

Care2Report: Data management for a multimodal corpus

Master thesis

Author: Brandon Koffijberg BSc (b.koffijberg@students.uu.nl), 6312810

Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands.



Utrecht University

Submitted in fulfillment of the requirements for the degree of: Master of Business Informatics.

Date: Utrecht, February 7, 2020
Version: 2.0 (Final version)

First supervisor:
Prof. dr. Sjaak Brinkkemper

Second supervisor:
Dr. G. Wagenaar

Daily supervisor:
Ph.D. Candidate Sabine Molenaar

Acknowledgements

First and foremost I would like to thank my 1st and 2nd supervisors for all their support during this research. I would like to thank Prof. Dr. Sjaak Brinkkemper for all the useful feedback, guiding, meetings and personal interest. I want to thank dr. G. Wagenaar for his enthusiasm about this research, as well for the provided comment and discussions. Secondly, a special thanks to Sandra van Dulmen and Otto Lange for their time and needed answers during the interview. Thirdly, I want to thank Sabine Molenaar and all the researchers within C2R program for their feedback during the group meetings. They provided me helpful knowledge. Finally, I would like to thank my parents and my girlfriend for all their support during my whole study and master thesis.

Table of contents

CHAPTER 1 INTRODUCTION	1
1.1 INTRODUCTION RESEARCH BACKGROUND	1
1.2 PROBLEM STATEMENT	2
1.3 RESEARCH OBJECTIVE	4
1.4 RESEARCH SCOPE.....	5
1.5 RESEARCH QUESTIONS	6
1.6 RELEVANCE	8
1.7 RESEARCH OUTLINE.....	8
CHAPTER 2 RESEARCH METHOD	9
2.1 DESIGN SCIENCE – RESEARCH FRAMEWORK	9
2.2 DESIGN CYCLE – RESEARCH APPROACH.....	9
2.2.1 <i>Literature research protocol</i>	10
2.2.2 <i>Semi-structured interview</i>	12
2.2.3 <i>Case study</i>	13
2.2.4 <i>Comparison golden standard</i>	13
2.3 VALIDITY & RELIABILITY.....	13
2.4 RESEARCH EXECUTION	14
CHAPTER 3 LITERATURE ON RESEARCH DATA MANAGEMENT	15
3.1 BASIC TERMINOLOGY.....	16
3.1.1 <i>Data terminology and definitions</i>	16
3.1.2 <i>Multimodal VS Multimedia</i>	17
3.2 FROM MULTIMODAL SPEECH AND VIDEO INPUT TO TEXT IN C2R.....	20
3.3 MULTIMODAL ANALYSIS & TESTING.....	24
3.4 DATA MANAGEMENT	26
3.4.1 <i>DAMA Framework</i>	26
3.5 RESEARCH DATA MANAGEMENT	29
3.5.1 <i>DMC Framework</i>	30
3.5.2 <i>Research data management in a larger institution</i>	32
3.6 STORAGE IN A MULTIMODAL CORPUS.....	34
3.6.1 <i>Traditional vs multimodal database</i>	34
3.6.2 <i>Metadata for multimodal storage</i>	37
3.6.3 <i>Designing issues multimodal database</i>	38
3.7 TESTING ENVIRONMENT FOR SOFTWARE TESTING	39
CHAPTER 4 DATA AND SET-UP	41
4.1 REQUIREMENTS SPECIFICATION	41
4.2 POLICIES, CODES OF CONDUCT AND LAWS.....	43
4.2.1 <i>Rules and processes for testing privacy sensitive data</i>	45
4.2.2 <i>Using sensitivity test data from real patients</i>	46
4.2.3 <i>Rules and processes of using research data within the University</i>	47
4.3 DATA CREATION STEPS SET-UP, PROCESS AND RECORDING OF THE MEDICAL CONSULTATIONS	48
4.3.1 <i>Recording medical consultation</i>	48
4.3.2 <i>Creating test data</i>	50
4.3.4 <i>Creating transcript of the recordings</i>	52
4.3.5 <i>Finding important words and sentence structure</i>	53
4.4 DATABASE DESIGN IN YODA.....	56
4.4.1 <i>Metadata structure</i>	58
4.4.2 <i>Data management plan - (DCC)</i>	60
CHAPTER 5 CASE STUDIES	62
5.1 EXECUTION C2R SOFTWARE WITH SIMULATED AND REAL MEDICAL CONSULTATIONS	62

