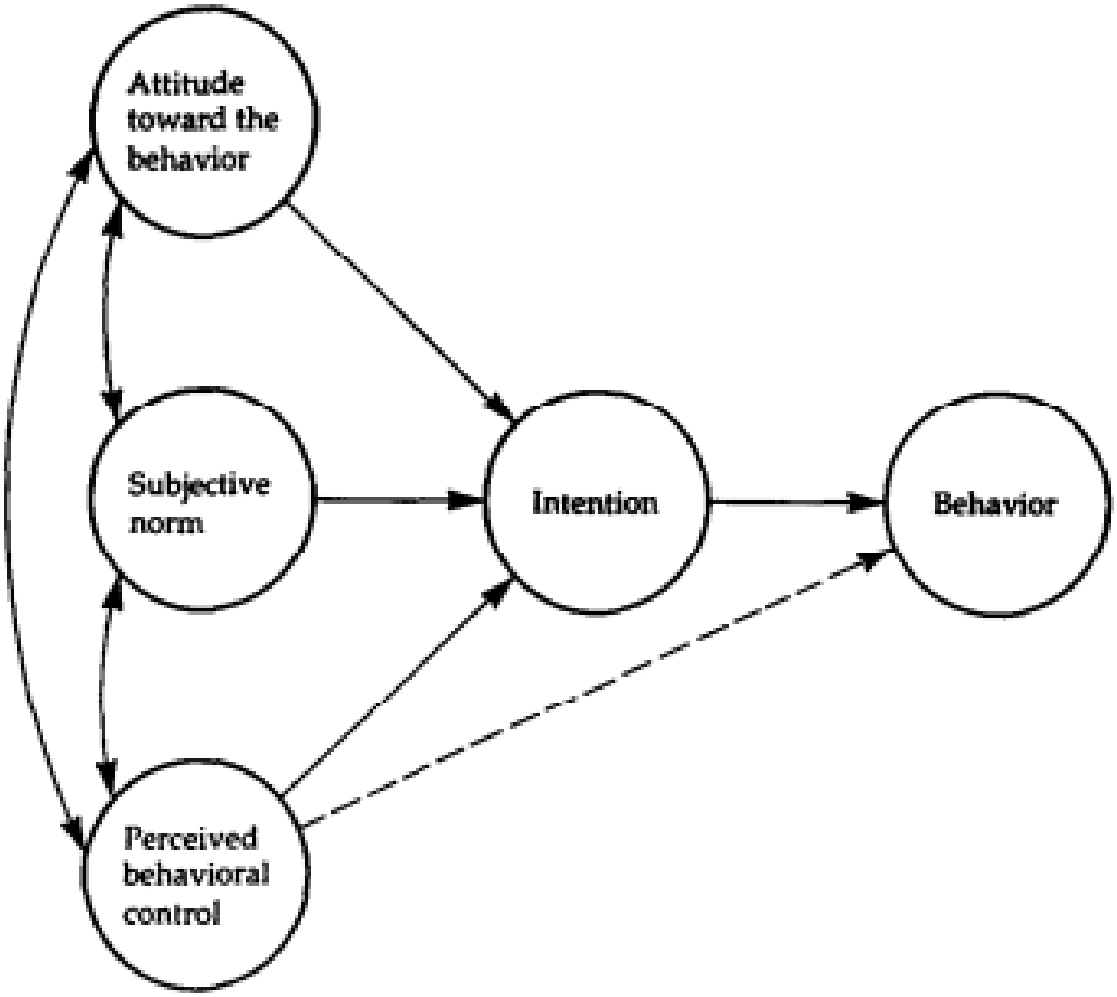

Figure 2.1 Theory of planned behavior (Ajzen, 1991)

2.3 Kingdon's agenda setting model

Kingdon's agenda setting model also will be used for this research. Political scientists have created various theories and models to study local policy making processes. However, John W. Kingdon's multiple streams model has been at the forefront for over more than thirty years in public agenda-setting. His multiple streams model composes three main streams: problem,

policy and political (Quirk, 1986) as illustrated in Fig.2.2. Policy change can happen when all three streams are aligned and support one another. Otherwise, the policy change might be difficult or even unlikely to achieve any results (Quirk, 1986). Moreover, policy change can happen if "policy window" is open and the opportunity is taken at the given time. This is the short period in which circumstances are right to push forward new policy (Quirk, 1986). Kingdon's model for agenda-setting of public policy will be used to address the healthcare situation in Kazakhstan by his main three streams.

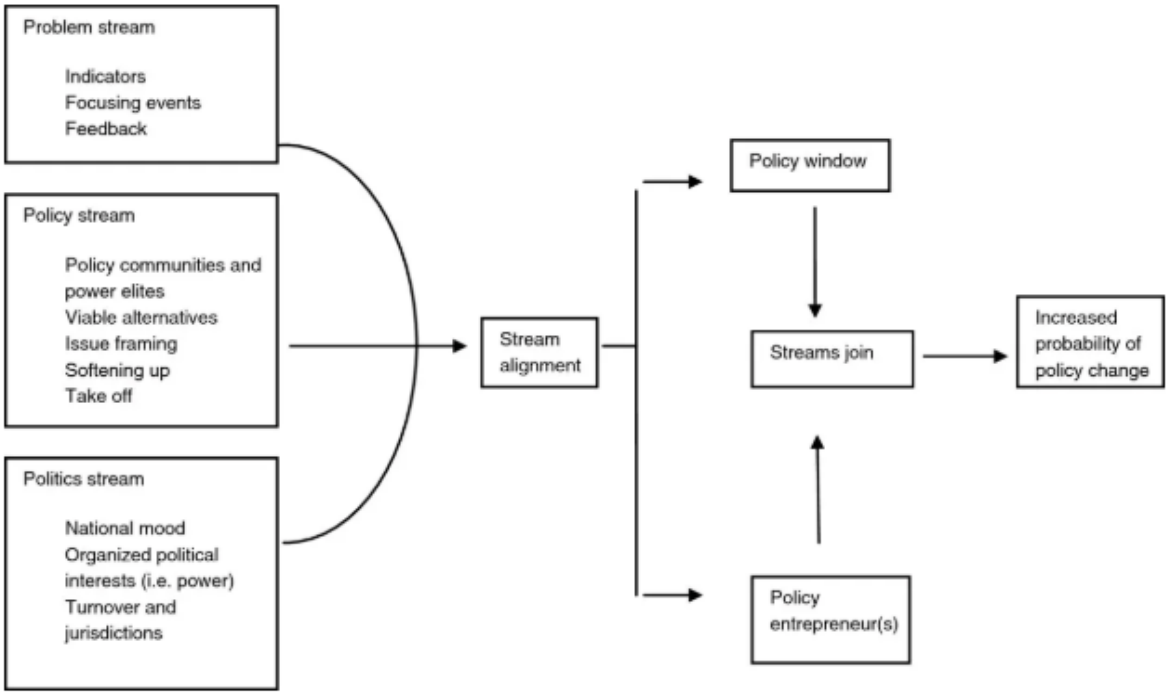

Figure 2.2 John W. Kingdon's multiple streams model (Gagnon & Labonté, 2013)

2.3.1 Problem stream

The first stream is the problems stream where public issues are identified occupying the attention of people working on governmental positions or decision-makers. Problems are defined more or less through systematic indicators which show that there is a problem (J. Kingdon, 1995) less from crucial events (Quirk, 1986). In the case of this research, data is difficult to find, access, interoperate and reuse by machines and humans. Thus, informational value of data has not been sufficiently used. Furthermore, a report published by the EU concluded that

the cost of missing FAIR research data could turn into billions (European Union, 2019). Applying FAIR Data Principles in research and healthcare in Kazakhstan can save huge amounts of money and allow more efficient use of data.

2.3.2 Policy stream

The second stream is the policy stream where policies are formed and proposed by people in government or people around government (J. Kingdon, 1995). People are outside government can also influence on policy making processes. The domain of experts including people from government, researchers, academics gather ideas, refine and propose solutions which relate to the problems identified in the problems streams (J. W. Kingdon, 2013). They offer solutions to the problem(s) and alternatives to them, as well as trying out the ideas of others and exchanging them, thereby offering completely new or altered solutions (J. W. Kingdon, 2013). In this case, The FAIR Data principles were stemmed from other concepts such as, Semantic Web and Linked Data which refer to connecting structured data on the Web (Bizer et al., 2009). This also can be seen from early FAIR examples where RDFs [1] and widespread ontologies or dictionaries are prominently used (Mons et al., 2017).

2.3.3 Political stream

The last political stream is composed of such influential things as national mood, internal and external changes of the country, changes of administration which can affect the overall mood of the country and political willingness to develop any agenda. A short period in which all streams come together and circumstances are encouraging to propose policy changes is called a "policy window" (Quirk, 1986).

2.3.4 Policy entrepreneurs

According to Kingdon's model, entrepreneurs play a crucial role in 'softening up' the system and connecting to the problem, policy and politics streams. Policy change cannot happen

[1] RDF is a standard model for data interchange on the Web

without policy entrepreneurs contribution (Gagnon & Labonté, 2013). This group of people push their proposals or push attention to particular problem(s). Proposals can be made when the time is good or bad, but an open policy window gives them a special opportunity, thus they have keep their proposals ready to advocate them when political climate is good (Wilson, 1993).

2.4 Conclusion

The Theory of Planned Behavior has been applied to this research. Intentions to perform behaviors of users/patients, healthcare professionals, policies and technical level will be conducted as independent variables of the theory of planned behavior. These constraints are given to understand how they affect the central factors in the theory of planned behavior such as intended behavior and behavior. These resources and opportunities give the likelihood of behavioral achievement of this research and will be properly reviewed from the given perspectives.

Kingdon's agenda-setting model has been applied to this study since it determines three streams for new ideas to get on a policy agenda. As a problem stream, the informational value of data has not been utilized appropriately due to problems with the findability, accessibility, interoperability and reusability of data. As a policy stream, FAIR Data Principles to make better use of data. As a political stream, a new e-health program for 2020–2025 ("State Health Development Program for 2020-2025", 2019), in which data interaction is one of the priorities, and the Comprehensive Cancer Control Plan for 2018–2022 (MoH & KazSRIOR, 2018), aimed at gaining international experience for better cancer treatment.

Any issues on the agenda in the process of developing public policy are determined by people inside and outside governments and consist of a stream of problems, a stream of politics, and a political stream. When three streams come together, a specific problem becomes prominent on the agenda, proposals corresponding to the problem attract attention, and then a policy change becomes possible. They can promote a particular idea (solution or policy), and they can activate the political will to accept the proposed idea as relevant to solve the recognised problem (van Reisen et al., 2019). In this regard, members of the Ambassadors

IN and VODAN IN act as policy entrepreneurs whose main aims are promoting FAIR Data Principles to facilitate digital data-driven science.

Apart from that, the spread of the virus that causes COVID-19 has been recognized as a pandemic these days [2] which has devastating impact across the world. Using machine learning and AI approaches can better address the issue through discovering meaningful patterns, and therefore, being data in a FAIR manner can alleviate this process. Consequently, the COVID-19 pandemic can be attributed to a focusing event that causes policy changes on specific issues, in particular, data policy changes that may affect better patient care or deal with future outbreaks. Focusing events are sudden events that attract the attention of the public and politicians, which serves as a potential incentive for policy change. These events are an important opportunity for politically disadvantaged groups looking for policy changes to push forward their proposals (Birkland, 1998). Concerning this, VODAN IN [3] might play an important role that was created to help to deal with the epidemic. This can be done through FAIRifing data from electronic health records to create proper FAIR Data and use distributed learning where machines will be able to learn new patterns from existing data.

[2] http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic

[3] https://www.go-fair.org/implementation-networks/overview/vodan/

Chapter Three

Research methodology

3.1 Research design

Exploratory design, feasibility and case study research design are used for this study, since no previous FAIR Data studies were conducted for Kazakhstan. One of the goals of this exploratory research is to understand the entire picture of the current situation of the digital healthcare of Kazakhstan, which platforms are used, and availability of digital health data. The next goal is determining the conditions for FAIR Data Implementation. Lastly, possible limitation and obstacles for the implementation of FAIR Data to provide direction for future research.

In design study, it is relevant to explore the situation before defining the design process in more detail. The fuzzy front end will be conducted for earlier stages of the research and design research will be applied for the development part of the project. The fuzzy front end is an important and early stage of an innovation process. It starts from the idea generation to either its approval or disapproval (Murphy & Kumar, 2002). The early phases of the process innovation usually has significant effect on the entire process and the results. Predevelopment activities impact the design, total costs and the timing of the project. At the early stages, the degree of freedom in design and influence on project outcomes are high, optimization can be made with little effort and can influence positively on the entire process. This benefit is limited by the amount of information and its certainty in comparison to later phases of the innovation process. Clear decisions cannot be made at this stage unless the necessary information is gathered (Herstatt & Verworn, 2004). For this, the data published by the government of

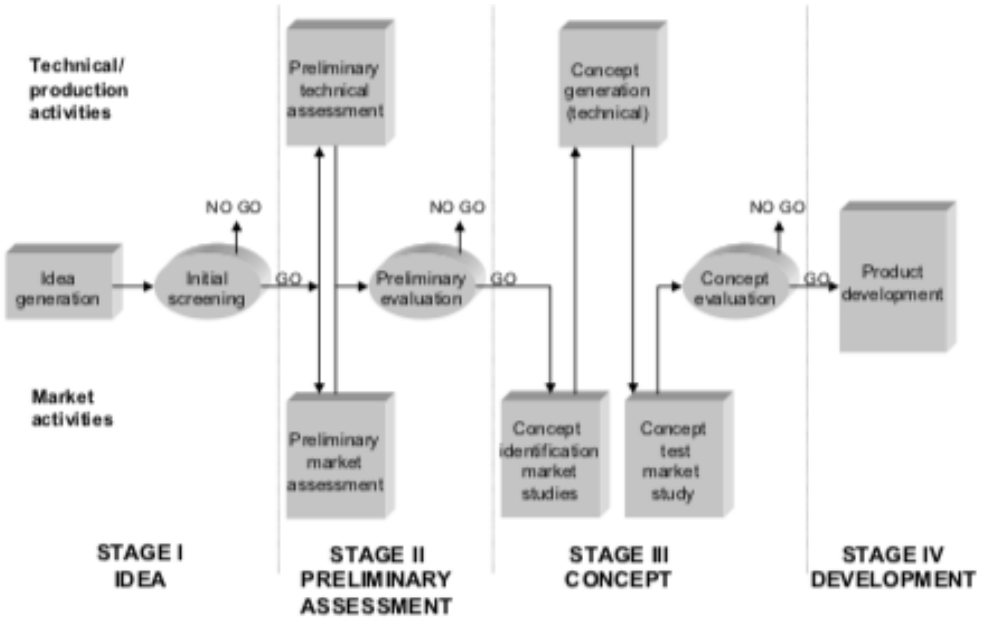

Figure 3.1 The "stage-gate"-process (Herstatt & Verworn, 2004)

Kazakhstan, policies on health, regulatory frameworks, as well as digital health data available on the Internet are checked. Apart from that, the contacts are made with the representatives of the Ministry of Health. Based on all this, decisions will be made with further stages. For the study, a "stage-gate" process model was selected. In this case, the preliminary development activities are divided into sub-stages, Fig.3.1, from idea generation to concept evaluation, and at each stage decisions are made to continue or terminate the project. A product development stage at the end of this model is only reachable if every stage ends with a "GO" decision (Herstatt & Verworn, 2004). The "stage-gate" process model was chosen as it has proven helpful in the case of incremental innovation (Herstatt & Verworn, 2004).

Incremental innovation is implied to the firm when market and technical uncertainty are low. This includes the expansion of the product line, small product improvements which could result in a competitive advantage of the product. (Herstatt & Verworn, 2004). Since digital health data in Kazakhstan remains fragmented and unrelated to global open science, we can assume that market uncertainty is low for adopting the FAIR Data Principles. Similarly, since the introduction of the Internet of FAIR Data and Services (IFDS) has taken place in different countries and institutions, this knowledge will be used and applied, so we can also assume that technical uncertainty for the expansion of FAIR Data Principles is low.

"Front-loading" problem solving strategy is also chosen as it improves the development performance by solving the problems at earlier phases. This is made by transferring knowledge from former projects, in particular, problems occurring during the development process of other projects (Herstatt & Verworn, 2004). The "front-loading" problem solving-approach will be used and for that I will consider FAIR-implementation projects in other countries and their problems occurring during the implementation. This may positively influence on the lead-time of the product development process.

Feasibility studies are used in this research, to determine whether an innovation is appropriate for further testing of the relevance and sustainability of this intervention (Bowen et al., 2009). The main question of the feasibility research design is "does it work?" under actual conditions compared to other practices (Bowen et al., 2009). There is a need to study a feasibility of FAIR-Data interventions in a specific country, in this case, Kazakhstan with in-depth research of the state programs and accessibility of digital health data and platforms. Appropriate areas of focus are the implementation of FAIR Data and its practicality. Firstly, Implementation is focused on the extent, likelihood and manner in which technology can be implemented. Second, practicality is about exploring whether the technology can be established with limited resources, in this case, availability of data on digital health and data policies of the country.

A design study will be applied for this research, and FAIR Data Point [1] will be designed using digital health data of Kazakhstan.

A case study will also be applied for this study with the particular focus on Kazakhstan, to explore difficulties, challenges and possibilities for a FAIR Data Point. Kazakhstan was chosen for this study because there is no FAIR Data Studies made previously, and there is still some room for improvement of digital health data infrastructure and for Open Science Kazakhstan. To make the case study feasible, an internship was obtained from the Ministry of Health of Kazakhstan to understand how digital healthcare works in Kazakhstan and what challenges Kazakhstan faces, what has been done and how FAIR data principles can be applied to improve healthcare field of Kazakhstan.

[1] FDP is a software that allows data owners to expose datasets in a FAIR manner

3.2 Literature search

Our first goal was to get an overview of the latest data exchange initiatives in the field of digital health in Kazakhstan, its policies, regulatory frameworks. Later, when contacts with the Ministry of Health were established, initiatives in the field of oncology were studied, and it became known that in oncology there is a room for improving the data structure. For the designing FDP, data on breast cancer and pancreatic cancer were taken from the Ministry of Health. To this end, the literature search was done on Google to find related articles and any publications on the web, at the same time Google Scholar was used to getting articles on certain scientific topics, international organizations websites of World Health Organization (https://www.who.int/) and World Bank (https://www.worldbank.org/), the site of the MOH Kazakhstan (https://www.dsm.gov.kz/) and Republican Center for Health Development of the Ministry of Health of Kazakhstan (https://www.rcrz.kz).

An initial literature review was carried out to point out the potential literature sources, terminology, and keywords that are related to FAIR data use in healthcare. It was also decided to focus on practical implementation and examined confidentiality issues, and / or policies. The following search criteria were used: health data, digital health data Kazakhstan, cancer data Kazakhstan, oncology, cancer, data sharing, data exchange, personalized medicine, interoperability platform, information dissemination.

The inclusion criteria were as follows:

Applicable to digital health Kazakhstan

Directly applicable to cancer (Kazakhstan/global)

Directly applicable for data exchange (Kazakhstan/global)

Directly applicable to clinical and / or medical data (Kazakhstan/global)

We also addressed ethical issues, and / or privacy issues, and / or policies.

The research's scope is limited to publications that were published in the last 5-7 years. This is because of the relatively recent evolution of FAIR data and healthcare analytics. Thus, any publications that are less than 2 years old will have more value as compared to those published 4 or 5 years ago. High-quality research is conducted on reliable resources that are

deemed valid for use.

3.3 The internships

For a better understanding of the FAIR Data Principles an internship was carried out at the GO FAIR International Support and Coordination Office (GFISCO). In addition, to understand how GFISCO supports the development of infrastructure for machine-readable research data and other digital resources. At GFISCO the coordination is made by various stakeholders through Implementation Networks (IN) on three main pillars: GO CHANGE, GO BUILD and GO TRAIN (GO FAIR). The main goal for this internship was to understand how FAIR Data Principles can address Kazakhstan's health-related issues. This internship helped to create a better understanding of FAIR Data and what are the possible solutions it might offer for various topic domains and countries. During the internship the research was introduced to how FAIR Data works technically, politically and what are the possible challenges for its adoption and expansion throughout the world.

The second internship was undertaken at the Ministry of Health in Kazakhstan. This internship also was helpful to understand the health data infrastructure within the MOH, particularly, how the MOH works towards reaching the seamless integration of digital health systems, and how they are adapting international medical standards. This also concerns the obstacles MOH is facing these days and how it wants to handle the issues especially in reaching interoperability of health data. Furthermore, the notes regarding reactions on the deployment of FAIR innovation was taken in a research logbook to assess the attitude and culture towards adopting FAIR innovation in Kazakhstan.

3.4 Data collection

The data for this study was obtained from the Ministry of Health of Kazakhstan when I was doing an internship there. I decided to select cancer data for the deployment of FDP, as there are a lot of people suffering from cancer in Kazakhstan and around the world. I wanted to

address the cancer issue by providing a better data infrastructure for healthcare organizations and cancer research. Furthermore, there was even the opportunity to obtain data from other diseases, but I decided to focus on cancer data due to the time constraints of this project. Thus, I was able to get cancer data with certain parameters, which were discussed with experts in the field of cancer. Each parameter of the obtained data was discussed with the thesis supervisors and other specialists in the field of medicine so that the data can bring value for research.

This research uses data from the Electronic Register of Cancer Patients (EROB) Kazakhstan. The request to the Ministry of Health consisted of 4 types of cancer: breast cancer, pancreatic cancer, glioblastoma and sarcoma that were diagnosed between 2013 and 2018. Although the initial request to the Ministry of Health consisted of these 4 types of diseases, the request returned only two types of diseases diagnosed with pancreatic and breast cancer. This happened mainly due to that sarcoma and glioblastoma are the International Classification of Diseases for Oncology (ICD-O-3) morphological type names [2], and the codes were needed to be specified in accordance to International Classification of Diseases (ICD-10) Code [3]. For designing the prototype version of FAIR Data Point all real data in Russian are translated into English using ICD-O-3 codes and the prototype will be presented locally without uploading to the internet for security purposes.