5.1.1 Modeling the software environment	62
5.1.2 Selecting test scenarios.....	63
5.1.3 Running and evaluation the test scenarios.....	64
5.1.4 Create a ground truth	69
5.2 COMPARISON OF THE C2R AND THE REAL CONSULTATION SOAP	69
5.2.1 Explanation of the results	71
5.3 ANALYZING THE SENTENCES AND STRUCTURE USED FOR TESTING	74
5.3.1 Irrelevant sentences and word blacklist	74
5.3.2. All the important sentences marked in the transcript with RIAS	75
5.3.3 Structure of the consultation	76
5.4 FUNCTIONAL DESIGN DATA EXTRACTION	77
5.5 STORAGE OF TRAINING, TESTING AND VALIDATION DATA FOR ACTION RECOGNITION	82
CHAPTER 6 VALIDATION OF THE GENERATED REPORT	83
CHAPTER 7 DISCUSSION & CONCLUSION	84
7.1 CONCLUSION	84
7.2 LIMITATIONS.....	86
7.3 FUTURE RESEARCH	87
7.4 IMPORTANT FINDINGS	87
REFERENCES	88
APPENDIX.....	92

List of Figures

Figure number	Figure description
Figure 1	Process overview C2R
Figure 2	Process storing experiment results on Yoda
Figure 3	Overview research questions within the research
Figure 4	Iteration problem solving in design science research
Figure 5	Design cycle
Figure 6	Backward snowballing approach
Figure 7	Overview of the Microservice Architecture Care2Report system
Figure 8	Example of text to formal representation
Figure 9	Life cycle for research data
Figure 10	Data Management and Curation
Figure 11	University storage network
Figure 12	Hierarchically organized storage for multimedia databases
Figure 13	Setup recording real situation
Figure 14	Setup recording simulated situation
Figure 15	comparison general consultation structure vs RIAS code made by van der Kooi & Lim (2019)
Figure 16	Comparison between SOAP and Rias made by van der Kooi & Lim (2019)
Figure 17	Overview of the current Yoda layout
Figure 18	The metadata from construction of Yoda
Figure 19	Overview of the videos within a sub-folder
Figure 20	Example of a generated report by the C2R prototype based on a consultation transcript
Figure 21	Route of the most used microanalyzers within C2R
Figure 22	EMR created by the general practitioner consultation 1
Figure 23	EMR created by the general practitioner consultation 2
Figure 24	Steps for the creation of the blacklist
Figure 25	Overview of the relation between words from a medical consultation
Figure 26	Birds eye view of the functional technical architecture
Figure 27	Functional design for extracting of data, results and logs
Figure 28	Overview of all the data that must be gathered in the testing environment
Figure 29	The position of the C# module in the testing environment
Figure 30	Ontology as used in the current prototype
Figure 31	The position of the C# module in the testing environment

List of tables

Table 1	Comparison of the terms multimodal vs multimedia
Table 2	Salient features of traditional - and multimodal databases (partially reprinted and adjust from “Multimedia databases”)
Table 3	The interfaces used by testers for testing software.
Table 4	Overview of the recordings
Table 5	Correct input structure of the transcript in the C2R system
Table 6	Error in interpret the transcript by the C2R system
Table 7	Comparison general structure and SOAP for a medical consultation
Table 8	Placing patient or doctor in the transcript
Table 9	Examples of the triples created by the microanalyzers
Table 10	Placing patient or doctor in the transcript
Table 11	Examples of the triples created by the microanalyzers
Table 12	Number of items included for each section of the SOEP format, with TPs/FPs/FNs for the generated reports.
Table 13	Analysis of relevance and completeness of generated reports

List of Abbreviations

C2R	Care2Report
CBS	Statistics Netherlands
EMR	Electronic Medical Record
fte	Fulltime-equivalent
CDM	Clinical Data Management
CRF	Case Report Forms
SCDM	Society for Clinical Data Management
HCI	Human-Computer-Interaction
DMC	Data Management and Curation
UU	Utrecht University
WGBO	Wet Geneeskundige Behandelovereenkomst
FAIR	Findable, Accessible, Interoperable and Reusable
DAMA	Data Management Association

Chapter 1 Introduction

This research project is part of a larger research program called Care2Report (C2R). The aim of the program is to develop a software platform for automated medical reporting to remove writing and typing to make notes during or after a(n) interaction/consultation between humans. By using speech recognition, action recognition technology and Bluetooth measurement to automate medical reporting. The C2R program will be executed by PhD, Master and Bachelor students. Every research project generates new insights and knowledge for the C2R system. The current project is focused on the rules, processes and policies around the data management from testing data and the storage of research data. The C2R system will reduce the writing time for professionals in the field. The system is focused on the medical field. When the medical field can be supported with the system, other fields will be explored such as vets and national police. The test data that is made for the test environment will be medical data only. In this chapter the research background, problem statement and outline of this research will be presented.