3.5 Implementation

To deploy the FAIR Data Point, the FDP code provided on GitHub was used (FAIRDataTeam, 2020). Data on cancer was used to carry out this study. The data were provided by the Ministry of Health and not intended for distribution.

[2] used in tumour or cancer registries for coding the site (topography) and the histology (morphology) of neoplasms (WHO, 2013)

[3] The foundation for the identification of health trends and statistics globally, and the international standard for reporting diseases and health conditions (WHO, 2004)

3.6 Obstacles

The originally obtained datasets from the MOH were stored in Russian. Columns which contain description and specific parameters of diseases translated from Russian into English using the International Classification of Diseases for Oncology (ICD-O-3) and International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). Personal history of patients was manually translated into English. This caused some difficulties during the translation process since the volume of the datasets was not small.

Chapter Four

Digital Health policy of Kazakhstan

Kazakhstan is one of the emerging economies which is aimed at becoming a member of the Organization for Economic Cooperation and Development (OECD) by 2050. In 2012, Kazakhstan has begun on a reform and investment program aimed at bringing the country to the list of the most developed countries in the world. This strategic goal, established by the 'Kazakhstan-2050' Strategy set a new political course for the further development of the country within the sectors of social and economic development (Nazarbayev, 2012). Likewise, the Strategy set new goals for the healthcare field and served as fundamental for later health-related programs. Currently, Kazakhstan is working on improving primary healthcare, embedding obligatory social health insurance providing accessibility and better quality to health services, and harmonisation of health data infrastructure. However, in comparison to the OECD and post-Soviet states, key health indicators in Kazakhstan are still at the lower stage (Obermann et al., 2016). Thus, reforming the healthcare field and addressing a number of health-related issues still require work.

4.1 Fundamental documents regulating access to data

Access to data/information, including public access to open data, are regulated by three fundamental documents. First, by the Constitution of the country which guarantees the right of citizens to access state data ("The Constitution of Kazakhstan", 1998). The second one, the data privacy law that is aimed to protect an unauthorised publication of personal data

("On Personal Data", 2013). The last one, by the Law on Access to Information that explicitly requires all state bodies to publish data sets in special public depositories in a proactive manner ("On Access to Information", 2015), and regulates the publishing of governmental data in special depositories (Kassen, 2019). These documents provide political, technological, and administrative regulation for the open data movement in Kazakhstan (Kassen, 2019).

4.2 What is the policy (legislation) and origin of the policy?

One of the fundamental documents for digital health in Kazakhstan is Strategy "Kazakhstan - 2050" which was issued by the first President of Kazakhstan in 2012. It was aimed at building unique standards of medical services, improving material and technical equipment of medical organisations, and focusing more on informational support work with the population. Applying smart medicine, remote diagnostic and electronic medicine introduced as a solution for such a geographically large country (Nazarbayev, 2012).

As a response for the Strategy-2050, in the beginning of 2013 the Program 'Informational Kazakhstan - 2020' was launched with various objectives in different sectors, including healthcare. The Program set new goals for the entire healthcare field, and included the term e-health / digital health and determined approaches for further digitisation of healthcare. It states that including ICT into the healthcare system will help to grow the quality of medicine and its services, as technology enables remote monitoring of patients, a better spread of information among patients, access to medicines ("IK", 2013). The main goals of this Program were to develop electronic medical cards of patients, connect all healthcare organisations to the unified healthcare network with 100 percent share, integrate all information systems of healthcare organisations into a single integration platform, increase computer literacy among medical staff and provide all medical employees with computers by 2020 ("IK", 2013). Thus, Information Kazakhstan 2020 was significant for the further development of e-health Kazakhstan, and there was a necessity to rethink the entire Concept of e-health at that time (MoH

Kazakhstan, 2013).

Apart from that, in 2016 Kazakhstan started the step-by-step implementation of OECD standards in the health sector by introducing National Health Program 2016–2019 Densaulyk ("Densaulyk State Program", 2016). One of the main goals of this program is to strengthen primary health care implementing public health policies for the prevention and treatment of diseases at the primary care level and create a more effective and financially sustainable health system, and introduce a compulsory social health insurance (Birtanov, 2016). It was also aimed to address some problems related to the emergence of new institutional structures and new functions, which in turn implies an increase in information needs, under which existing ICT tools (information systems and databases) should be adapted, and appropriate data collection processes and regulatory legal acts developed.

In December 2019, another e-health development program for 2020–2025 was introduced in Kazakhstan ("State Health Development Program for 2020-2025", 2019). Ensuring high-quality and affordable healthcare through the formation of a commitment to a healthy lifestyle among the population and the development of public health services; improving the quality of medical care; sustainable health system development ("State Health Development Program for 2020-2025", 2019). Along with this, the integration of health information systems and information systems medical organizations remain the main goal of the digital healthcare space. Thus, delivering high quality healthcare services is requires better data infrastructure which is still on the development stage.

In addition to all of the above, personalized medicine is another goal of the new e-health program ("State Health Development Program for 2020-2025", 2019). According to the e-health Program (2019) it will allow to identify and predict the course of the disease at the preclinical stage, carry out preventive measures, thereby reducing the cost of treatment and rehabilitation of preventable diseases. Thus, for this it is extremely important to have an appropriate data infrastructure. Personalized medicine uses AI, which requires interoperable data, and the more patient data available, the better and more accurate the treatment. Thus, accessing and sharing patient data internationally can better address this issue. For this, in Kazakhstan it is necessary to adopt regulations on e-health, the preservation and exchange of

data in the country and the use of data abroad.

4.2.1 Regulatory frameworks for digital health in Kazakhstan

Up to this point, Kazakhstan did not have a legislative framework providing legal regulation of the digitization of healthcare ("State Health Development Program for 2020-2025", 2019). Ensuring legal regulation is one of the goals of the E-Health Program for 2020-2025. Moreover, neither electronic health passports nor e-health is in any way reflected in the current Kazakhstan legislation (RCRZ, 2018). Therefore, for the smooth and stable development of e-health in Kazakhstan it is necessary to launch regulations on e-Health. Under the new e-Health Program, legislation must be developed to guarantee safety of information of patients, who is responsible for the information that they enter into the system, and for the leakage of personal data of patients.

The work on the regulatory framework for healthcare digitalization includes: issues of access to data, storage, privacy protection, quality assurance of the technologies and software products to be used ("State Health Development Program for 2020-2025", 2019). It also includes tools and standards for collecting and exchanging health data ("State Health Development Program for 2020-2025", 2019).

The implementation of the regulatory framework is carried out jointly with international organizations to coordinate and ensure compliance with the international regulatory frameworks ("State Health Development Program for 2020-2025", 2019). For that, the commitments made in the field of healthcare have been taken into account, promoting Kazakhstan's main initiatives abroad (promoting the Astana declaration [1]), as well as ensuring the transfer of knowledge and new technologies to the health sector through gaining international experience ("State Health Development Program for 2020-2025", 2019).

[1]The Astana Declaration it is a new declaration that defines the course of development of primary health care, taking into account the challenges and opportunities of the current time

4.2.2 Policies on cancer

According to the Comprehensive Cancer Control Plan for 2018-2022 the Step 59 states that Kazakhstan should expand the range of clinical trials, creating clinical trial registries that might provide stakeholders with access to the main results of clinical trials. This may positively affect improvement of public health and medical decision making. The participation of specialists in international clinical trials and medical research would be beneficial for the country to apply the latest international scientific and technological developments for the diagnosis and treatment of cancer patients (MoH & KazSRIOR, 2018). In this regard, the Comprehensive Plan refers to World Health Organisation's (WHO) idea on creating a database of clinical trials, and about the need to create international standards regarding the data of controlled studies, including technical aspects. Sharing clinical trial data is a moral obligation of data owners or researchers to share data responsibly (Alfonso et al., 2017). According to Alfonso et al. (2017) creation of a comprehensive and accessible clinical trial database - the reuse of data from at-risk patients participating in the study is the rationale for this global effort. Thus, this Comprehensive Cancer Control Plan shows that Kazakhstan is ready to participate in global efforts on creating a database of clinical trials for cancer and to contribute to the improvement of science worldwide as well as within Kazakhstan.

4.3 Implementing Digital Health in Kazakhstan

At the level of clinical / population and research data, the opening up medical data, sharing and linking of large amounts of datasets of medical data allows to semantically link and enrich data on symptoms, diseases, diagnosis, treatment methods and prescriptions, offering the potential to improve care for people and population groups as well as more effective semantic access to the evidence base (Kostkova et al., 2016). In this regard, Kazakhstan is actively working towards reaching semantic interoperability of medical data, and this can be seen from a number of health reforms in recent years. The new e-Health Program 2020–2025 (2019) has set targets for ensuring data compatibility, and this shows that there is still room for improvement in the

health data infrastructure.

The interoperability platform is one of the main components of the new e-health architecture in Kazakhstan towards reaching semantic interoperability, designed to provide the technological infrastructure for the development of e-health. In 2013, as part of a World Bank project, Swiss Tropical and Public Health Institute evaluated the information systems of the MOH of Kazakhstan (MoH Kazakhstan, 2013). This evaluation helped to develop the Concept of e-Health which provided the abandonment of obsolete technologies, revision of goals and priorities, and rejection of monopolization within healthcare. According to the Healthcare Concept 1.0 (2013) the Ministry's main focus was shifted from the collection of analytical data to the formation of an integrated information environment. The main goal of this Concept 1.0 is the formation of an integrated information environment that ensures the involvement and access to the necessary information of all the main actors of the healthcare system (MoH Kazakhstan, 2013). By the Concept, the digitalization of the leading clinical processes at the regional and local levels should be provided by local executive bodies. The Ministry, in turn, is developing national systems aimed at financing and management issues, as well as providing mechanisms for the exchange of medical information through the creation of an EHR.

On December 18, 2015, a contract for the supply of the Health Interoperability Platform was signed with Croatian Ericsson Nicola Tesla as part of the World Bank Project. The completion and commissioning of the pilot operation are planned in mid-2018. However, the implementation of integrated information environment has faced a number of unexpected challenges, and designing the platform took more time than expected (Abishev, 2018), and currently the implementation of the Platform has been finished, but the integration of information systems is still necessary for achieving interoperability of medical data.

4.3.1 Electronic Health Records

One of the key points of digital health programs launched in Kazakhstan is Electronic Health Passports (Electronic Health Records) which is aimed at storing, collecting and analysing patients data in one place at the level of body authorities. Electronic health passports were

launched in 2019 (Birtanov, 2019) which was implemented as part of the Densaulyk 2016-2019 state program, and was also synchronized with the Digital Kazakhstan state program, in fulfillment of the order of the Head of State from 2018 to switch to paperless maintenance medical records, medical organizations ("Densaulyk State Program", 2016).

4.4 The acceptance of FAIR Principles by the Theory of Planned Behavior

The theory addresses the attitudes towards acceptance and usability of FAIR Guiding Principles by users/healthcare workers. According to Abishev (2018), the medical staff of the country is very conservative population, therefore the digitization of healthcare is very difficult and problematic. Systematic approaches, training and broad outreach are necessary for a smooth transition. Thus, the attitude of healthcare workers to the FAIR Principles can be complex as it is a new technology for them, but by understanding the value that it can bring to them and society, they might be more engaged and interested in. Since the ideas of FAIR Data Principles are data being distributed and thus, there will be many data points that need to be managed by data competence centres for which health workers also play a big role, it is necessary to get their interest in the usage of technology. On the other hand, Kazakhstan is a country with a centralized government structure, which is characterized by top-down decision-making power and a governance system (Liebert et al., 2013). Thus, the interest of people in high positions is also an important component towards performing such behavior.

As for the subjective norm, data policies have been studied, and it is clear that Kazakhstan is on its way to achieving full digitization in healthcare. This can be seen from a number of health programs that have been launched in recent years. Data management legislation has been verified and there are data exchange laws and a law on personal data, however there is no policy regarding the use of data for research. This means that Kazakhstan still needs to move forward in this direction.

At the technical level, databases are still fragmented, as they are stored in heterogeneous

information systems that are not connected to each other. One of the important achievements that was achieved was that Kazakhstan switched to a paperless environment. Thus, data is stored in electronic database repositories, which means that the FAIR principles can be applied to solve this fragmentation problem along with using the information value of the data.

4.5 Internship outcomes

At the moment, the legislation does not provide a responsible institution for checking private medical information systems for compliance with the standards of the Republic of Kazakhstan. Such a body should have appropriate competences in the field of healthcare and IT-technologies, as well as the authority for such activities. Also, the body responsible for training in the field of standardization is not provided for - these two factors act as a mechanism for postponing and hindering the implementation of standards, and therefore, the overall efforts to achieve interoperability of healthcare information systems in Kazakhstan has been hindered.

4.6 Challenges of the digitization

In this chapter, the main challenges of the health digitization will be discussed.

4.6.1 Data infrastructure issues

Data infrastructure could be organised in various approaches, including centralised and federated architecture as shown in Fig.4.1[2]

[2]https://academic.oup.com/view-large/figure/137283004/bbz044f2.tif

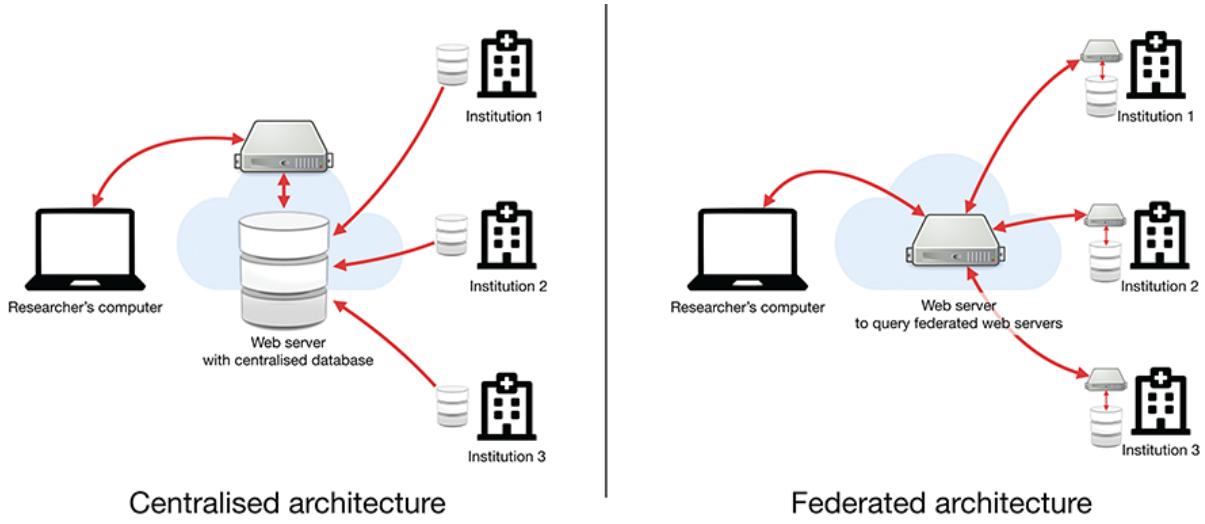

Figure 4.1 Centralised vs federated architecture (Vesteghem et al., 2019)

In the centralized architecture, every organization must upload their data to the data storage centre. In contrast, in the federated architecture, data doesn't leave the respective organization and stay with them. However, each organization must design an interface to make data discoverable, and it does not need to be accessible (Vesteghem et al., 2019). Kazakhstan's healthcare industry is facing significant progress in adopting modern information technologies, including the creation of several portals, improving the provision of computer equipment. However, developed and implemented applications are aimed only at solving specific issues of financing and managing the healthcare system. The existing databases are fragmented, implying that health data is not interoperable on the national level. The most significant tasks of the MoH and its structural divisions were automated by implementing 22 information systems at the national level. They are designed to collect statistical information and provide funding for the health sector (Birtanov, 2017). Information systems are only connected on the level of the Ministry of Healthcare, and data are collected in the context of individual cases, diagnoses or levels of care. Under these conditions, the data remains locally, i.e. fragmented at the level of one organization, and there remains the need to re-record medical data: on paper, in a local MIS, in the Information Systems of the Ministry of Health. (On people's health and healthcare system, September 2018). Thus, fragmented data is insufficient to support clinical decisions at the local/patient level, and impedes the integration and continuity of various levels and the health service. In this regard, semantic interoperability is an essential factor in obtaining

the advantages from electronic health record systems to strengthen the quality and safety of patient care, clinical research, public health, and health service management (European Union, 2013). Therefore, achieving semantic interoperability might give healthcare providers better access to all the necessary patient data on the national level and ensure timely and safe patient care. The review by the OECD (2018) states that for the conducting overview of healthcare, they requested health data for data analysis, the data that served as the basis for the review was incomplete and contradictory, and there was doubt on their quality and reliability. The data do not meet international standards and lags behind the OECD countries in terms of the effectiveness of using available data to improve the health system. Along with that, the exchange of information between medical institutions at different levels is very limited and is a significant obstacle to improving the integration and coordination of treatment activities ("OECD Reviews of Health Systems", 2018).

4.6.2 Policy issues

One of the next significant issues is a clear gap between policy initiatives and measuring the effects they have (Obermann et al., 2016). Failure to measure progress towards policy goals will seriously affect the visualization and effective communication of the meaning and consequences of proposed reforms (Obermann et al., 2016). Kazakhstan suffers from a lack of independent monitoring and evaluation of reforms. There are several comprehensive strategies, but monitoring of the health-care programs is difficult because there is no institution which might produce a more independent assessment of the health-care programs (Birtanov, 2016). Thus, the government has launched a number of programs, to address urgent health related-issues, however there have not been done any independent evaluation on the efficiency of those reforms. Although reforms are widespread and often implemented, very little attention has been given to the assessment of real progress in their effective implementation ("OECD Reviews of Health Systems", 2018). There is very little information about reform problems and progress towards achieving the intended results. Although there are clear systems for monitoring the "achievements" of implementation of various reforms, the number of people trained and the

number of institutions participated ("OECD Reviews of Health Systems", 2018). Reforms have not been analysed by independent third parties and the impact of reforms have not been systematically assessed ("OECD Reviews of Health Systems", 2018). There was not enough existing data to measure progress as reforms developed ("OECD Reviews of Health Systems", 2018). Thus, data remains difficult to interpret and analyse.

4.6.3 Open Science policies

Open science is a growing movement. It is expanding throughout the world, and more and more institutions are giving priority to its promotion and acceptance worldwide (Mons et al., 2017). While most countries are taking steps towards open science and the possibility of reusing research data, Kazakhstan still lacks legislation on open science and the possibility of reusing research data. Even though Kazakhstan supports and promotes Open Data movement through the publication of state datasets in depositories (Kassen, 2017), there is no law on open science and the possibility of reusing research data.

4.7 Conclusion

For collecting digital health data it is necessary to build an infrastructure for data exchange and advanced use of data ("State Health Development Program for 2020-2025", 2019). Implying the formation and presentation of high-quality data for all levels of the healthcare system is applicable with the FAIR Principles. Thus, applying FAIR technology for the healthcare in Kazakhstan is relevant and feasible in terms of the objectives set in the e-health Program fro 2020-2025 years. For this, it is necessary to introduce e-health rules in Kazakhstan, but before that, with a bottom-up approach, policy entrepreneurs, in this regard, Ambassadors IN and VODAN IN may address this issue and promote FAIR as a solution.

Chapter Five

FAIR-ness of digital health data of Kazakhstan

To integrate information systems and build a single data repository with one entry point, an integration platform is being developed together with the World Bank (The World Bank, 2017). The architecture of the platform is given in Fig.5.2. The platform was supposed to go into pilot mode at the end 2017 [1]. However, the project is still expected to launch.

In this chapter, existing healthcare information systems and digital health data of Kazakhstan will be discussed and assessed according to the FAIR guidelines: findability, accessability, interoperability and reusability. There are 22 online information systems functioning in Kazakhstan which designed to collect all health-related data [2] for statistical and analytical purposes, and provide funding for the health sector [3]. Apart from that, there is the Open Data Portal (data.egov.kz) which enables data sharing of different datasets on the state portal and opens up opportunities for stakeholders to participate in governmental activities and contribute to transparency of state related services. It is one of the components of the Open Government, created to provide interested citizens with access to different data sets originating from the state bodies of Kazakhstan.

[1] https://www.inform.kz/en/healthcare-ministry-ibm-to-introduce-ai-to-kazakhstan-s-oncology_a3044271

[2] http://ezdrav.kz/posetitelyam/kratkoe-opisanie-informatsionnykh-sistem

[3] https://www.inform.kz/en/healthcare-ministry-ibm-to-introduce-ai-to-kazakhstan-s-oncology_a3044271

5.1 Information systems of Kazakhstan

The description of information systems are given on the e-Health website [4], since the databases are only accessible by healthcare organisations, information systems will be assessed by the first facet of FAIR and cannot be assessed for other facets, such as accessability, interoperability and reusability principles.

The number of information systems available in Kazakhstan is given on the website with their description, links and main goals. However, if we get to the links themselves, then there will be no (meta) data. We can assume that this happens for security purposes, and in this study, the list of information systems will be provided:

1. Register of pregnant and women of childbearing age: to monitor indicators of health status of pregnant women and women of childbearing age. Users of the system are ambulatory-level medical organizations (https://www.eisz.kz).

2. Register of acute coronary syndrome: it is aimed for registration of patients with the cardiovascular diseases, and for further monitoring of patients, detection and registration of violations of diagnostic algorithms and treatment protocols, the formation of statistical and analytical reporting for operational management decisions (https://www.eisz.kz).

3. Information system "Electronic register of inpatients": it is intended to ensure the speed of data collection for the formation of a unified, centralized, information database on treated cases, containing all the information necessary to finance inpatient and inpatient care for medical services rendered (https://ersb.eisz.kz/).

4. Information System "Medical Services Quality Management System": It is intended for information support of the functions of management, structural divisions and individual specialists of governing bodies and healthcare organizations in planning, accounting, analysis and management (organization), as well as for monitoring the quality of medical services provided for various operations for governing bodies (https://sukmu.eisz.kz).

5. Information System "Drug Support": for prescribing medicines by medical stuff, and the interaction of medical organizations providing outpatient care and pharmacy organizations

[4] http://ezdrav.kz/posetitelyam/kratkoe-opisanie-informatsionnykh-sistem

providing free drugs; accounting of the provided free medicines and medical products in pharmacy organizations; the ability to use the data entered into the System for further statistical and analytical purposes (https://islo.eisz.kz/).

6. Population Register: It is intended for the formation of a single centralized information database on the actual number of people attached to each health organization providing primary health care (https://www.eisz.kz).

7. Information System Electronic Register of Cancer Patients (EROB): The system is intended for keeping records of patients of oncological dispensaries, maintaining and processing information on the volume of medical care provided to patients with cancer diseases. EROB consists of a financial and production unit. The production part is system for entering data from cancer patients throughout the country. Also in this block, the acquisition of expensive chemotherapy drugs, the consumption and planning of medicines is monitored. Accounting for the movement of cancer patients in the hospital and on an outpatient basis. The financial unit covers all the payment documents of dispensaries providing medical care to patients (https://www.erob.eisz.kz).

7. Information system "Electronic register of dispensary patients": the database stores information of patients, who are under dispensary registration, allows observation of patients and determine the need for free drug supply at the outpatient level (https://www.eisz.kz). It has 7 main subsystems.

8. Subsystem "National register of diabetes" of the information system "Electronic register of dispensary patients": The system is designed to automate the process of collecting and processing data from patients with diabetes for the first time taken into account, deregistered and registered in the dispensary. The register allows to obtain reliable information about the incidence, mortality of patients with diabetes, the effectiveness of treatment and preventive measures(https://www.eisz.kz).

9. Subsystem "National register of tuberculosis patients" of the information system "Electronic register of dispensary patients": the system is designed to automate the process of collecting and processing data on tuberculosis patients first recorded, deregistered and registered in the dispensary, as well as information about the medical examination, a map of the patient

with tuberculosis, information on laboratory tests (https://www.eisz.kz.).

10. Subsystem "Register for patients with chronic kidney failure" of the information system "Electronic register of dispensary patients": The system is designed to automate the process of collecting and processing data on dispensary observation of patients who need hemodialysis, and patients with an allotransplanted kidney who need supportive immunosuppressive therapy throughout their lives (https://www.eisz.kz).

11. Subsystem "Electronic register of drug addiction patients" of the information system "Electronic register of dispensary patients": the system is designed to automate the process of collecting and processing data of drug addicted patients first taken to the register, deregistered and registered in the dispensary. The register allows to receive reliable information about the incidence, disability, mortality of drug addicts, the effectiveness of treatment and preventive measures (https://www.eisz.kz).

12. Subsystem "Register of patients with viral hepatitis" of the information system "Electronic register of dispensary patients": The system is designed to monitor the effectiveness of the detection, treatment and prevention of viral liver diseases in the Republic of Kazakhstan (https://www.eisz.kz).

13. Subsystem "Register of mental patients" of the information system "Electronic register of dispensary patients": The system is designed to automate the process of collecting and processing data from patients with mental illness registered/deregistered in the dispensary. The register allows to receive reliable information about the incidence, disability, mortality of patients with mental illness, the effectiveness of treatment and preventive measures (https://www.eisz.kz).

15. Hospitalization bureau: it is intended to provide information on planned hospitalization of patients in the country with information about free beds in hospitals; about patients on the waiting list for planned hospitalization; about patients hospitalized in hospitals or who are denied hospitalization.

16. Information System "Resource Management System" of the Ministry of Health of the Republic of Kazakhstan: the portal is designed to automate the tasks performed by specialists of health organizations in planning, monitoring, accounting and analysis of the activities of

health organizations in terms of managing material and human resources (www.eisz.kz).

17. Information system "Management system for drug support, drug monitoring": a unified database of medicines registered and approved for medical use in the Republic of Kazakhstan (https://sulo.eisz.kz).

18. Information System "Medical Technology Management System": data on medical equipment located in healthcare organizations, automation of the process of generating an application for the purchase of medical equipment, as well as further monitoring of its use in healthcare organizations (https://sumt.eisz.kz).

19. Information System "Unified Payment System": it was designed to automate the process of payment for medical care provided at the stage of primary health care and consultative and diagnostic care. An additional purpose of the System is the implementation of mutual settlements between medical organizations.

20. "Electronic registry of services at the outpatient level" of the information system "Portal Outpatient care": the system is designed to generate personalized data about the patient, information about his visits to outpatient organizations. The object of automation for the System is the field of activity of medical healthcare organizations providing outpatient care (https://www.eisz.kz).

21. An additional Component to the primary health care: the portal aimed at decreasing maternal and child mortality rates, early detection of cancer diseases of visual localization, tuberculosis control, ensuring the continuity of medical care between levels: primary and inpatient, development of socially oriented medicine (https://skpn.eisz.kz).

22. Information system of medical organizations "Clinic".

5.2 Analysing information systems

All digital health related data are stored in the above given information systems and are used to collect data for statistical, analytical and decision-making processes. Despite the fact that they are aimed at storing health data of patients, diseases, drugs and allocation of health facilities, they are not yet integrated with each other and do not communicate (“Results of

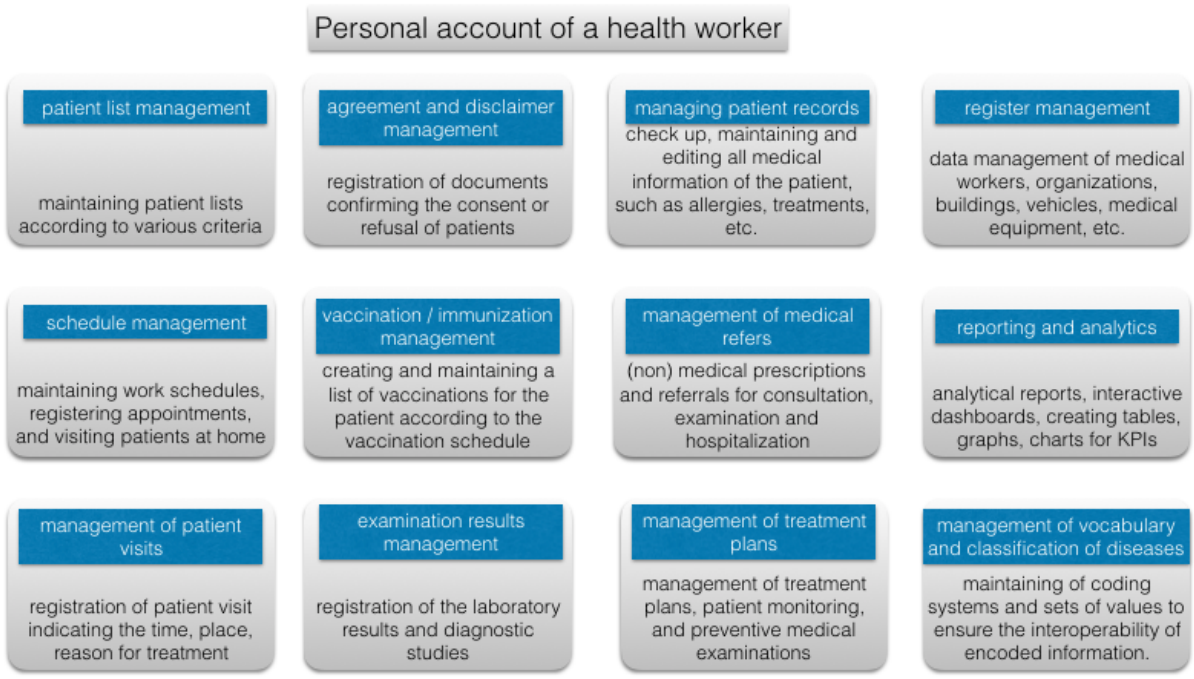

Figure 5.1 Functionality of the personal account of a health worker [5]

Densaulyq Program", 2020). Although the functionalities of health workers (Figure 5.1) are provided on the website, there is no data that could provide certain database parameters. By the first aspect of FAIR, information systems are accessible only through search services and are presented on the state healthcare website in Russian, and they have no global unique and persistent identifier. All health related data are only accessible by healthcare organisations and medical staff. Thus, an accurate analysis of information systems and their features cannot be made in this study. Therefore, an assessment of healthcare data which is given on the state portal will be done in the next section of this chapter.

5.3 Health related data available on the Open Data Portal of Kazakhstan

An essential aspect of implementing FAIR Guidelines is to measure the level of FAIRness using specific metrics that quantify FAIRness (Wilkinson et al., 2018). For that, the open data portal that shares health-related data will be evaluated on the basis of FAIR data using the FAIR

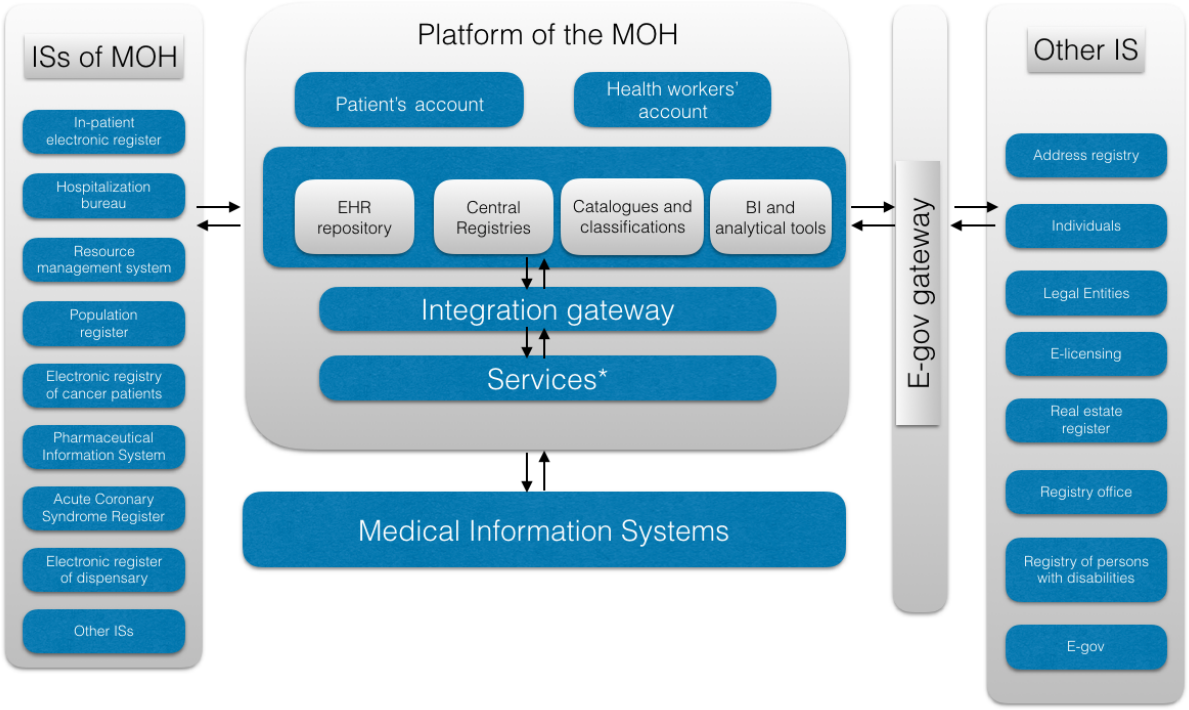

Figure 5.2 Interoperability Platform Architecture [6]

Evaluation Services developed by Wilkinson[7].

There are more than 3 thousand datasets available on the Open Data Portal (data.egov.kz). Most available health related datasets provide only statistics, such as a number of medical organisations, medical personal available in different regions and etc. There is no any data that can be used for research purposes. If the user of the website wants to receive a certain dataset which is not provided on the portal, he / she can send a request which can be done only by citizens / residents of Kazakhstan who has an electronic digital signature (EDS). The main issue of the given portal is that no one knows what kind of data he / she can request, because there is no metadata describing types of data available for request.

If we move to the Healthcare section, the first thing we notice is that the data sets are not classified, most of them that are in the healthcare category are not related to healthcare at all. Although metadata is accessible and machine-readable, the data presented does not add value for this study. Any website user who wants to get online access to particular data repository must provide an API key, and this option is available only for citizens / residents

[7]https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/!/

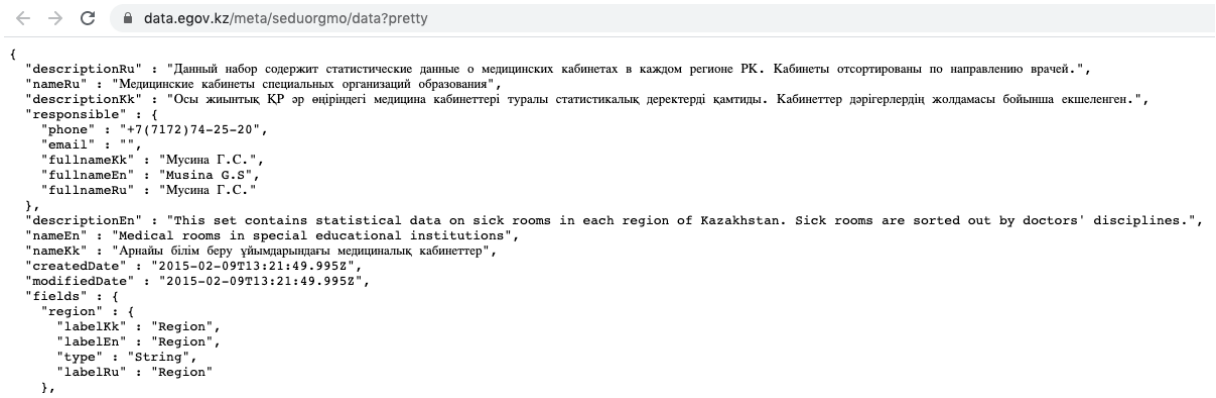

Figure 5.3 How metadata of the state portal looks like

of the country who have Kazakhstani digital signatures. However, it is possible to download a small piece of data which makes possible to see the parameters of database.

5.4 Assessment of the FAIRness of health-related data

To assess the FAIRness of the above given digital resource the FAIR Maturity Evaluation Service (https://fairsharing.github.io/) is used for assessment according to FAIR Data Principles. It allows digital resources to be assessed objectively and transparently.

5.4.1 Provenance of the policy

Opening up state depositories to public is caused by the Law on Access to Information ("On Access to Information", 2015). This was made to make governmental services transparent to public and involve citizens to participate in governmental activities. Other than that there is no data, in particular, digital health data available for various stakeholders, including academia and independent researchers.

5.5 F - Findability

In Kazakhstan, the only health-related digital data available through the Internet is the open data portal. There are more than three thousand data sets presented. However, it turns out that data sets provide generic statistics and do not store any medical data, except for the

information related to generic healthcare information, such as the number of diseases presented by the year, number of cases of different illnesses, number of medical rooms in various regions and specialists presented in various cities. All datasets presented on the website has machine-readable metadata, and presented in three languages, including Kazakh, Russian and English. However, the datasets are accessible under presence of EDS / or in case users do not have EDS they can download data from the website with the limited row (100) numbers.

5.5.1 An analysis of F1 - F4

Persistent, globally unique identifiers, resolvable on the Web are one of the main aspects of data being in a FAIR manner. The assessment of the open data portal was carried out in accordance with the principles of F1-F4, 4 out of 8 points were successful, and can be found at this link: https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/!/evaluations/3345.

According to the first principle of F, in particular F1, meta(data) needs to be assigned with globally unique and persistent identifiers. Globally unique and persistent identifiers allows meta(data) to be discovered globally (Wilkinson et al., 2016). The assessment of the Open Data Portal was done using the FAIR Maturity Evaluation Service. It found that a unique identifier is applied for metadata resource with type of 'uri', a number of characters which identifies a particular resource without any ambiguity (Wikipedia, 2020). However, the unique identifier of the data/metadata resource is unlikely to be persistent. While testing known URL persistence schemas (purl, oclc, fdlp, purlz, w3id, ark) the metadata GUID does not conform with any known permanent-URL system. The assessment was unable to locate the data identifier in the metadata using any property/predicate reserved for this purpose. The link for the data has not found as it is located in another place as shown in Fig.5.3.

The second principle of facet F is that metadata is described in a rich manner so that human and machine can understand what stores the exact dataset (Wilkinson et al., 2016). The reason following this principle is that someone should be able to find data based on the knowledge provided by their metadata, even though the data identifier is not provided. Therefore, compliance with F2 helps people to locate the data, and increase the further re-use

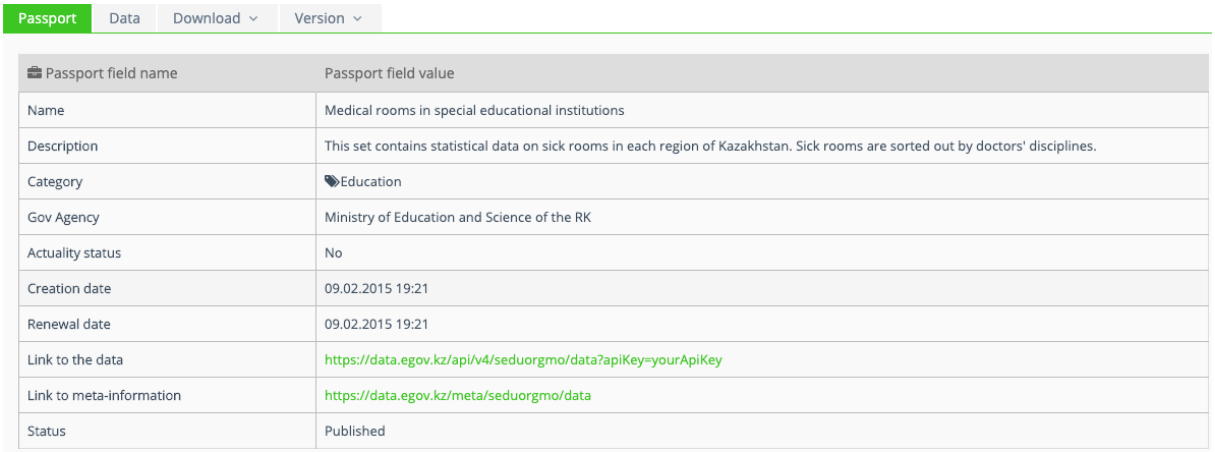

Figure 5.4 How data and metadata are presented on the state portal

and citations of the same data set (Wilkinson et al., 2016). In this regard, metadata of the Open Data Portal was also tested to see whether a machine is able to find structured metadata. This could be RDF, embedded json, json-ld, or content-negotiated structured metadata such as RDF Turtle. It showed that linked-data style structured metadata was found on the website. Metadata was also tested whether a machine was able to find 'grounded' metadata. For instance, metadata terms that are in a resolvable namespace, where resolution leads to a definition of the meaning of the term. Examples include JSON-LD, embedded schema, or any form of RDF. Thus, the F2 aspect works perfectly on the website giving more chances to be explored by other interested individuals.

F3 principle states that metadata consists of the identifier of the data set (Wilkinson et al., 2016). Both metadata and data are located in separately, mentioning a data set's globally unique and persistent identifier in the metadata[8]. This was tested on the state portal to see if the metadata contains the unique identifier to the data. The Evaluator was unable to locate the data identifier in the metadata. Also, the test was conducted to see if the metadata contains the unique identifier to the metadata itself. The identifier https://data.egov.kz/meta/seduorgmo/data?pretty was not found in either the structured or unstructured content from resolving that identifier. Although metadata does not store the dataset identifier, the link for the dataset is stored on the page where the main data description is presented.

[8]https://www.go-fair.org/fair-principles/

Rich metadata and identifiers will not guarantee 'findability' of data through the internet. Despite the fact that data resources are perfectly operating F1-F3 principles, data can still not be discovered if they do not use indexing or any other methods that will help to quickly find data through search engines. Principle F4 states that data can be found through the search engine (Wilkinson et al., 2016). In this regard, indexing helps to easily find data across the internet and it works for almost all ordinary data, although scholarly research data requires more better approach for indexing (The FAIR Maturity Evaluation Service). The FAIR Maturity Evaluation Service uses Microsoft Bing to test portal whether a machine is able to discover the resource by search. It was unable to discover the metadata record by search in Bing. Thus, this could be potential problem for data being findable through the internet as data might not be seen by stakeholders only because they do not know that data exists on the internet.

According to the FAIR Maturity Evaluation Services, the open data portal has 4 out of 8 successful points. Principle F1 works partially, F2 works fully, and F3 and F4 principles do not work. The only presence of partial F1 and F2 do not ensure that data sets will be discovered through the internet. Therefore, the probability of finding the necessary data sets from the given portal is low, and the chances to find data from state bodies is higher on the state portal itself.

5.5.2 Challenges

Although metadata (F2) is available on the state portal for all existing datasets, other facets of FAIR are not applied to meta(data) or are partially applied. This makes it difficult to discover meta(data) globally. Another problem is that, although various data sets are provided on the governmental website, they contain general statistics, and most of the data provided are not up to date, not properly classified. Apart from all of the above given, the portal provides the opportunity to obtain a specific dataset that is not presented on the portal. However, it does not provide a list of (meta)data available for request.

5.6 A - Accessibility

The second aspect is "A" is not necessarily mean open, it means that data is open under specific conditions. The same portal will be assessed in this chapter. The open data portal is initially intended for citizens and provides full functionality (a request for a specific data set, any request for updating, uploading data) of the website only for users who have EDS.

5.6.1 The current state of the art

The Open Data Portal is created for citizens of Kazakhstan, and it requires EDS for the authorization. Being authorized means that the user can have some privileges than usual quest on the website. However, it doesn't give lots of privileges, as data sets themselves only provide general statistics. The A facet was tested by the FAIR Maturity Evaluation Service and it is available through this link: https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/!/evaluations/3363

The Principle A1 states that meta(data) are retrievable by their identifier using a standardised communication protocol (Wilkinson et al., 2016). Thus, FAIR data retrieval should be gained without specific tools or communication methods. In terms of sensitive data, it is required to specify contact details of the data provider since fully mechanised protocol will not be secure for these kind of data. According to the A1.1 principle the protocol is open, free and universally implementable. For data to be maximally reused, the protocol should be with no-cost and open, therefore, globally implementable to promote data retrieval. In terms of the Open Data Portal, metadata may be retrieved by an open and free protocol. This evaluation of the second facet of FAIR passes InChI Keys, DOIs, Handles, and URLs. The identifier of the website is of type uri, which is resolvable by an open protocol. The evaluation was unable to locate the data identifier in the metadata using any (common) property/predicate reserved for this purpose. Data set is located in different link and may be retrieved by the presence of EDS. A1.2 principle states that the protocol allows for an authentication and authorisation where necessary (Wilkinson et al., 2016). The evaluation tested metadata for the ability to implement authentication and authorization in its resolution protocol. The GUID of the metadata is a uri, which is known to be allow authentication/authorization. It was also tested a discovered data

GUID for the ability to implement authentication and authorization in its resolution protocol. It also searches the metadata for the Dublin Core 'accessRights' property, which may point to a document describing the data access process. No data identifier was found in the metadata record. However, the data itself is in a different link.

A2 Principle states that metadata should be accessible even when the data is no longer available (Wilkinson et al., 2016). The evaluation was done to test if the metadata contains a persistence policy, explicitly identified by a persistence Policy key (in hashed data) or a predicate in Linked Data. It was unable to find a persistence policy using any approach. This means that metadata will not be discoverable for users when data itself is no longer available.

5.6.2 Challenges

Digital medical data of Kazakhstan is available only to medical organizations and citizens themselves to access their own data. The rest of the data for any other purposes are not findable through the internet. The data presented on the website do not add value to any research, as well as cancer research, since this website does not have research data.

5.7 Interoperability and Reusability

Various data repositories and datasets should be interoperable between each other in order to allow increasingly complex questions to be answered. In such a case, interoperability of data takes place in two levels, including syntactical and semantical. Through the introducing FDPs which produce findable, accessible, interoperable and reusable data owners are able to share their data and address this interoperability issues. The Open Data Portal is machine-actionable which makes data to be syntactically interoperable and reused. However, data quality, including the range of data sets provided is quite low. This means that syntactically data sets can operate with each other as well as reused, but data sets themselves do not make any sense as they do not provide any data that can be used for research purposes.

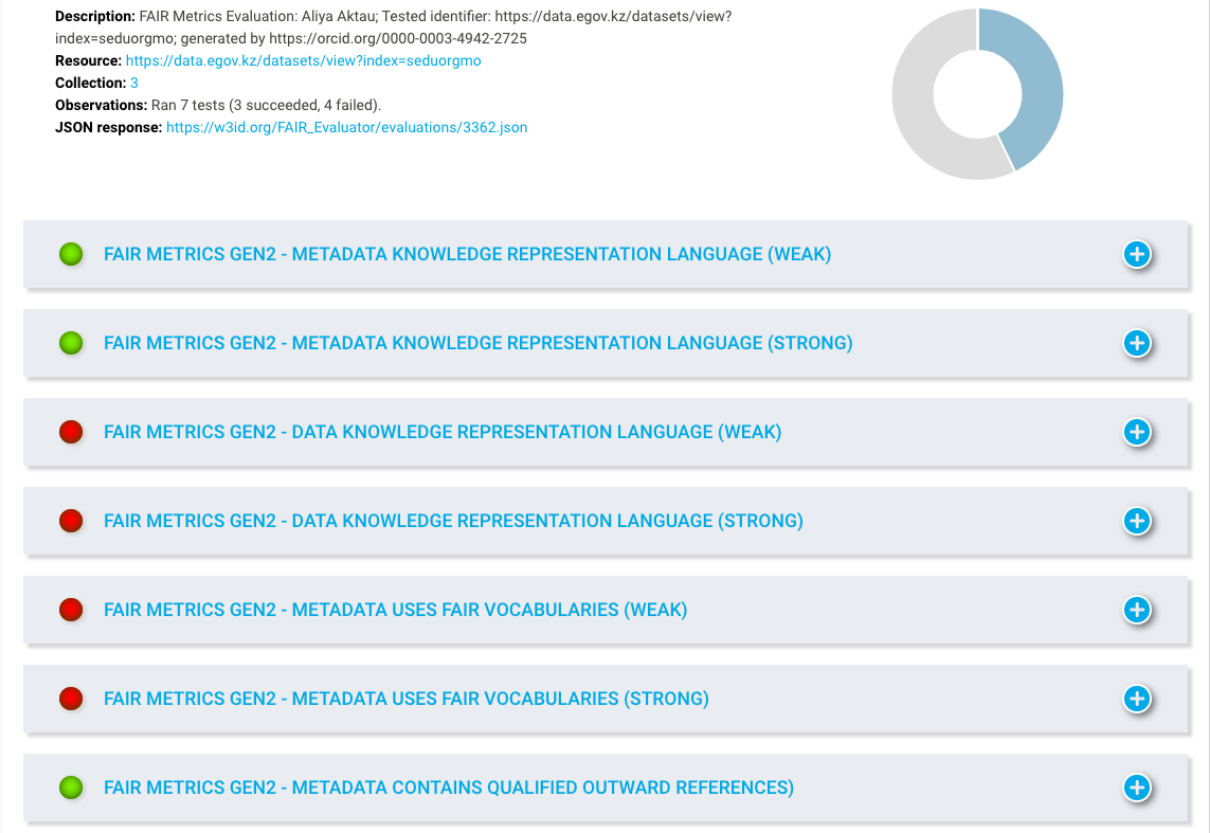

Figure 5.5 An evaluation of the state portal data on I aspect

5.7.1 An analysis of I + R

The I facet was assessed by the FAIR Evaluator, and 3 out of 7 points succeeded the test.

I1 Principle states that meta(data) use a formal, accessible, shared, and broadly applicable language for knowledge representation (Wilkinson et al., 2016). For better findability and interoperability of data sets, it is essential to use (1) commonly used controlled vocabularies, ontologies, thesauri (globally unique and persistent identifiers, see F1) and (2) a data model (a well-defined framework for (meta)data) [9].

The dataset tested if the metadata uses both weak and strong formal languages broadly applicable to knowledge representation as shown in Fig.5.5. This test accepts everything that can be represented as structured data and terms are semantically-grounded in ontologies (FAIR Evaluation Services). As a result of the test, linked data was found, thus it has passed successful assessment.

[9] https://www.go-fair.org/fair-principles/

The data set also tested if the data uses both weak and strong formal languages broadly applicable for knowledge representation. The Evaluator was unable to locate the data identifier in the metadata using any (common) property/predicate reserved for this purpose (FAIR Evaluation Services). Even though the test could not find any link to the data in the metadata, the link for the data is located on the general description level. However, the test was not able to evaluate this link.

The I2 facet states that meta(data) use vocabularies that follow the FAIR Principles (Wilkinson et al., 2016). The controllable vocabulary that used to describe datasets needs to be documented and well-regulated using globally unique and persistent identifiers. The documentation should be easily findable and accessible to anyone interested in using the data set. In such a case, using the FAIR Data Point might be one the examples of I2. The FAIR Evaluator tests if the linked data/metadata uses terms that resolve. 0 of the first 6 predicates discovered in the linked data could be resolved. The minimum to pass this test is 50 per cent, and it does not meet the requirement. Furthermore, the Evaluator tests if the linked data/metadata uses terms that resolve to linked (FAIR) data. 0 of the first 6 predicates discovered in the metadata resolved to Linked Data data. The minimum to pass this test is 50 per cent.

The I3 facet states that meta(data) include qualified references to other meta(data) (Wilkinson et al., 2016). To be more precise, specifying if one data set builds on another one, if supplementary data sets are needed to complete the data, or if the corresponding information is stored in another data set. The scientific links connecting the data sets need to be described. Besides, all data sets need to be properly cited, including their globally unique and persistent identifiers. The FAIR Evaluator tests if the metadata links outward to third-party resources, and it only tests metadata that can be represented as Linked Data. 6 of the 6 triples discovered in the linked metadata pointed to resources hosted elsewhere. This means that test successfully passed all necessary requirements.

5.8 R facet of FAIR

The R1 facet (meta)data are richly described with a plurality of accurate and relevant attributes (Wilkinson et al., 2016). R1 focuses on the ability of a machine or a human to determine if the data is valuable in a particular circumstance.

R1.1 (meta)data are released with a clear and accessible data usage license (Wilkinson et al., 2016). Although 'I' aspect covers the components of technical interoperability, R1.1 deals with legal interoperability, covering the aspect of what usage rights attached to data. The Evaluator tested if the linked data/metadata contains an explicit pointer to the license. This includes xhtml, dvia, dcterms, cc, data.gov.au, and Schema license predicates in linked data, and validates the value of those properties. According to the test no License property was found in the metadata. The second test verification point checks if the metadata contains an explicit pointer to the license. this point differs from the above one, as this one is 'weak' test, and uses a case-insensitive regular expression, and scan both key/value style metadata, as well as linked data metadata. According to test results,no License property was found in the metadata.

R1.2 (meta)data are associated with detailed provenance.

R1.3 (meta)data meet domain-relevant community standards.

5.9 Conclusion

Digital health data of Kazakhstan is only accessible for medical organisations, and citizens of the country can only access their own data on the Open Government portal. The health-related information systems are only described on the website of the MOH and are not accessible in terms of particular parameters. The data presented on the open data portal do not add value to this study, since only general health statistics are offered, which do not add any value to these hypotheses related to the area of cancer, due to the inaccessibility of cancer data.

Digital medical data of Kazakhstan is available only to medical organizations, and citizens of the country who can access their data on the Open Government portal. Health-related

information systems are described only on the website of the Ministry of Health and are not accessible in terms of specific parameters. The data presented on the open data portal do not add value to this study, since only general health statistics are offered, which do not add any to any research due to the inaccessibility of data.

Chapter Six

Comparing Cancer Data with Global Cancer Data

In this chapter, cancer data obtained from the Ministry of Health of Kazakhstan will be compared with international cancer databases to verify their alignment. The data which is necessary for an adequate test of hypotheses of the given research was taken from the Patient Record System, in particular, Electronic Register of Cancer Patients (EROB) Kazakhstan. The dataset includes information on patient status and disease phenotype. Patient status comprises, for instance, demographic information such as age at onset, gender, disease type, disease confirmation method, medication, and medication outcomes. The phenotype of the disease is characterized by morphology and topography. Morphology details the cellular structure of cancer, while topography determines its location. Usually, this data is collected by medical personnel and stored in electronic medical records (EHR) or in the context of research and clinical trials where EHR is defined as all patient medical data available in electronic format. The data sets do not allow the identification of individuals as any data which can refer to individuals' personal data were removed from the data sets.

6.1 Data description

The data sets received from the Ministry of Health consisted of two main files. The first file includes patients registered with a cancer diagnosis for the first time, and the second

file consists of patients diagnosed with secondary cancer, such as patients with relapse or a diagnosis of another type of cancer. Initially, the data was obtained in Russian language with 65534 and 65535 lines, respectively. The dataset was obtained using the diagnosis, morphology and topography of diseases in accordance with the standards of the World Health Organization, such as ICD10, ICD-O-3. For the translation of the above given parameters International Classification of Diseases for Oncology (ICD-O) was used for coding the site (topography) and the histology (morphology) of neoplasms (WHO | (ICD-O-3)) [1].

The datasets contain two types of diseases, such as pancreatic cancer and breast cancer with several cases of sarcoma[2]. The data was retrieved from the EROB with the above given illnesses during the years 2013-2018, and with the below parameters: as in Fig.6.1. with the columns: date of birth, gender, ethnicity, location (region) , date of diagnosis, duration of hospitalization, date of death (deregistration), reason for deregistration, as in Fig.6.2. with the columns: method of confirming the diagnosis, type of cancer, subtype of cancer (morphology, topography), stage of cancer, clinical group, cancer location, treatment prescribed, treatment outcome whether there was a relapse. Each row of the data set has a unique identifier, which shows how the same patient underwent treatment several times with the time period for hospitalisation.

[1] https://apps.who.int/iris/bitstream/handle/10665/96612/9789241548496_eng.pdf

[2] A usually aggressive malignant neoplasm of the soft tissue or bone, and a rare kind of cancer, but in this case cancer appears in breast/pancreas and spread (metastasize) to other parts of the body

u_id	Regional Dispen	Location (region	Date of birth	Gender	Ethnicity	Registration dat	Hospitalization c	Discharge date	Deregistration
414927609	Almaty Oncology Center	Almaty city	Mon Feb 20 00:00:00 CET 1967	Female	Uighurs	Wed Aug 24 12:49:00 CEST 2016	Thu Aug 03 00:00:00 CEST 2017	Fri Aug 04 00:00:00 CEST 2017	Fri Apr 13 00:00:00 CEST 2018
413249067	Almaty Regional multidisciplinary clinic	Almaty region	Mon Oct 11 00:00:00 CET 1954	Female	Kazakhs	Fri Jul 29 23:03:00 CEST 2016	Wed May 17 00:00:00 CEST 2017	Mon May 22 00:00:00 CEST 2017	Not indicated
411647023	Shymkent Oncology Center	Shymkent	Fri Sep 10 00:00:00 CET 1954	Female	Kazakhs	Thu Dec 22 21:55:00 CET 2016	Tue Jun 20 00:00:00 CEST 2017	Fri Jun 23 00:00:00 CEST 2017	Not indicated
402568150	East Kazakhstan Regional Multidisciplinary Center of Oncology and Surgery	East Kazakhstan region	Mon Oct 09 00:00:00 CET 1950	Female	Russians	Mon Sep 02 09:00:00 CEST 2013	Mon Nov 25 00:00:00 CET 2013	Wed Dec 04 00:00:00 CET 2013	Wed Oct 12 00:00:00 CEST 2016
411060144	Shymkent Oncology Center	Shymkent	Mon Aug 16 00:00:00 CET 1965	Female	Kazakhs	Tue Apr 09 09:00:00 CEST 2013	Mon Jan 06 00:00:00 CET 2014	Fri Jan 17 00:00:00 CET 2014	Mon Feb 10 00:00:00 CET 2014
15300000000000000	Kyzylorda Regional Oncology Center	Kyzylorda region	Fri Apr 02 00:00:00 CET 1954	Female	Kazakhs	Mon Nov 14 23:39:00 CET 2016	Wed May 31 00:00:00 CEST 2017	Tue Jun 13 00:00:00 CEST 2017	Not indicated
420295134	North Kazakhstan Regional Oncology Dispensary	North-Kazakhstan region	Wed Dec 13 00:00:00 CET 1950	Female	Russians	Thu Oct 27 16:11:00 CEST 2016	Thu Jan 05 00:00:00 CET 2017	Wed Jan 11 00:00:00 CET 2017	Not indicated
416540693	Almaty Oncology Center	Almaty city	Tue Apr 01 00:00:00 CET 1952	Female	Russians	Fri Apr 05 00:00:00 CEST 2013	Fri Apr 01 00:00:00 CEST 2016	Mon Apr 04 00:00:00 CEST 2016	Not indicated

Figure 6.1 Personal Data of Patients

Diagnosis	Diagnosis confir	Morphological ty	Topography	Stage	Clinical group	Cancer localizat	Prescribed treat	Treatment outco	Relapse
C50	morphological	carcinoma, NOS	C50 BREAST	II b	III	unknown	chemotherapeutic	improvement	other
C50	morphological	infiltrating duct carcinoma, NOS (C50._)	C50 BREAST	II b	III	unknown	chemotherapeutic	improvement	continued treatment of the primary tumor
C50	morphological	infiltrating duct carcinoma, NOS (C50._)	C50 BREAST	II stage	II	unknown	chemotherapeutic	no change	continued treatment of the primary tumor
C50	cytological	adenosquamous carcinoma	C50.9 Breast, NOS	II a	II	Not indicated	chemotherapeutic	improvement	continued treatment of the primary tumor
C50	morphological	neoplasm, malignant	C50 BREAST	II stage	II	unknown	chemotherapeutic	improvement	continued treatment of relapse
C50	morphological	adenocarcinoma, NOS	C50 BREAST	IV stage	IV	unknown	beam	improvement	primary tumor treatment
C50	morphological	carcinoma, NOS	C50 BREAST	II stage	II	unknown	chemotherapeutic	improvement	continued treatment of the primary tumor

Figure 6.2 Medical Data of Patients

In the data obtained, cancer types are described in the ICD-10 Code. The first diagnosis found in the dataset is C25, which corresponds to pancreatic cancer, and it has several child codes found in the data obtained.

C25.0 Malignant neoplasm of head of pancreas

C25.1 Malignant neoplasm of body of pancreas

C25.2 Malignant neoplasm of tail of pancreas

C25.3 Malignant neoplasm of pancreatic duct

C25.4 Malignant neoplasm of endocrine pancreas

C25.7 Malignant neoplasm of other parts of pancreas

C25.8 Malignant neoplasm of overlapping sites of pancreas

C25.9 Malignant neoplasm of pancreas, unspecified (WHO, 2004).

The second type of cancer found in the obtained data is breast cancer, and it corresponds to the code C50 in ICD-10. This type of cancer comes with nine subcodes, but in our datasets only these three child codes are found.

C50.0 Malignant neoplasm of nipple and areola

C50.1 Malignant neoplasm of central portion of breast

C50.2 Malignant neoplasm of upper-inner quadrant of breast (WHO, 2004).

Several cases of sarcoma, such as sarcoma (NOS), pleomorphic rhabdomyosarcoma, leiomyosarcoma (NOS), epithelioid leiomyosarcoma, myxoid leiomyosarcoma, alveolar rhabdomyosarcoma were detected from the database of pancreatic and breast cancer cases in the "morphological type" column. This could be used as a separate dataset to study cases of sarcoma. In this regard, the advantages of creating FDP for rare cancer diseases are also evident for which data is available in a minimal number of centres around the world. No single organization, and in most cases, no single country, has enough patients to conduct general studies or clinical trials on rare cancers. Combining all these data through FDPs may influence not only precision treatment of cancer patients, but also this could better address the issue with the rare diseases.

6.2 Comparison with international cancer database

The database obtained from the MOH will be compared to large-scale collaborative cancer project such as the Cancer Genome Atlas (TCGA)[3]. The database is registered on FAIRsharing[4] which lists over a thousand FAIR-related data and metadata standards (Sansone et al., 2019). TCGA is a joint effort led by the National Institutes of Health (NIH) which collects, characterizes, and analyzes cancer samples to improve the prevention, diagnosis and treatment

[3] https://portal.gdc.cancer.gov/

[4] "FAIRsharing is an informative and educational resource that describes and interlinks community-driven standards, databases, repositories and data policies" (Sansone et al., 2019).

of cancer [5].

Data available through the GDC (https://portal.gdc.cancer.gov/) provides researchers with access to standardized clinical and genomic cancer research data. It comprises clinical data, genomic information and high-level sequence analysis of the tumour genomes (Pavlopoulou et al., 2015). Clinical section composes demographic data which is shown in Fig.6.1. (gender, race, ethnicity, birth date, vital status), diagnoses/treatment which is shown in Fig.6.2. (icd-10 code of the disease, classification of tumour, the structure and activity of a patient's tumor, age at diagnosis, treatment received, and outcomes) and exposure(alcohol and smoking history, weight and height parameters of the patient) subsections. Biospecimen[6] section stores mostly all data related to blood, urine, tissue, cells, DNA, RNA, or protein[7]. It also stores information on somatic Mutations.

Data available through the FDP are also offering the same categories as GDC, including medical and personal history of patients. However, this could be improved by providing more data regarding tumour specifications, such as the size, grade of the tumour and etc. Moreover, most international data repositories are unlocking the genomic data of cancer patients, which help to better understand genetical causes of cancer. This study does not include genomic data of patients due to the scope of the project and the limited time. Despite this, including genomic data of cancer patients as further work could significantly affect further treatment of oncology both globally and within Kazakhstan. This might influence the entire treatment in precision oncology[8], since it uses a precise knowledge of the structure and activity of a patient's tumor genome to suggest particular therapies, thereby providing meaningful therapeutic responses (Jensen et al., 2017). Since Kazakhstan is adopting IBM Watson[9] to provide tailored treatment for patients, this also might be a big contribution for the given approach. For that, digitizing

[5]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga

[6]A material sample obtained from a biological object/living organism for testing, diagnostic, distribution, treatment or research objectives

[7]https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biospecimen

[8]Precision oncology(medicine) is an approach which uses genomic data of cancer patients, and thus, leads to better diagnosis and treatment strategies that tailored for patients with various tumours

[9]https://www.inform.kz/en/healthcare-ministry-ibm-to-introduce-ai-to-kazakhstan-s-oncology$_a$3044271

```
"exposures": [
  {
    "years_smoked": null,
    "state": "released",
    "cigarettes_per_day": null,
    "exposure_id": "7cbedf3a-c92c-5eea-9497-362e705179e4",
    "alcohol_history": "Not Reported",
    "alcohol_intensity": null,
    "height": null,
    "bmi": null,
    "updated_datetime": "2019-07-31T21:59:49.395320-05:00",
    "submitter_id": "TCGA-AN-A046_exposure",
    "weight": null,
    "created_datetime": null
  }
],
"demographic": {
  "demographic_id": "001b6933-5d74-5eb9-bdfe-93995f809818",
  "state": "released",
  "vital_status": "Alive",
  "gender": "female",
  "race": "white",
  "year_of_birth": 1942,
  "days_to_birth": -25158,
  "age_at_index": 68,
  "year_of_death": null,
  "ethnicity": "not hispanic or latino",
  "updated_datetime": "2019-07-31T21:59:49.395320-05:00",
  "submitter_id": "TCGA-AN-A046_demographic",
  "created_datetime": null
},
```


Figure 6.3 How data are presented on the website GDC Data Portal. Part 1.

clinical data as well as genomics data in a FAIR manner may also have implications for other aspects of development within the country. For instance, digitizing the genome of Kazakhstani citizens could be one of the key solutions for studying the history of Kazakhstan with the second breath, and its general capabilities along with national security (Zhabagin, 2018). Therofore, making data available in a FAIR manner might enforce other fields of study, not only healthcare field.

```
"diagnoses": [
  {
    "icd_10_code": "C50.9",
    "submitter_id": "TCGA-AN-A046_diagnosis",
    "state": "released",
    "progression_or_recurrence": "not reported",
    "days_to_recurrence": null,
    "tumor_grade": "not reported",
    "classification_of_tumor": "not reported",
    "age_at_diagnosis": 25158,
    "ajcc_pathologic_t": "T2",
    "site_of_resection_or_biopsy": "Breast, NOS",
    "updated_datetime": "2019-08-08T16:23:06.992266-05:00",
    "ajcc_pathologic_n": "N0",
    "last_known_disease_status": "not reported",
    "days_to_last_follow_up": 10,
    "created_datetime": null,
    "days_to_last_known_disease_status": null,
    "prior_treatment": "No",
    "ajcc_pathologic_m": "M0",
    "treatments": [
      {
        "treatment_effect": null,
        "submitter_id": "TCGA-AN-A046_treatment_1",
        "state": "released",
        "treatment_intent_type": null,
        "treatment_type": "Pharmaceutical Therapy, NOS",
        "treatment_anatomic_site": null,
        "days_to_treatment_start": null,
        "days_to_treatment_end": null,
        "treatment_outcome": null,
        "treatment_or_therapy": "no",
        "treatment_id": "8c021e82-8589-521b-8e1a-806ed57b8544",
        "created_datetime": "2019-04-28T14:06:15.305677-05:00",
        "regimen_or_line_of_therapy": null,
        "therapeutic_agents": null,
        "updated_datetime": "2019-07-31T21:59:49.395320-05:00",
        "initial_disease_status": null
      },
      {
        "created_datetime": null,
        "treatment_intent_type": null,
        "treatment_or_therapy": "no",
        "treatment_id": "61b73b83-2962-5cb9-947b-07722d11e202",
        "therapeutic_agents": null,
        "updated_datetime": "2019-07-31T21:59:49.395320-05:00",
        "submitter_id": "TCGA-AN-A046_treatment",
        "treatment_type": "Radiation Therapy, NOS",
        "state": "released"
      }
```


Figure 6.4 How data are presented on the website GDC Data Portal. Part 2.

Kazakhstan has not obtained standard requirements and protocols for biomedical institutions. Although hundreds of samples have been identified from published data, they are still not standardized (Momynaliev & Imanbekova, 2014). Consequently, an establishing of biobank with standardized requirements could create a better environment for quality research (Momynaliev & Imanbekova, 2014). The National Center for Biotechnology has already begun a biobank with more than 1,500 blood samples, intending to create a biobank including around 10,000 blood samples of healthy volunteers (Momynaliev & Imanbekova, 2014). The creation of biobanks can be a positive contribution to the development of medical science in Kazakhstan. In this regard, creation of biobanks in a FAIR manner connected to cancer data might significantly affect research quality in Kazakhstan and around the world.

6.3 Conclusion

Data which presented in this study is a part of cancer data which are stored in the information systems of Kazakhstan. Therefore, within the scope of this study we cannot make explicit comparison to global cancer datasets. Our datasets in comparison to the global cancer data has the same categories such as, personal data and medical data. However, personal data can be improved by adding more personal history of patients, including smoking history, alcohol, any genetic problems and etc. Medical data might have more specifications on tumour and genomic data of patients. Thus, the inclusion of genomic data and the personal history of patients in the FDP can provide a more accurate picture of the causes of diseases and provide better treatment for cancer patients.

Chapter Seven

Designing a FAIR Data Point

This study has two main objectives: firstly, to exchange data in a FAIR manner, thus allowing other data users to discover metadata and get access to actual data under well-defined protocols. Secondly, integration into larger data sets will provide an opportunity to contribute to global cancer research.

7.1 Objectives

The research is adopting FAIR Data Point developed by FAIRDataTeam, which is available on the github. FAIR Data Point is a data registry that provides data and metadata using FAIR Data principles. It helps data owners to expose their data in a FAIR manner, and also allows data users to find metadata and access them if license conditions allow. Although FAIR Data Point could be applied for many knowledge domains, in this research we will be focusing on the cancer data of patients with breast cancer, pancreatic cancer and several cases of Sarcoma. The FDP prototype will use these types of diseases to test the feasibility of this study.

The objective of the designing FAIR Data Point is to illustrate how oncological data on breast cancer, pancreatic cancer and Sarcoma can be assigned machine-readable metadata, to enable them to be discoverable by individuals and machines. The basis of this project is that the Kazakhstan Cancer Center is ready to receive knowledge at the international level, sharing experience with others. The development of this FDP can contribute to cancer research worldwide, as well as in Kazakhstan, thereby improving the health of millions of people around

the world. Deploying FDP for healthcare of Kazakhstan has two main approaches: (i) to be used as a stand-alone web application, where data owners share their own data and data consumers get access to the particular data, thus, contributing to Kazakhstani research and medicine, (ii) to be the part of larger interoperability systems providing accessibility functionality globally.

7.2 Implementation

This chapter consists of implementation steps which has been done to deploy the FDP. A video demonstrating the FTP deployment procedure is available here: https://yadi.sk/i/o$_7 I - 7ZJPLwEnw$.

7.2.1 Implementation choices

Even though FAIR Principles are widely known and are being accepted by many communities, the implementation choices still remain a global challenge (Sustkova et al., 2019). The common representation of data and agreeing on community standards is one of the challenges that FAIR has faced today (Wise et al., 2019). To solve this issue, the FAIR Convergence Matrix has been launched in 2019 by the GO FAIR Community (Sustkova et al., 2019) that allows to use and reuse resources used by other communities. It was aimed to make a complete list of existing resources that communities have clearly chosen to implement FAIR, and analyze these resources to determine optimal reuse tactics (Sustkova et al., 2019). The resources provided on the FAIR Convergence Matrix consist of policies, technologies, data, metadata, ontologies, data models, and standards (Sustkova et al., 2019). Using and reusing resources provided by the FAIR Convergence Matrix will alleviate implementation of FAIR, and specifically, present data in a common way.

7.3 FAIRification process

FAIRification process of non-processed data takes several steps. Requirements for findability and accessibility can be achieved at the metadata level, whereas interoperability and reusability

require more works at the data level as shown in Fig.7.1 [1]. These steps will be described below because they are important to get FAIRified data, all description is taken at this link: https://www.go-fair.org/fair-principles/fairification-process/.

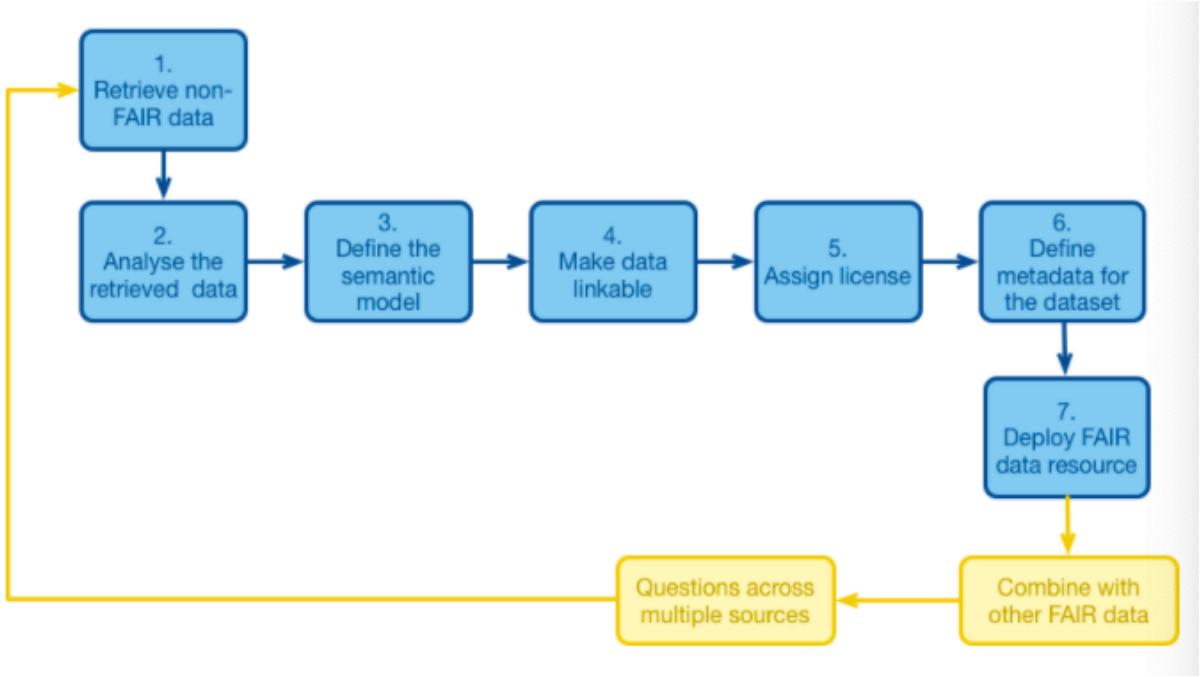

Figure 7.1 The FAIRification process of data [2]

The first step of the FAIRification process is to retrieve the data to be FAIRified. The next step focuses on the analysis of the database structure, whether the database is relational or not, as well as field names. The third step is defining the semantic model which can determine dataset accurately and in a computer-actionable way. This also includes capturing terms, URIs and descriptions. Making data linkable is the next step. This step can be done using the Semantic Web and Linked Data technologies. The fifth step is to configure access licenses. What kind of usage rights need to be attached to the data, who can use data without any ambiguity. The sixth step is about defining rich metadata for the dataset. The last step is publishing FAIRified data together with the metadata and a license. Thus, metadata can be found by search engines and the data can be accessed under well-defined licenses.

The semantic model, as shown in Fig.7.2 was built on the basis of data obtained from the Ministry of Health in order to make data linkable.

[1] https://www.go-fair.org/fair-principles/fairification-process/

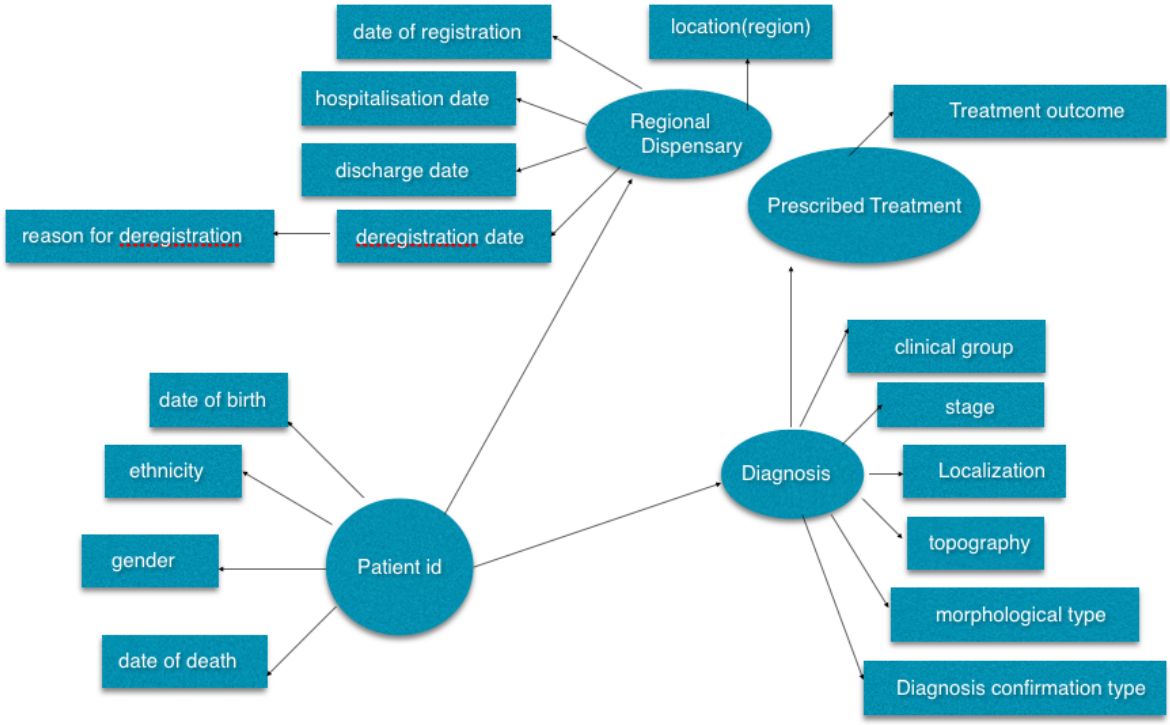

Figure 7.2 The Semantic data model

The Fig.7.3 illustrates how the semantic data model can be built in RDF Turtle.

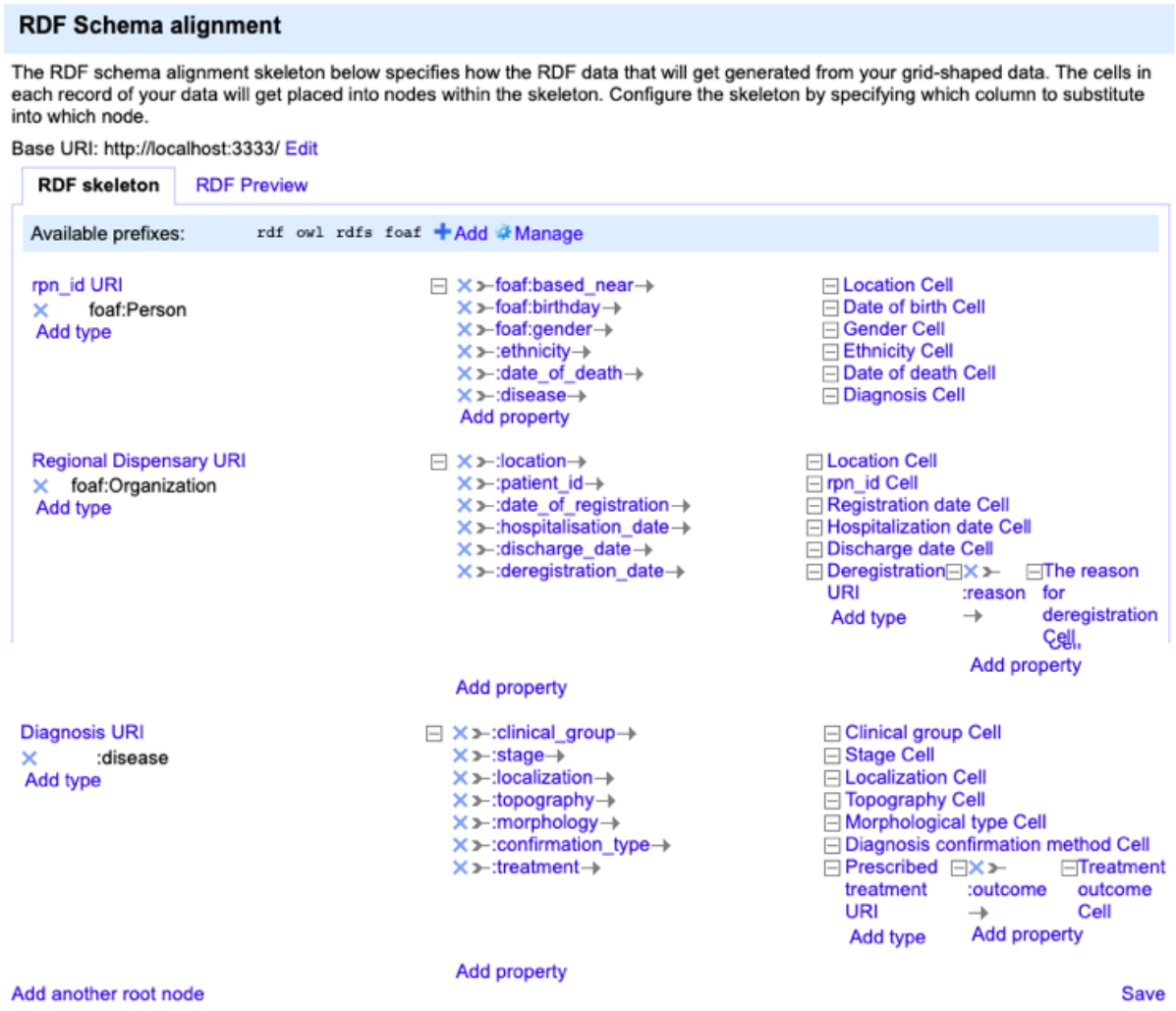

Figure 7.3 Skeleton of Semantic data model

In the Fig.7.4 RDF preview code is presented.

Figure 7.4 Semantic data model written in RDF Turtle

7.3.1 OpenRefine

For the FAIRification of the data obtained from the Ministry of Health OPenRefine tool was used. OpenRefine is a software tool that enables preprocessing data: cleaning it; transforming it from one format into another; and extending it with web services and external data [3]. It is important to note that private data can be stored on computer until the owner of data wants to share or collaborate with anyone. It operates by running a small server on a computer and the interaction with it can be done through the browser. Datasets from the OpenRefine can be downloaded in these formats: tab separated (tsv), comma separated (csv), Excel (xls , xlsx), JSON, XML, RDF as XML, Google Spreadsheets [4]. Thus, datasets can be converted to machine-actionable formats, and therefore, can be interoperable and reusable for further needs. OpenRefine also can be used to connect and expand datasets with various web services [5]. In this regard, the OpenRefine FDP extension was used to link datasets in OpenRefine with

[3]https://openrefine.org/

[4]https://datacarpentry.org/openrefine-socialsci/aio.html

[5]https://openrefine.org/

FDP itself. Using this extension, data can be easily FAIRified directly in OpenRefine [6].

7.3.2 FAIR Data Point Components

The deployment of FDP includes these components as shown in Fig.7.5.

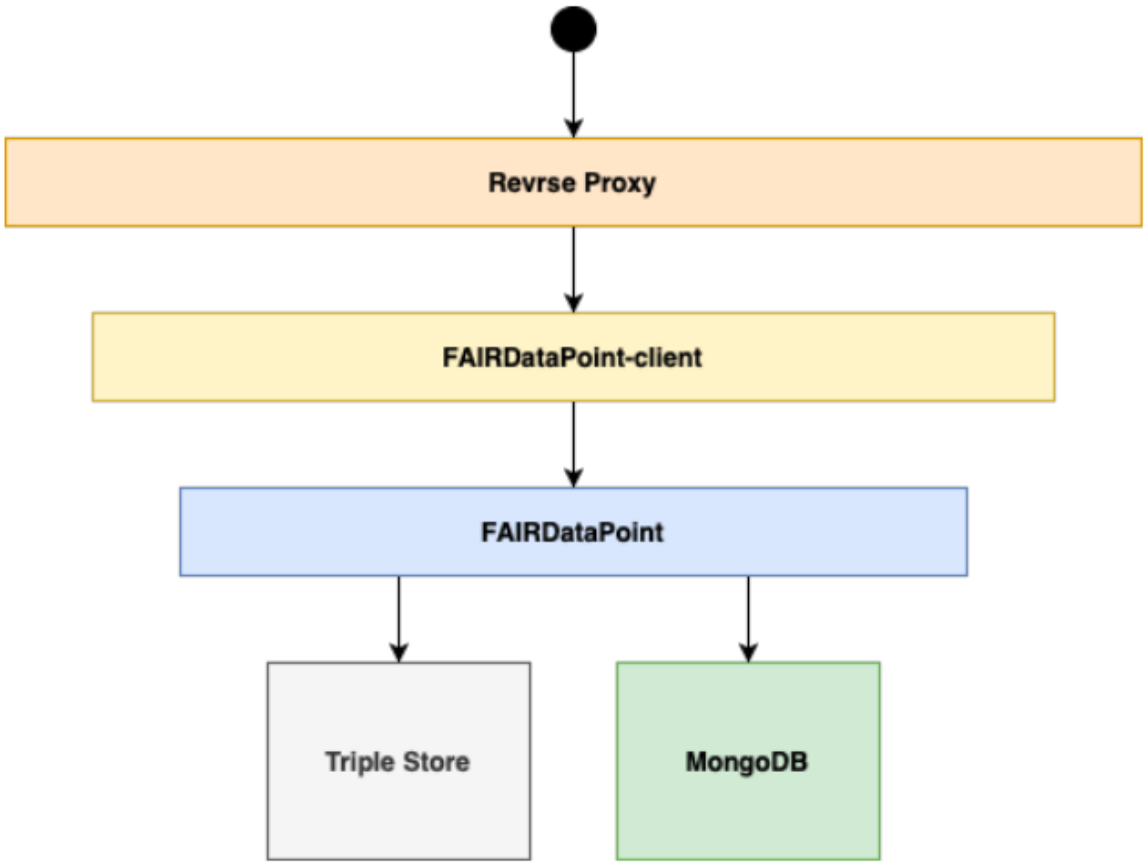

Figure 7.5 FAIR Data Components [7]

Triple store: Semantic data in the FAIR Data Point needs to be stored somewhere. Triple store is default place where semantic data is stored [8].

MongoDB: a place where information about user accounts and their roles are stored.

FDP: FAIRDataPoint is the core component which manages the entire business logic and any operations with the semantic data.

FAIRDataPoint-client: this component is in charge of providing a user interface for hu-

[6]https://readthedocs.org/projects/fairdatapoint/downloads/pdf/latest/

[8]https://readthedocs.org/projects/fairdatapoint/downloads/pdf/latest/

mans. It acts as a reverse proxy in front of the FAIR data point, which decides whether the request is for machine-readable data and transfers it to FAIRDataPoint.

Reverse Proxy: in a production deployment, there is usually a reverse proxy that handles HTTPS certificates, so connecting to FAIR Data Point is protected.

7.3.3 Metadata Specification

The FAIR data point (FDP) is a web-based application for creating, storing and serving metadata that meets all requirements of FAIR principles. The FDP prototype was built based on Figure 7.6, which is part of the architecture shown in Figure 7.5. The structure of FDP represented by ArchiMate's[9] Application layer notation, and inside the FAIR Data Point box from the top-down. The existing API consists of two parts, such as the Metadata Provider API and FAIR Data Accessor API. Both of them are a public interface to provide access to their services. The Data Accessor Service provides the Data Access function to provide access to the data in a FAIR Format. The Metadata Provider Service provides a Metadata Retrieval function with sub-functions:

- FDP Metadata Retrieval is returning the FDP Metadata. The FDP Metadata consists of a list of Catalog Metadata.
- Catalog Metadata Retrieval is returning the Catalog Metadata. The Catalog Metadata consists of a list of Dataset Metadata.
- Dataset Metadata Retrieval is returning the Dataset Metadata. The Dataset Metadata consists of a list of Distribution Metadata, and it can have a Data Record Metadata information.
- Distribution Retrieval is returning the Distribution Metadata. The Distribution Metadata is information about the representation of a dataset. For instance, the file format, download URL, size or access.

[9]The ArchiMate is an open-source modelling and sketching tool that meet Enterprise Architecture standard

- Data Record Metadata is returning the Data Record Metadata. The Data Record Metadata is information about the structure and content of the dataset. For instance, types of columns in the table, range, relations, or domains.

The detailed information about each of metadata object listed above are given in the metadata section inside of the Appendix.

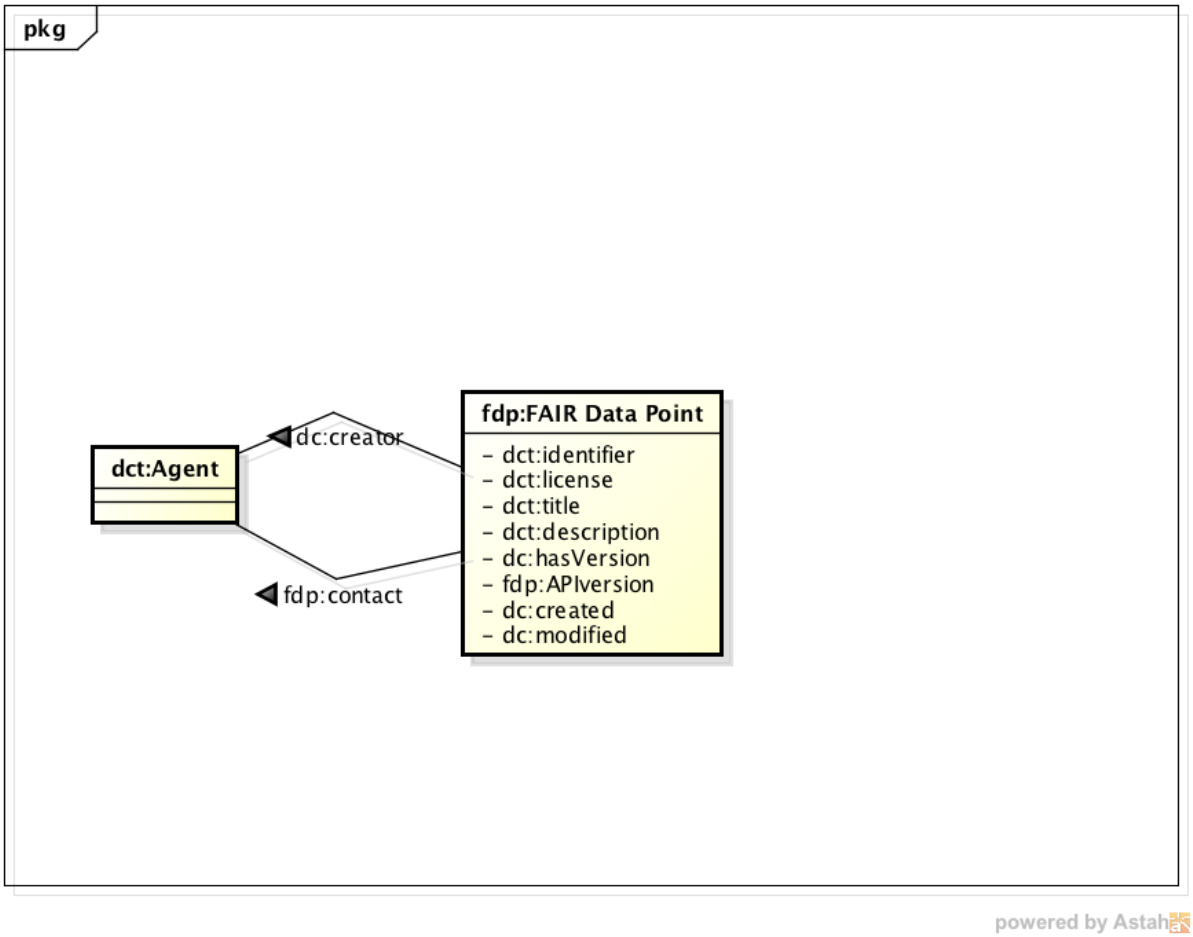

Figure 7.6 Mapping dc and dct terms (FAIRDataTeam, 2020)

FDP prototype consists of three parts, such as a server, client and storage, and each part has its function. For instance, the client-side is a graphical user interface that provides intuitive and clear guidance on how to work with metadata. The server-side then acts as the connection point between the client-side and the storage side to handle the entire user request properly. Before building the prototype, we had to meet software requirements. We installed an environment such as Docker, Java and Maven. Then, we pulled the docker images by creating a compose file. The images are already in the cloud which uploaded by FAIRDataTeam, and by

setting the correct parameters and address, we got the FDP. Finally, we successfully launched FDP locally on the machine.

7.3.4 Metadata of Datasets

The metadata presented on a mock FDP stores information about three types of diseases with unique identifiers and descriptions for each dataset. To create metadata about the datasets OpenRefine[10] software with the FAIR[11] extension has been used, it makes possible to connect to our FDP and manage metadata. Using OpenRefine, we pre-processed our data. The data were received in the Russian language. It was presented in two files with both of them 65000 rows. Each line of the dataset has unique id which shows how the same patient was treated several times with the given time reference for each treatment. All the data were translated into English. Then, we uploaded datasets into the OpenRefine. In Figure 7.7, you can see the columns and rows of our datasets that can be altered, deleted or added. After pre-processing of the datasets, we connected them to the FDP and created metadata as shown in Figure 7.8. Note that fully detailed information about metadata object can be found in the Appendix.

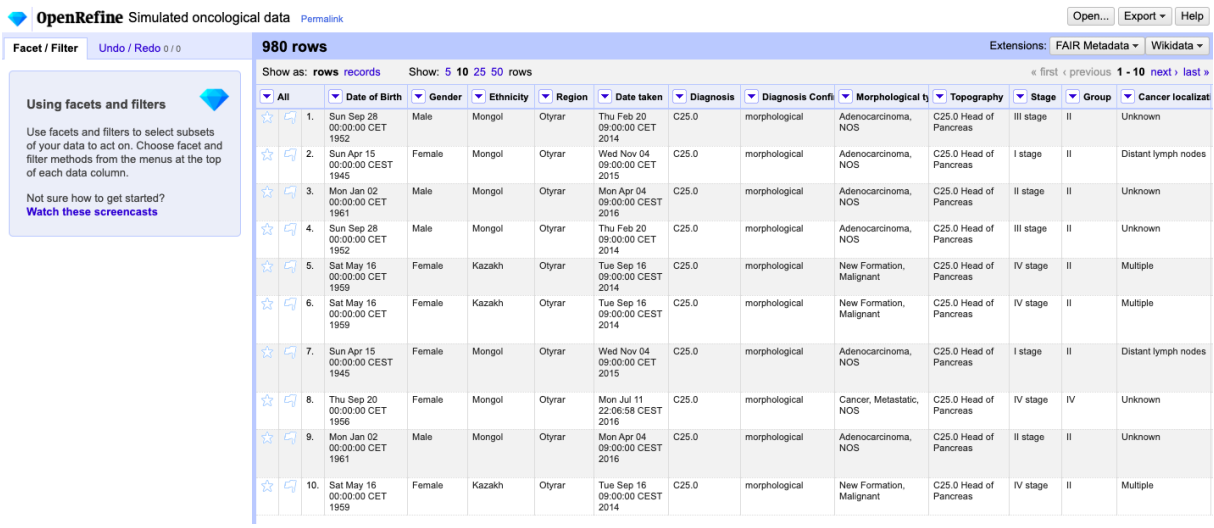

Figure 7.7 OpenRefine's user interface

[10]OpenRefine is a tool for working with data to transform, clean and extending with web services.
[11]OpenRefine extension to support FAIR Data Point with FAIR Metadata.

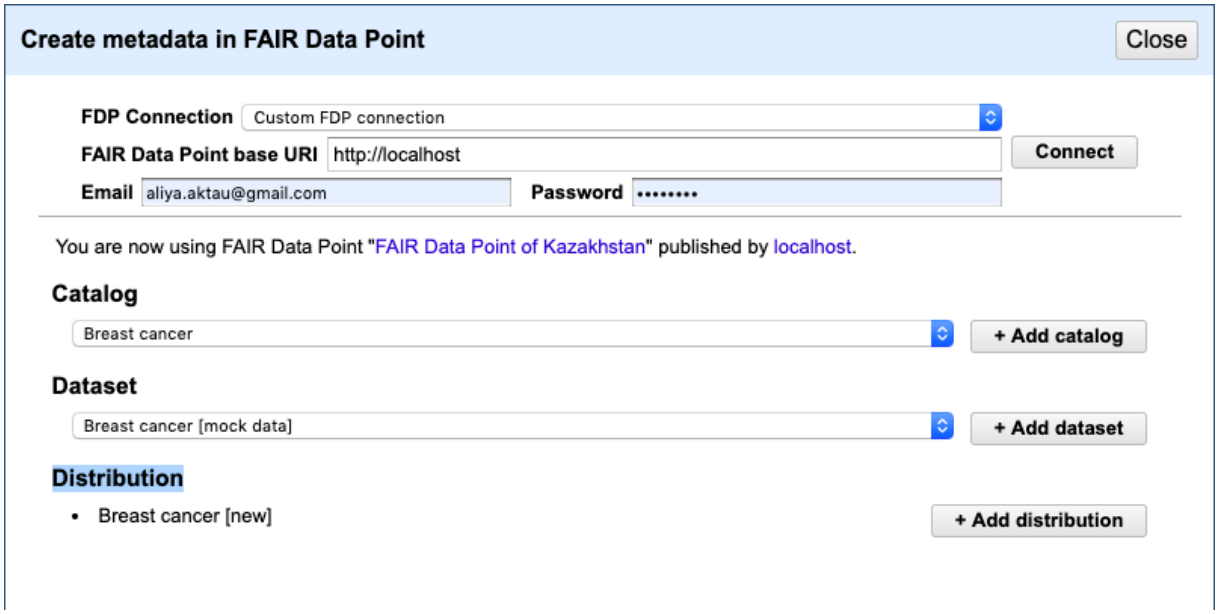

Figure 7.8 Creation of FAIR metadata

7.4 Results

Deploying the FDP for this study was enabled using Open Refine with the FDP extension, which allows to manage data, and FDP that is designed in accordance with the FAIR Guiding Principles. All datasets are stored in Open Refine, and metadata was created when Open Refine was connected to FDP. Metadata was written for all datasets with an explicit description giving detailed information about each of them, including disease type, time clause and whether patients diagnosed for the first time or with the secondary cancer. As a result, we got a website, as illustrated in Figure 7.9. The website represents the list of catalogues of the dataset. Catalogues store information about datasets themselves and distributions where through the link data can be downloaded if licenses allow. The distribution page is shown in Fig. 7.10., and using the 'download' button, users can download the dataset or get a contact of dataset's owner or linked to the website of the data owner, depending on how sensitive the data is. Another thing about the website is that it can provide the ability to create users with a role, such as administrator and user, as shown in Figure 7.11. The administrator has access to all metadata created by users and can manage information. Users can control only the information that was created by them.

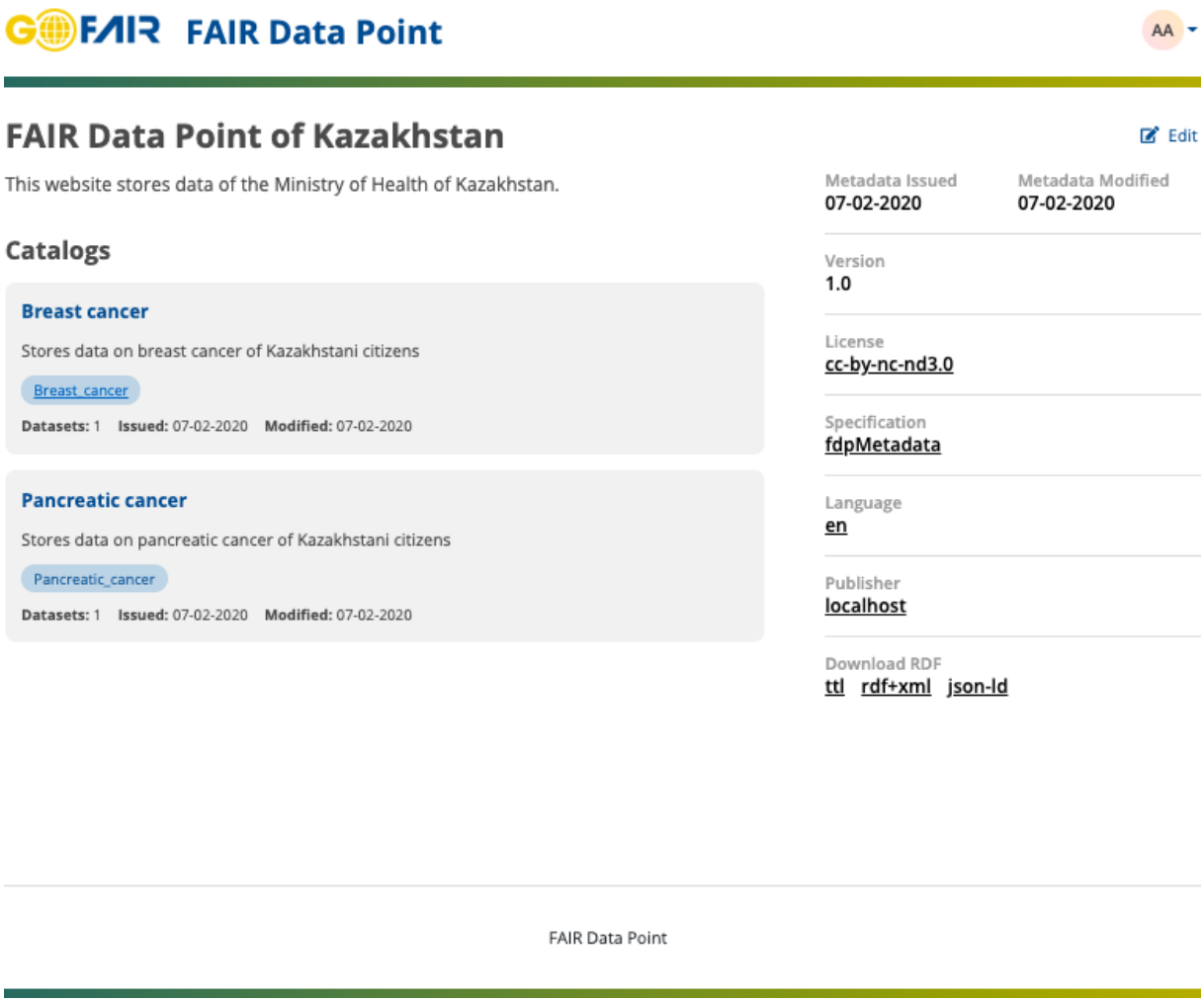

Figure 7.9 Main page of FDP

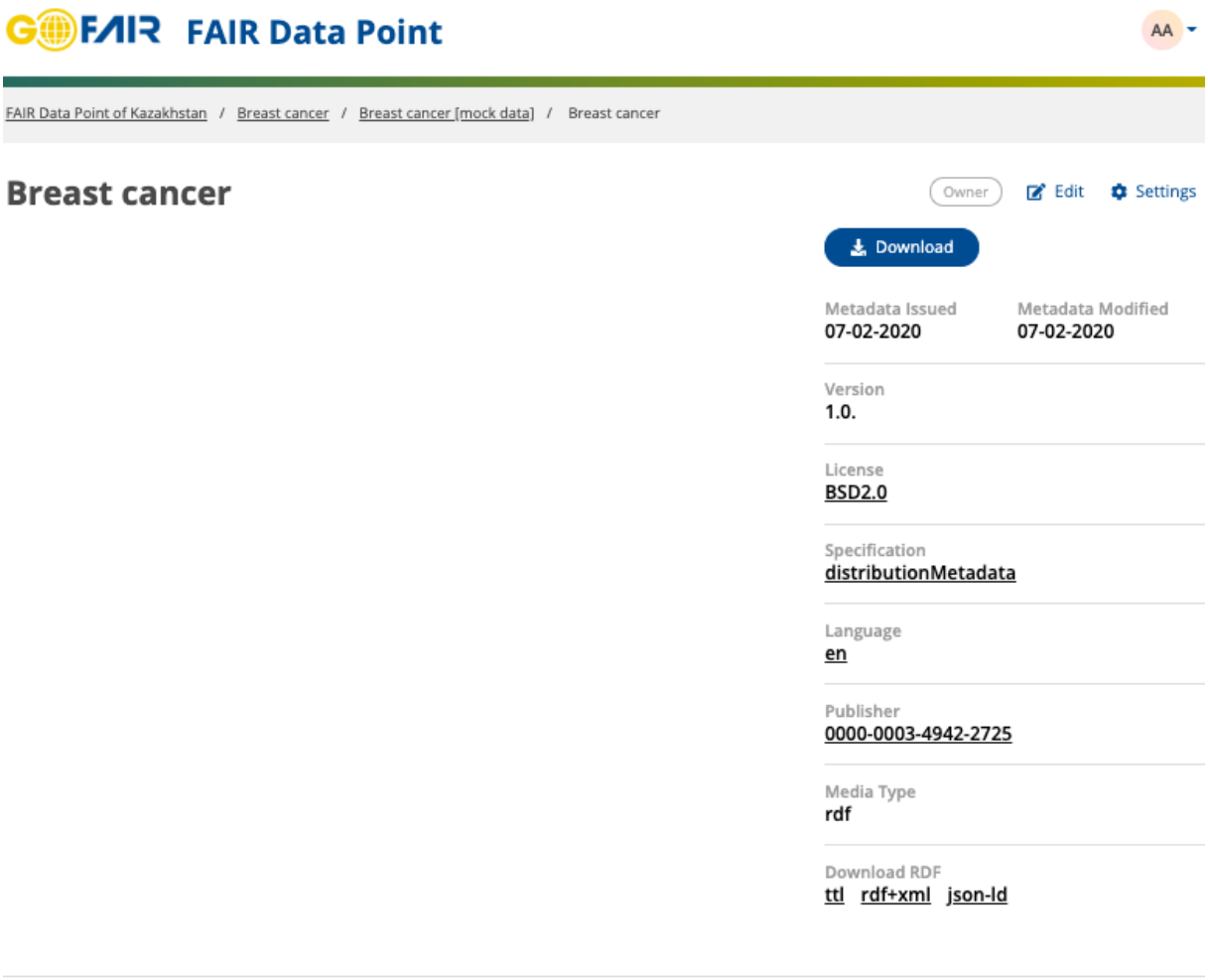

Figure 7.10 Distribution page of FDP

7.5 Obstacles

The data was obtained from the Ministry of Health contained two files: (i) cancer data of patients who first registered at the clinic and (ii) cancer data of patients registered with the secondary cancer. The datasets were written in Russian language and translation of them was done through international standards ICD-10, ICD-0-3. Since the number of lines in each file was about 65,000 lines with approximately 15-20 columns, the time spent on translation was extremely long. Therefore, for applying of FDP for Kazakhstan will require extra effort for translation and delivering data in English.

H

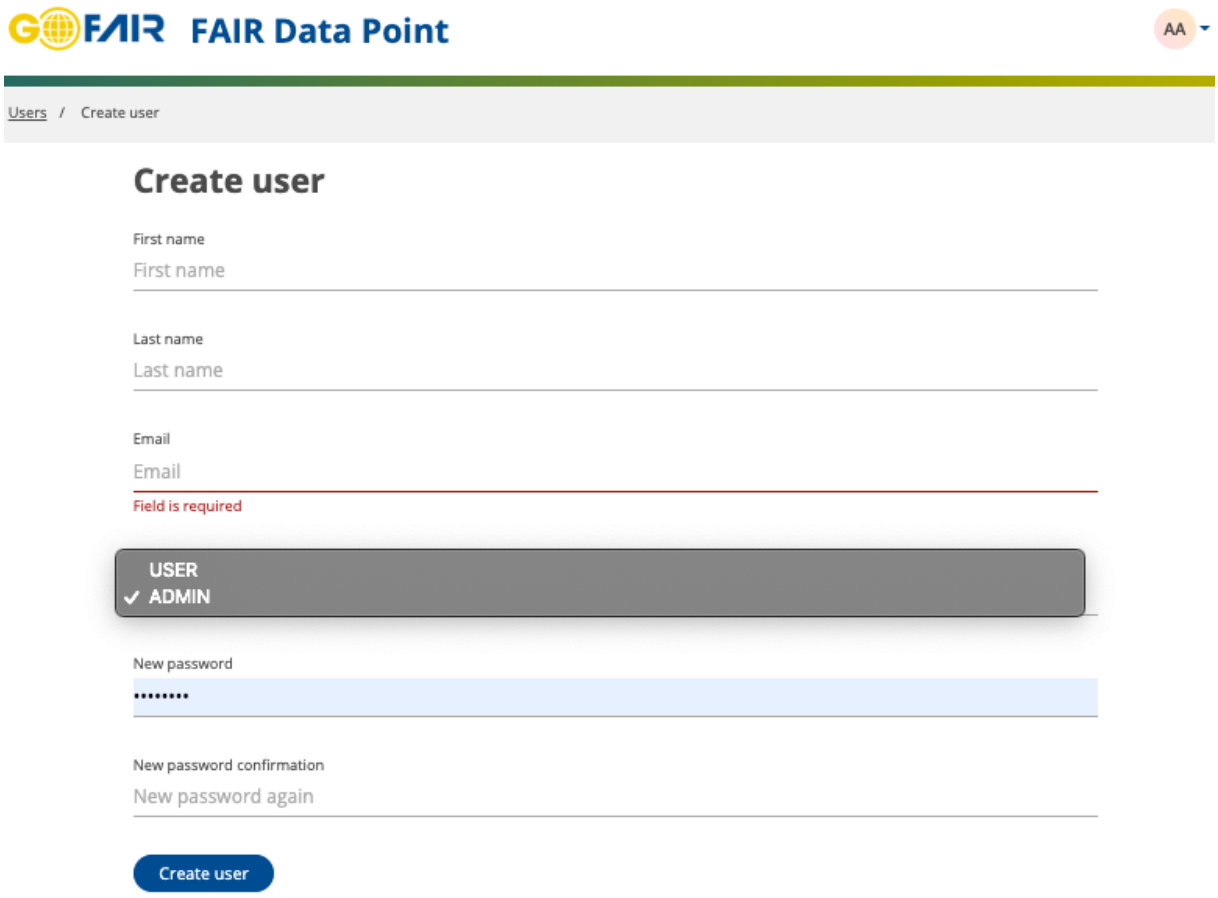

Figure 7.11 User creation page of FDP

7.6 Conclusion

This chapter was aimed at deploying the FAIR Data Point to illustrate how it can be used in Kazakhstan's healthcare. Deployment of FDP have been done through the Specification which provided on the GitHub (FAIRDataTeam, 2020). The main reason for applying FAIR Principles for health data of Kazakhstan is to allow data owners to expose datasets along with data users to discover offered datasets. The next goal is to integrate with larger data sets to contribute to global cancer research and help Kazakhstan benefit from being connected to global science. Deployment of FDP has not taken too long as all steps are provided by the FAIRDATATEAM on the GitHub, and everything is written there is intuitively clear. Moreover, if anyone has any questions regarding the implementation part they will answer to any question and support where it is necessary.

The deployment of the FDP has been done using clinical data from EROB Kazakhstan. However, most international datasets such as, TCGA (The Cancer Genome Atlas, ICGC (International Cancer Genome Consortium) and NCI-MATCH (National Cancer Institute's Molecular Analysis for Therapy Choice) provide data access linking genomic and clinical data/results (Siu et al., 2016). Therefore, as further research, it would be useful to ensure the exchange of clinical and genomic data to facilitate access to innovative targeted treatments within Kazakhstan and globally.

As described above, the translation of the datasets required much time. However, we might assume that since Kazakhstan adopted SNOMED-CT terminology at the beginning of 2019 [12] alignment of health data of Kazakhstan with the FAIR-based databases should not take much effort and time.

The illustration of this FDP shows the possibility of applying FAIR Principles to Kazakhstan's healthcare. Ontologies, technologies, and standards are important aspects of data exchange for full data interoperability. For this, the resources provided in the FAIR Convergence Matrix can be used to successfully disseminate data and exchange data.

[12] https://www.snomed.org/our-stakeholders/member/kazakhstan

Chapter Eight

Discussion and Conclusion

Since 2012, the active development of digital healthcare in Kazakhstan began with the "Strategy 2050", which set new goals for digitizing healthcare. Most of the subsequent national health development programs were implemented in accordance with Strategy 2050, such as Concept 1.0, Salamatty Kazakhstan, Information Kazakhstan 2020 and Densaulyk 2016-2019. In December 2019, another e-health development program for 2020–2025 was introduced ("State Health Development Program for 2020-2025", 2019), for which the 2050 Strategy also serves as a fundamental document. Under the new Program, the integration of health information systems is still necessary, although the goal of launching the integration platform began with Concept 1.0 in 2013. According to Obermann (2016), Kazakhstan does not have an independent organization that could analyze government health programs and determine whether goals were achieved properly or not, and which approaches work and which do not. This is evident from the repetition of the same tasks following one program after another which is launching an integration platform where communicates with each other. Thus, this study was done to address this issue along with since the informational value of data is not used enough.

Two cases of introducing FAIR are provided in this study. The first one, FDP as a standalone application for storing data within the country, the second one, as a part of large integrated data store. For that, initiatives of Kazakhstan on digital health and cancer initiatives were studied. The study mostly addressed aspects of availability and interoperability of digital health data: the ability to access data (syntactic interoperability); understand the data once retrieved (semantic interoperability); Along with this, testing the feasibility of deploying FDP

for digital healthcare in Kazakhstan was another goal of this study.

This study was also conducted to understand how Kazakhstani digital health data and cancer data, can be processed using FAIR data principles. In this regard, FDP prototype was created to store (meta)data and make possible to exchange data on the global level. Thus, this may make better precised oncology for patients of Kazakhstan and globally. For this, FDPs with cancer data throughout the world store data and act as access points and allow AI models to visit them. In other words, data integration happens due to the fact that data accessible to data consumers as one integrated data store.

One of the major problem in Kazakhstan is that data are not yet communicating due to fragmentation problems, since they are stored on various entry points and gathered at the level of government bodies for analytical purposes. Patient data is not accessible from various entry points and data stored on the level of healthcare organisations does not communicate with each other. The MOH has been working on the integration platform of health data. For this, the integration platform was introduced in the Concept 1.0 in 2013, and the implementation continued with the program Densaulyk 2016-2019, Informational Kazakhstan 2020, and even the platform itself has been finished by 2020, the entire integration of datasets is still one of the main objectives of the e-health state program for 2020-2025. One of the significant results of the above given programs was switching to paperless medical documentation and presenting Electronic Health Passports, which have been in use since 2019. Even though Electronic Health Passports were launched, the integration platform has not yet been fully set up which requires further steps towards full interoperability.

Apart from the data interoperability, data must be able to communicate in the meaning. In this regard, Kazakhstan is adopting international standards, such as SNOMED-CT, ICD10, ICD-O-3. The cancer data obtained for this study do not provide all dataset parameters existing in the EROB, thus a proper analysis cannot be made on other parameters of database. However, the use of international standards partially were in place in the obtained data from the MOH. All the columns in the datasets are stored in Russian, including personal and medical data of patients. Although disease names, morphological type and topography of diseases are stored according to international standards, some names of them are stored in Russian language

without ICD-10, ICD-O-3 codes. Thus, datasets require human interaction in order to translate them and to make machine-actionable. Based on this, language barrier for effective data understanding and sharing might be another problem which persist in Kazakhstan. In order to make medical data available as FDP on the global level datasets need to be translated into English language using international codes.

8.1 Addressing digital health initiatives using Kingdon's Agenda Setting Model

This section will be addressing the current state of digital healthcare in Kazakhstan by Kingdon's three main streams.

8.1.1 Problem stream

Since 2013, Kazakhstan's main focus in the field of digital healthcare has been given to the formation of an integrated information environment that can serve as the basis for personalized and preventive medicine (MoH Kazakhstan, 2013). The country has begun work on the integrated data infrastructure to improve people-centred health systems. To this end, 22 health information systems provide statistics and analytics for better decision making. Although the implementation of the Interoperability Platform has been finished data remains fragmented on the level of healthcare organizations and does not interact with each other, and this caught the attention of the MOH ("State Health Development Program for 2020-2025", 2019). According to WHO in Kazakhstan (2018) (WHO, 2018), one of the top three challenges within the country is building overall capacity in handling data. The amount of data is growing rapidly in recent years, and to solve this problem there should be an emphasis on the quality of the data and it should be easily accessible for interested parties and medical personnel, and interacting with each other to better provide medical services, as well as medical research. In accordance with the MoH reforms and so far unsuccessful efforts to introduce the interoperable platform ("OECD Reviews of Health Systems", 2018), the data are still not findable, accessible, interop-

erable and, therefore, cannot be and reused by healthcare organizations and remain a problem in Kazakhstan. Resolving this in a FAIR manner would result in the saving of billions of euros (European Union, 2019).

Addressing the high burden of noncommunicable diseases is another major concern for Kazakhstan (WHO, 2018). The biggest issues are heart disease, cancer and other noncommunicable diseases (Birtanov, 2016). Among the causes of mortality in Kazakhstan, cancer mortality is in the second place (MoH & KazSRIOR, 2018). Although mortality rates are declining in comparison to previous years, cases of newly reported cancers are increasing (Kaidarova, 2019). The government has made several attempts which aimed at improving the oncological situation within the country, including the organization of oncological care, early detection of cancer diseases and reforms to address the issue. The recent Comprehensive Cancer Control Plan (2018) carries out a series of procedures to participate in international clinical research and trials to learn global practices and gain knowledge for better cancer treatment (MoH & KazSRIOR, 2018). In this regard, the willingness of Kazakhstan to gain global experience in cancer can be achieved through global exchange of cancer data using FAIR Data Principles for better diagnosis and treatment of cancer.

8.1.2 Policy stream

Most health related issues in Kazakhstan are identified by the Ministry of Health, and the Ministry of Health pursue reforms and policies to align with the national strategies (Obermann et al., 2016). To address the issue of cancer situation the MOH together with the Kazakh Research Institute of Oncology and Radiology launched the Comprehensive Cancer Control Plan for 2018-2022 years which considers participation of Kazakhstan in medical science on the global level (MoH & KazSRIOR, 2018). Thus, this may serve as the reason for the development of FDP for digital healthcare in Kazakhstan, contributing to the development of cancer treatment both in Kazakhstan and around the world.

8.1.3 Political stream

A strong political will to improve health outcomes is evident in a number of the above reforms. In this regard, the government of Kazakhstan is willing to develop health data infrastructure which might serve as a fundamental for medical care and medical research. For this research, the political stream is the digital health Concept 2013-2020, Densaulyk 2016-2019 and State Health Development Program for 2020-2025 years ("State Health Development Program for 2020-2025", 2019) endorsing interoperability of data on the national level and the Comprehensive Cancer control Plan for 2018-2022 years which is aimed at the participation of controlled studies and exchanging experience on international level (MoH & KazSRIOR, 2018). FAIR principles to be used for Kazakhstan on matters regarding cancer data sharing on an international level.

8.1.4 Policy entrepreneurs

Policy entrepreneurs might play a big role in proposing possible solutions for problems, which come from problem(s) stream. Promoting the principles of FAIR Data for digital health of Kazakhstan through the group of people, in this case, GO FAIR Ambassadors IN might be a big contribution. The IN interested in the improvement of global medical care and medical science through promotion and adoption of FAIR Data Principles.

When the three streams merge and a window of opportunity appears, then it is time for the introduction of FAIR in Kazakhstan. This should be used when the "policy window" is open. To this end, Ambassadors IN and VODAN IN members put forward their proposals and establish contacts with the health authorities and the Ministry of Health, to show how the FAIR can solve data related issues and how they can solve the problem associated with COVID-19.

In order FAIR to be reality in Kazakhstan awareness of individuals about FAIR needs to be addressed with broad reach. As long as they are well-informed and understand the changes FAIR might bring to the country their interest in FAIR may increase. However, as people in Kazakhstan are used to top-down approach for most changes, people in government need to be

involved as well. Combining both bottom-up and top-down approaches might bring significant results in addressing the attitudes towards acceptance of FAIR in Kazakhstan, therefore this can lead to the implementation of additional data policies on FAIR.

8.2 Obstacles

Since the main requirement of FDP is data being distributed, one of the main problems for Kazakhstan may be the transition from a centralized data warehouse to distributed data warehouses. Reversing such approach can be a difficult task from the point of view of all the efforts made after Concept 1.0 since 2013, the subsequent programs and funds that were spent to achieve interoperability of data. However, FAIR technology in Kazakhstan can be adopted by storing data in a centralized data repository and linking to other FDPs globally. In addition, there are also cultural barriers to transfer an organization from a protective and segmental data mentality into a way of thinking about sharing data with all relevant stakeholders (Wise et al., 2019). Disclosure of such a mentality can be achieved by combining top-level commitments with a bottom-up approach by academics (Wise et al., 2019) and in this case, policy entrepreneurs can play a significant role in changing attitudes towards the FAIR of people around government.

Another issue is related to the regulatory framework which Kazakhstan does not have at present. However, e-health legislation is being developed as part of the new e-health program for 2020-2025, so it is a matter of time. In order to implement the FAIR in Kazakhstan, it is necessary to promote the FAIR and involve people in the government with proposals that FAIR might bring to Kazakhstan, and thus, FAIR technology can comply with the regulatory framework.

8.2.1 The acceptance of technology

Data exchange standards proposed by the MoH to be used by medical organizations for data interaction are listed on the website of Electronic Health Center Kazakhstan [1]. In this regard,

[1] http://ezdrav.kz/dlya$_p ostavshikov_m is$

HL7 Clinical Document Architecture (CDA) [2] has been introduced to exchange patients' data and ensure data compatibility with the centralized database. During the internship at the Ministry of Health, it was said that acceptance of this standard by suppliers of Medical Information Systems is difficult. Thus, the lack of motivation of software developers to adapt data exchange standard can negatively affect on technology acceptance. Furthermore, healthcare organisations must see benefit of adopting a single standard for data exchanging. Therefore, the adoption of FAIR Data should be carried out very carefully, not through obligations, but through an understanding of the full value of the technology and what it can bring to society, Kazakhstan's research, global research and health.

8.3 Conclusion

For the principles of FAIR to become a reality in Kazakhstan, it is necessary to combine the efforts of various organizations, such as the government responsible for making major decisions, researchers and entrepreneurs, to make this possible and to formulate public policy based on FAIR. The opportunities presented by the FAIR principles to the healthcare in Kazakhstan are numerous. Recognition of the use of FAIR principles will depend on how they see these opportunities in terms of how easy the use will be, what impact they will have on their work and whether they will see benefits. Acceptance factors and public policy formulation can ensure the inclusion of FAIR in digital healthcare systems, where it can improve access to medical data and precision medicine.

Adopting the FAIR principles will help solve the problems of data fragmentation in Kazakhstan, which will lead to the correct use of the information value of existing data and the provision of the best treatment for patients. Therefore, FDP is relevant and necessary to address the issues related to health. Deployment of FDP is feasible from the point of view of technical readiness, and the number of health care reforms carried out aimed at improving the data infrastructure and improving health indicators in Kazakhstan. However, understanding

[2] CDA is standard for representing structured clinical documentation on patients for the purposes of health information exchange ("HL7 CDA Certification | HL7 International", 2019)

and accepting the FAIR requires a different attitude and social norms, therefore, the introduction should be carried out from the bottom up together with people in high positions, by promoting its value, which it can bring to healthcare and society in Kazakhstan.

8.4 Limitations

This study was conducted only using data on patients with breast cancer and pancreatic cancer. It was limited to these types of cancer due to the size of the project and the time given. Therefore, the addition of other types of cancer can further contribute to both global cancer research and Kazakhstan's cancer research along with cancer treatment. In addition, linking patient genomic cancer data to clinical information (description of patient symptoms, previous history) will help to develop precision oncology, which is aimed at bringing individual treatment for each patient. This helps to find out how genes determine the symptoms of the disease and provide accurate treatment. Furthermore, deploying the FDP using other types of diseases will also help to improve health indicators globally and in Kazakhstan.

APPENDIX

.1 Metadata

Ontology	Term name	Datatype	Required/Optional	Description
RDF	rdf:type	IRI	*Required*	Required to be of type r3d:Repository
DC terms	dct:title	String	*Required*	Name of the repository with the language tag
	dct:hasVersion	String	*Required*	Version of the repository
	dct:description	String	Optional	Description of the repository with the language tag
	dct:publisher	IRI	*Required*	Organisation(s) responsible for the repository
	dct:language	IRI	Optional	
	dct:license	IRI	Optional	
	dct:conformsTo	IRI	Optional	The specification of the repository metadata schema (for example ShEx)
	dct:rights	IRI	Optional	
	dct:references	IRI	Optional	Reference to documentation (API or otherwise).

FDP ontology	fdp:metadataIdentifier	IRI	*Required*	Identifier of the metadata entry. Define new sub property 'metadataID' for dct:identifier
	fdp:metadataIssued	DateTime	*Required*	Created date of the metadata entry
	fdp:metadataModified	DateTime	*Required*	Last modified date of the metadata entry
RDF Schema	rdfs:label	String	Optional	Name of the repository with the language tag
RE3Data	r3d:institution	IRI	Optional	
	r3d:startDate	DateTime	Optional	Release date of the repository

Table 1 FAIR Data Point metadata description table (**noauthor_fairdatateamfairdatapoint-spec_nodate**)

The FAIR Data Point metadata code representation (**noauthor_fairdatateamfairdatapoint-spec_nodat**

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix fdp: <http://rdf.biosemantics.org/ontologies/fdp-o#> .
@prefix lang: <http://id.loc.gov/vocabulary/iso639-1/> .
@prefix r3d: <http://www.re3data.org/schema/3-0#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
<http://136.243.4.200:8087/fdp> a r3d:Repository;
  dcterms:accessRights <http://136.243.4.200:8087/fdp/accessRights>;
  dcterms:conformsTo <http://rdf.biosemantics.org/fdp/shex/fdpMetadata>;
```

```
11    dcterms:description "This is a prototype FDP for hosting research and student projects
12    dcterms:hasVersion "1.0";
13    dcterms:language lang:en;
14    dcterms:license <http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0>;
15    dcterms:publisher <http://biosemantics.org>;
16    dcterms:title "FDP of biosemantics group";
17    fdp:metadataIdentifier <http://purl.org/biosemantics-lumc/fdp>;
18    fdp:metadataIssued "2017-05-23T09:43:15.57Z"^^xsd:dateTime;
19    fdp:metadataModified "2018-08-20T13:09:55"^^xsd:dateTime;
20    r3d:dataCatalog <http://136.243.4.200:8087/fdp/catalog/Biosamples>, <http://136.243.4.
21          <http://136.243.4.200:8087/fdp/catalog/textmining>;
22    r3d:institutionCountry <http://lexvo.org/id/iso3166/NL>;
23    r3d:repositoryIdentifier <http://136.243.4.200:8087/fdp#repositoryID>;
24    rdfs:label "FDP of biosemantics group";
25    dcterms:references <http://136.243.4.200:8087/fdp/swagger-ui.html> .
26  <http://purl.org/biosemantics-lumc/fdp> a <http://purl.org/spar/datacite/ResourceIdentif
27    dcterms:identifier "fdp" .
28  <http://biosemantics.org> a <http://xmlns.com/foaf/0.1/Organization>;
29    <http://xmlns.com/foaf/0.1/name> "Biosemantic group" .
30  <http://136.243.4.200:8087/fdp/accessRights> a dcterms:RightsStatement;
31    dcterms:description "This resource has no access restriction" .
32  <http://136.243.4.200:8087/fdp#repositoryID> a <http://purl.org/spar/datacite/Identifier
33    dcterms:identifier "176c810f-504a-421a-903b-742a70c2806a" .
```

Ontology	Term name	Datatype	Required/Optional	Description
RDF	rdf:type	IRI	*Required*	Required to be of type dcat:Catalog
DC terms	dct:title	String	*Required*	Name of the catalog with the language tag

	dct:hasVersion	String	*Required*	Version of the catalog
	dct:publisher	IRI	*Required*	Organisation(s) or Persons(s) responsible for the catalog
	dct:description	String	Optional	Description of the catalog with the language tag
	dct:language	IRI	Optional	
	dct:license	IRI	Optional	
	dct:issued	DateTime	Optional	Created date of the catalog entry
	dct:modified	DateTime	Optional	Last modified date of the catalog entry
	dct:conformsTo	IRI	Optional	The specification of the catalog metadata schema (for example ShEx)
	dct:rights	IRI	Optional	
	dct:accessRights	IRI	Optional	Description of the access rights, see Access rights rdf model
	dct:isPartOf	IRI	*Required*	Relation to the parent metadata.

FDP ontology	fdp:metadataIdentifier	IRI	Required	Identifier of the metadata entry. Define new sub property 'metadataID' for dct:identifier
	fdp:metadataIssued	DateTime	*Required*	Created date of the metadata entry
	fdp:metadataModified	DateTime	*Required*	Last modified date of the metadata entry
RDF Schema	rdfs:label	String	Optional	Name of the catalog with the language tag
FOAF	foaf:homepage	IRI	Optional	
DCAT	dcat:dataset	IRI	*Required*	List of dataset URLs
	dcat:themeTaxonomy	IRI	*Required*	List of taxonomy URLs

Table 2 Catalog metadata description table (**noauthor_fairdatateamfairdatapoint-spec_nodate**)

The catalog metadata code representation (**noauthor_fairdatateamfairdatapoint-spec_nodate**):

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix fdp: <http://rdf.biosemantics.org/ontologies/fdp-o#> .
@prefix lang: <http://id.loc.gov/vocabulary/iso639-1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
<http://136.243.4.200:8087/fdp/catalog/textmining> a dcat:Catalog;
  dcterms:accessRights <http://136.243.4.200:8087/fdp/catalog/textmining#accessRights>;
  dcterms:conformsTo <https://www.purl.org/fairtools/fdp/schema/0.1/catalogMetadata>;
```

```
  dcterms:description "Catalog for describing textmining datasets";
  dcterms:hasVersion "1.0";
  dcterms:isPartOf <http://136.243.4.200:8087/fdp>;
  dcterms:language lang:en;
  dcterms:license <http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0>;
  dcterms:publisher <http://biosemantics.org>;
  dcterms:title "Catalog for textmining datasets";
  fdp:metadataIdentifier <http://purl.org/biosemantics-lumc/fdp/catalog/textmining>;
  fdp:metadataIssued "2018-03-20T10:20:37.08Z"^^xsd:dateTime;
  fdp:metadataModified "2018-08-20T13:09:55"^^xsd:dateTime;
  rdfs:label "Catalog for textmining datasets";
  dcat:dataset <http://136.243.4.200:8087/fdp/dataset/gene_disease_association>;
  dcat:themeTaxonomy <http://dbpedia.org/resource/Text_mining>, <http://edamontology.org
<http://purl.org/biosemantics-lumc/fdp/catalog/textmining> a <http://purl.org/spar/datac
  dcterms:identifier "textmining" .
<http://biosemantics.org> a <http://xmlns.com/foaf/0.1/Agent>;
  <http://xmlns.com/foaf/0.1/name> "Biosemantic group" .
<http://136.243.4.200:8087/fdp/catalog/textmining#accessRights> a dcterms:RightsStatemen
  dcterms:description "This resource has no access restriction" .
```

Ontology	Term name	Datatype	Required/Optional	Description
RDF	rdf:type	IRI	*Required*	Required to be of type dcat:Dataset
DC terms	dct:title	String	*Required*	Name of the dataset with the language tag
	dct:publisher	IRI	*Required*	Organisation(s) or Persons(s) responsible for the dataset
	dct:hasVersion	String	*Required*	Version of the dataset

	dct:description	String	Optional	Description of the dataset with the language tag
	dct:conformsTo	IRI	Optional	The specification of the dataset metadata schema (for example ShEx)
	dct:issued	DateTime	Optional	Created date of the dataset entry
	dct:modified	DateTime	Optional	Last modified date of the dataset entry
	dct:language	IRI	Optional	
	dct:license	IRI	Optional	
	dct:rights	IRI	Optional	
	dct:accessRights	IRI	Optional	Description of the access rights, see Access rights rdf model
	dct:isPartOf	IRI	*Required*	Relation to the parent metadata.
FDP ontology	fdp:metadataIdentifier	IRI	*Required*	Identifier of the metadata entry. Define new sub property 'metadataID' for dct:identifier
	fdp:metadataIssued	DateTime	*Required*	Created date of the metadata entry

	fdp:metadataModified	DateTime	*Required*	Last modified date of the metadata entry
RDF Schema	rdfs:label	String	Optional	Name of the dataset with the language tag
DCAT	dcat:distribution	IRI	*Required*	List of distribution URLs
	dcat:theme	IRI	*Required*	List of concepts that describe the dataset
	dcat:contactPoint	IRI	Optional	
	dcat:keyword	String	Optional	Keyword(s) related to the dataset with the language tag
	dcat:landingPage	IRI	Optional	Home page of the dataset

Table 3 Dataset metadata description table (**noauthor_fairdatateamfairdatapoint-spec_nodate**)

The dataset metadata code representation (**noauthor_fairdatateamfairdatapoint-spec_nodate**):

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix fdp: <http://rdf.biosemantics.org/ontologies/fdp-o#> .
@prefix lang: <http://id.loc.gov/vocabulary/iso639-1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://136.243.4.200:8087/fdp/dataset/gene_disease_association> a dcat:Dataset;
  dcterms:accessRights <http://136.243.4.200:8087/fdp/dataset/gene_disease_association#a
```

```
11    dcterms:conformsTo <https://www.purl.org/fairtools/fdp/schema/0.1/datasetMetadata>;
12    dcterms:description "High-throughput experimental methods such as medical sequencing a
13    dcterms:hasVersion "1.0";
14    dcterms:isPartOf <http://136.243.4.200:8087/fdp/catalog/textmining>;
15    dcterms:language lang:en;
16    dcterms:license <http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0>;
17    dcterms:publisher <http://biosemantics.org>;
18    dcterms:title "Gene disease association (LUMC)";
19    fdp:metadataIdentifier <http://purl.org/biosemantics-lumc/fdp/dataset/gene_disease_ass
20    fdp:metadataIssued "2018-03-20T10:30:18.662Z"^^xsd:dateTime;
21    fdp:metadataModified "2018-08-20T13:09:55"^^xsd:dateTime;
22    rdfs:label "Gene disease association (LUMC)";
23    dcat:distribution <http://136.243.4.200:8087/fdp/distribution/gene_disease_association
24      <http://136.243.4.200:8087/fdp/distribution/gene_disease_association_html>, <http://
25    dcat:keyword "GDA", "Gene disease association (LUMC)", "LWAS", "Text mining", "The Exp
26      "The Implicitome";
27    dcat:theme <http://dbpedia.org/resource/Text_mining>, <http://semanticscience.org/reso
28  <http://purl.org/biosemantics-lumc/fdp/dataset/gene_disease_association> a <http://purl.
29    dcterms:identifier "gene_disease_association" .
30  <http://biosemantics.org> a <http://xmlns.com/foaf/0.1/Agent>;
31    <http://xmlns.com/foaf/0.1/name> "Biosemantic group" .
32  <http://136.243.4.200:8087/fdp/dataset/gene_disease_association#accessRights> a dcterms:
33    dcterms:description "This resource has no access restriction" .
```

Ontology	Term name	Datatype	Required/Optional	Description
RDF	rdf:type	IRI	*Required*	Required to be of type dcat:Distribution

DC terms	dct:title	String	*Required*	Name of the data distribution with the language tag
	dct:conformsTo	IRI	Optional	The specification of the distribution metadata schema (for example ShEx)
	dct:license	IRI	*Required*	Link to the license description
	ddct:hasVersion	String	*Required*	Version of the distribution
	dct:issued	DateTime	Optional	Created date of the distribution entry
	dct:modified	DateTime	Optional	Last modified date of the distribution entry
	dct:rights	IRI	Optional	
	dct:description	String	Optional	Description of the description with the language tag
	dct:accessRights	IRI	Optional	Description of the access rights, see Access rights rdf model
	dct:isPartOf	IRI	*Required*	Relation to the parent metadata.

FDP ontology	fdp:metadataIdentifier	IRI	*Required*	Identifier of the metadata entry. Define new sub property 'metadataID' for dct:identifier
	fdp:metadataIssued	DateTime	*Required*	Created date of the metadata entry
	fdp:metadataModified	DateTime	*Required*	Last modified date of the metadata entry
RDF Schema	rdfs:label	String	Optional	Name of the data distribution with the language tag
DCAT	dcat:accessURL	IRI	*Required*	A landing page, feed, SPARQL endpoint or other type of resource that gives access to the distribution of the dataset
	dcat:downloadURL	IRI	*Required*	A file that contains the distribution of the dataset in a given format
	dcat:mediaType	String	*Required*	The media type of the distribution
	dcat:format	String	Optional	
	dcat:byteSize	Decimal	Optional	

	dcat:keyword	String	Optional	Keyword(s) related to the dataset with the language tag
	dcat:landingPage	IRI	Optional	Home page of the dataset

Table 4 Distribution metadata description table (**noauthor_fairdatateamfairdatapoint-spec_nodate**)

The distribution metadata code representation (**noauthor_fairdatateamfairdatapoint-spec_nodat**

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix fdp: <http://rdf.biosemantics.org/ontologies/fdp-o#> .
@prefix lang: <http://id.loc.gov/vocabulary/iso639-1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://136.243.4.200:8087/fdp/distribution/gene_disease_association_nquads_gzip>
  a dcat:Distribution;
  dcterms:accessRights <http://136.243.4.200:8087/fdp/distribution/gene_disease_associat
  dcterms:conformsTo <https://www.purl.org/fairtools/fdp/schema/0.1/distributionMetadata
  dcterms:description "The complete set of all ~204 million associations (explicit and i
  dcterms:hasVersion "1.0";
  dcterms:isPartOf <http://136.243.4.200:8087/fdp/dataset/gene_disease_association>;
  dcterms:language lang:en;
  dcterms:license <http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0>;
  dcterms:publisher <http://biosemantics.org>;
  dcterms:title "Gene disease association (LUMC) nquads as gzip distribution";
```

```
20    fdp:metadataIdentifier <http://purl.org/biosemantics-lumc/fdp/distribution/gene_diseas
21    fdp:metadataIssued "2018-03-20T10:40:17.677Z"^^xsd:dateTime;
22    fdp:metadataModified "2018-08-20T13:09:55"^^xsd:dateTime;
23    rdfs:label "Gene disease association (LUMC) nquads as gzip distribution";
24    dcat:downloadURL <https://datadryad.org/bitstream/handle/10255/dryad.91060/gda-np.nq.g
25    dcat:mediaType "application/gzip" .
26  <http://purl.org/biosemantics-lumc/fdp/distribution/gene_disease_association_nquads_gzip
27    a <http://purl.org/spar/datacite/ResourceIdentifier>;
28    dcterms:identifier "gene_disease_association_nquads_gzip" .
29  <http://biosemantics.org> a <http://xmlns.com/foaf/0.1/Agent>;
30    <http://xmlns.com/foaf/0.1/name> "Biosemantic group" .
31  <http://136.243.4.200:8087/fdp/distribution/gene_disease_association_nquads_gzip#accessR
32    a dcterms:RightsStatement;
33    dcterms:description "This resource has no access restriction" .
```

REFERENCES

Abishev, O. (2018). Speech of the director of republican center for e-health on the day of digitalization in kazakhstan. Retrieved May 27, 2019, from https://digitalization.astanahub.kz/projects/64

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Alfonso, F., Adamyan, K., Artigou, J.-Y., Aschermann, M., Boehm, M., Buendia, A., Chu, P.-H., Cohen, A., Dei Cas, L., Dilic, M., Doubell, A., Echeverri, D., Enç, N., Ferreira-González, I., Filipiak, K. J., Flammer, A., Fleck, E., Gatzov, P., Ginghina, C., ... Lüscher, T. F. (2017). Data Sharing: A New Editorial Initiative of the International Committee of Medical Journal Editors. Implications for the Editors' Network. *Revista Portuguesa de Cardiologia*, *36*(5), 397–403. https://doi.org/10.1016/j.repc.2017.02.001

Birkland, T. A. (1998). Focusing Events, Mobilization, and Agenda Setting [Cambridge University Press]. *Journal of Public Policy*, *18*(1), 53–74. https://doi.org/10.1017/S0143814X98000038

Birtanov, Y. (2016). Kazakhstan gears up to launch social health insurance. *Bulletin of the World Health Organization*, *94*(11), 792–793. https://doi.org/10.2471/BLT.16.031116

Birtanov, Y. (2019). Electronic health passport available on eGov and mGov platforms. Retrieved April 5, 2020, from http://www.government.kz/en/news/press/elektronnyy-pasport-zdorovya-mozhno-posmotret-na-platformah-egov-i-mgov-e-birtanov

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data: The Story so Far. *International Journal on Semantic Web and Information Systems*, *5*, 1–22. https://doi.org/10.4018/jswis.2009081901

Bowen, D. J., Kreuter, M., Spring, B., Cofta-Woerpel, L., Linnan, L., Weiner, D., Bakken, S., Kaplan, C. P., Squiers, L., Fabrizio, C., & Fernandez, M. (2009). How we design feasibility studies. *American Journal of Preventive Medicine*, *36*(5), 452–457. https://doi.org/10.1016/j.amepre.2009.02.002

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M., Sherry, S., & Flicek, P. (2012). The

1000 Genomes Project: Data management and community access. *Nature Methods*, *9*(5), 459–462. https://doi.org/10.1038/nmeth.1974

The Constitution of Kazakhstan. (1998). Retrieved April 5, 2020, from http://www.akorda.kz/en/official_documents/constitution

Densaulyk state program for 2016-2019 years. (2016). Retrieved April 5, 2020, from https://primeminister.kz/ru/documents/gosprograms/gosudarstvennaya-programma-razvitiya-zdravoohraneniya-respubliki-kazahstan-densaulyk-na-2016-2019-gody

European Union. (2013). Semantic interoperability for better health and safer healthcare: Deployment and research roadmap for Europe. Retrieved October 18, 2019, from https://op.europa.eu:443/en/publication-detail/-/publication/9bb4f083-ac9d-47f8-ab4a-76a1f095ef15/language-en/format-PDF

European Union. (2019). Cost-benefit analysis for FAIR research data : Cost of not having FAIR research data. [ISBN: 9789279988868 Publisher: Publications Office of the European Union]. Retrieved April 9, 2020, from http://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1

FAIRDataTeam. (2020). FAIRDataPoint-Specification. Retrieved February 4, 2020, from https://github.com/FAIRDataTeam/FAIRDataPoint-Spec

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behaviour: An introduction to theory and research* (Vol. 27).

Gagnon, M. L., & Labonté, R. (2013). Understanding how and why health is integrated into foreign policy - a case study of health is global, a UK Government Strategy 2008–2013. *Globalization and Health*, *9*(1), 24. https://doi.org/10.1186/1744-8603-9-24

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591

Gurdasani, D., Barroso, I., Zeggini, E., & Sandhu, M. S. (2019). Genomics of disease risk in globally diverse populations. *Nature Reviews Genetics*, *20*(9), 520–535. https://doi.org/10.1038/s41576-019-0144-0

Herstatt, C., & Verworn, B. (2004). The 'Fuzzy Front End' of Innovation. https://doi.org/10.1057/9780230512771_16

Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges, *34*(1), 135–174. https://doi.org/10.1111/dpr.12142

HL7 CDA Certification | HL7 International. (2019). Retrieved April 12, 2020, from https://www.hl7.org/certification/cda.cfm?ref=nav

Implementation results of Densaulyq State Program, modern medical technologies, social insurance, or how healthcare system of Kazakhstan is improving. (2020). Retrieved

April 13, 2020, from https://primeminister.kz/en/news/reviews/implementation-results-of-densaulyq-state-program-modern-medical-technologies-social-insurance-or-how-healthcare-system-of-kazakhstan-is-improving

Jensen, M. A., Ferretti, V., Grossman, R. L., & Staudt, L. M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, *130*(4), 453–459. https://doi.org/10.1182/blood-2017-03-735654

Kaidarova, D. (2019). Oncology service in the Republic of Kazakhstan: Results and prospects. Retrieved April 11, 2020, from http://pharmnews.kz/ru/article/onkologicheskaya-sluzhba-v-rk-itogi-i-perspektivy_14840

Kassen, M. (2017). Open data in Kazakhstan: Incentives, implementation and challenges. *Information Technology & People*, *30*(2), 301–323. https://doi.org/10.1108/ITP-10-2015-0243

Kassen, M. (2019). Open Data Politics in Kazakhstan: Understanding a Tentative Advance of Civic Engagement in a Transitional Society. *Open Data Politics, DOI: 10.1007/978-3-030-11410-7_4, pp.69-98*. Retrieved April 5, 2020, from https://www.academia.edu/38276948/Open_Data_Politics_in_Kazakhstan_Understanding_a_Tentative_Advance_of_Civic_Engagement_in_a_Transitional_Society

Kingdon, J. W. (2013). *Agendas, alternatives, and public policies, update edition, with an epilogue on health care* (2nd [upd.] ed.). Harlow, United Kingdom, Pearson Education Limited.

Kingdon, J. (1995). *Agendas, Alternatives, and Public Policies*. HarperCollins College Publishers. https://books.google.kz/books?id=_gmSQgAACAAJ

Law on Access to Information. (2015). Retrieved April 5, 2020, from http://adilet.zan.kz/eng/docs/Z1500000401

Law on Personal Data and their Protection. (2013). Retrieved April 5, 2020, from http://adilet.zan.kz/eng/docs/Z1300000094

Liebert, S., Gondrey, S., & Goncharov, D. (2013). Public Administration in Post-Communist Countries: Former Soviet Union, Central and Eastern Europe, and Mongolia [Library Catalog: www.routledge.com]. Retrieved April 16, 2020, from https://www.routledge.com/Public-Administration-in-Post-Communist-Countries-Former-Soviet-Union/Liebert-Condrey-Goncharov/p/book/9781439861370

MoH Kazakhstan. (2013). E-health development concept for 2013-2020. Retrieved April 11, 2020, from http://www.rcrz.kz/index.php/ru/kontseptsiya-razvitiya-elektronnogo-zdravookhraneniya

MoH, & KazSRIOR. (2018). Comprehensive cancer control plan for 2018-2022 years. Retrieved April 5, 2020, from https://onco.kz/news/razrabotan-proekt-kompleksnogo-plana-po-borbe-s-onkologicheskimi-zabolevaniyami-na-2018-2022-gody/

Momynaliev, K., & Imanbekova, M. (2014). The need for standardized biobanks in Kazakhstan. *Central Asian Journal of Global Health*, *2*(Suppl). https://doi.org/10.5195/cajgh.2013.99

Mons, B. (2018). Data Stewardship for Open Science: Implementing FAIR Principles. Retrieved April 3, 2020, from https://www.crcpress.com/Data-Stewardship-for-Open-Science-Implementing-FAIR-Principles/Mons/p/book/9780815348184

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, *37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Murphy, S., & Kumar, V. (2002). The Front End of New Product Development: A Canadian Survey. *R&D Management*, *27*, 5–15. https://doi.org/10.1111/1467-9310.00038

Nazarbayev, N. (2012). Address by the President of Kazakhstan. Retrieved April 5, 2020, from https://www.inform.kz/en/article/2346141

Obermann, K., Chanturidze, T., Richardson, E., Tanirbergenov, S., Shoranov, M., & Nurgozhaev, A. (2016). Data for development in health: A case study and monitoring framework from Kazakhstan. *BMJ global health*, *1*(1), e000003. https://doi.org/10.1136/bmjgh-2015-000003

OECD Reviews of Health Systems: Kazakhstan 2018 | READ online. (2018). Retrieved November 8, 2019, from https://read.oecd-ilibrary.org/social-issues-migration-health/oecd-reviews-of-health-systems-kazakhstan-2018_9789264289062-en

Pavlopoulou, A., Spandidos, D. A., & Michalopoulos, I. (2015). Human cancer databases (Review). *Oncology Reports*, *33*(1), 3–18. https://doi.org/10.3892/or.2014.3579

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature News*, *538*(7624), 161. https://doi.org/10.1038/538161a

Quirk, P. (1986). Agendas, alternatives and public policies. *Journal of Policy Analysis and Management*, *5*(3), 607–613. http://search.proquest.com/docview/1761689531/

RCRZ. (2018). Kazakhstan medicine will become "smart"». Retrieved April 20, 2020, from https://inbusiness.kz/ru/news/kazahstanskaya-medicina-stanet-%5C%C2%ABumnoj%5C%C2%BB

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., & Thurston, M. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, *37*(4), 358–367. https://doi.org/10.1038/s41587-019-0080-8

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048

Siu, L. L., Lawler, M., Haussler, D., Knoppers, B. M., Lewin, J., Vis, D. J., Liao, R. G., Andre, F., Banks, I., Barrett, J. C., Caldas, C., Camargo, A. A., Fitzgerald, R. C., Mao, M., Mattison, J. E., Pao, W., Sellers, W. R., Sullivan, P., Teh, B. T., ... Voest, E. E. (2016). Facilitating a culture of responsible and effective sharing of cancer genome data. *Nature Medicine*, *22*(5), 464–471. https://doi.org/10.1038/nm.4089

State Health Development Program for 2020-2025. (2019). Retrieved April 11, 2020, from http://adilet.zan.kz/rus/docs/P1900000982#z17

The State program "Information Kazakhstan - 2020". (2013). Retrieved October 26, 2019, from https://tengrinews.kz/zakon/prezident_respubliki_kazahstan/hozyaystvennaya_deyatelnost/id-U1300000464/

Sustkova, H. P., Hettne, K. M., Wittenburg, P., Jacobsen, A., Kuhn, T., Pergl, R., Slifka, J., McQuilton, P., Magagna, B., Sansone, S.-A., Stocker, M., Imming, M., Lannom, L., Musen, M., & Schultes, E. (2019). FAIR Convergence Matrix: Optimizing the Reuse of Existing FAIR-Related Resources [Publisher: MIT Press]. *Data Intelligence*, *2*(1-2), 158–170. https://doi.org/10.1162/dint_a_00038

The World Bank. (2017). Health Sector Technology Transfer and Institutional Reform. Retrieved April 11, 2020, from https://projects.worldbank.org/en/projects-operations/project-detail/P101928

van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A. O., Kirkpatrick, C., & Mons, B. (2019). Towards the Tipping Point for FAIR Implementation. *Data Intelligence*, *2*(1-2), 264–275. https://doi.org/10.1162/dint_a_00049

van Soest; Oliver Kohlbacher; Lukas Zimmermann; Holger Stenzhorn; Md. Rezaul Karim; Michel Dumontier; Stefan Decker; Luiz Olavo Bonino da Silva Santos; Andre Dekker, O. B. A. C. J. (2020). Distributed analytics on sensitive medical data: The personal health train. *Data Intelligence*, *2*, 96–107. https://doi.org/10.1162/dint_a_00032

Vesteghem, C., Brøndum, R. F., Sønderkær, M., Sommer, M., Schmitz, A., Bødker, J. S., Dybkær, K., El-Galaly, T. C., & Bøgsted, M. (2019). Implementing the FAIR Data Principles in precision oncology: Review of supporting initiatives. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbz044

WHO. (2004). ICD-10 : International statistical classification of diseases and related health problems : Tenth revision, 2nd ed. world health organization. Retrieved April 5, 2020, from https://apps.who.int/iris/handle/10665/42980

WHO. (2013). WHO | International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3). Retrieved April 5, 2020, from https://apps.who.int/iris/handle/10665/96612

WHO. (2018). World Health Organization in Kazakhstan (2018) [Library Catalog: www.euro.who.int Publisher: World Health Organization]. Retrieved April 11, 2020, from http://

www.euro.who.int/en/countries/kazakhstan/publications/world-health-organization-in-kazakhstan-2018

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness, *5*. https://doi.org/10.1038/sdata.2018.118

Wilson, W. (1993). *Sociology and the Public Agenda*. Thousand Oaks, California. https://doi.org/10.4135/9781483325484

Wise, J., de Barron, A. G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., Mellino, G., Harrow, I., Smith, I., Taubert, J., van Bochove, K., Romacker, M., Walgemoed, P., Jimenez, R. C., Winnenburg, R., Plasterer, T., Gupta, V., & Hedley, V. (2019). Implementation and relevance of FAIR data principles in biopharmaceutical R&D, *24*(4), 933–938. https://doi.org/10.1016/j.drudis.2019.01.008

Zhabagin, M. (2018). How digitization of the genome can change our ideas about the world and about ourselves. Retrieved April 13, 2020, from https://biocenter.kz/en/news/media-about-us/735