1.1 Introduction Research background

Research data management is researched for the validation of the system C2R. The validation of the system that will be developed is tested by making use of testing data. The test data that is used will be managed.

Data management has been a research topic for decades, there are papers that date from 1967. Before computers data management comprised maintaining files and keep them up to date. The new information was added in the files that were stored so it could be re-used when needed. This was done for either the business, a library or academic research (Franks, 1967).

Data management is in these days still a widely discussed domain. Data management and specifically the technology that makes data management possible. Scientific data management is creating massive data stores to analyze and organize the data. This is needed because the volume of the data is almost doubling every year. All the new technologies and used instruments are also improving the data quality in a high tempo. The traditional database is lacking in supporting the scientific types of data (Gray et al., 2006).

Data management is about data, not only in text or numbers but also images, video and audio. Where it used to be text in files, can it be these days speech, video, audio or a combination. Katz et al. (2002) state to get a full understanding of this type of media areas of for example artificial intelligence, object recognition and speech transcription must be used. It is needed to capture the essence of the information in that type of documents (Katz, Lin, & Felshin, 2002). Scientific trail and instruments used in those trails creates lots of data. A scientist collect data from participants, this data must be carefully stored, organized accessed and documented. Metadata could manage this descriptive information about the data. In the description are the measurements, the way the measurement is executed, data origin, units and the layout of the data etc. (Gray et al., 2006). The enormous growth of information and the availability in multiple multimodal forms as text, image etc. give people more access to more information than before (Katz et al., 2002). There is a difference between multimodal and multimedia, where multimedia focus on information presentation, is multimodal more focus on the interpretation of different media (Lee, 1996). In the existing literature the terms multimodal and multimedia are sometime used interchangeably. Where multidata is, there is a multidata store. To store multimodal data there is a multimodal database needed. Multimodal databases are mostly used in machine learning/ artificial intelligence. Within the topic multi-modal storage, we found a gap in the literature. Multimodal storage is mostly focused on datasets for machine learning.

The current situation of research data management is that, the growth at this moment in research worldwide is increasing the requiring and transparency for developing data management plans. This is for preparing data to be shared and get public access. Before this is possible the sensitive data must be selected, managed and stored in a manner that the data is protected (Perrier & Barnes, 2018). (not all countries) But countries such as United States and United Kingdom have government level policies for research data management. The document will describe the expectations and responsibilities for the data management plan and sharing the research data. The document is used for major funding of a research project.

The way of storing in research data management by using a survey with 286 respondents (Anderson et al., 2007). In Biomedical Research Data Management, most of the researchers (59%) store their multimodal data digitally and approximately 34% still store their data hard copy. In the interviews for this survey two main themes were clear. The theme that is relevant for this research is, the difficulty to organize, store and retrieve research data. Non-specialized applications are used to store data such as text files, spreadsheets and shared files. The reasons of using these types of applications are not in favor of data management. It is because of the easy availability, easy to use and the simplicity of the layout. When the question was raised about what the most urgent problem with digital data is, more then 28 present answered (of 82 respondents) “*long-term digital archiving of a variety of data types*” (Anderson et al., 2007). The research stated additional challenge of storing sensitive data that there is a need to improve the current way of storing research data organization wide.

The gap in the existing knowledge is within the literature of the broadly discussed topic data management. The first paper of data management dates from 1967 all the way until 2019. The term is used in the context of (large) organizations, or between units within a company. There are no specific papers about multimodal databases for medical systems to our knowledge.

1.2 Problem statement

During a medical consultation there are several problematic moments for the care provider and for the patient concerning the documentation of patient medical information in the electronic medical record (EMR). The inconveniences have impact on the productivity and job satisfaction of the care provider. The patient does not get the full attention and time. Three issues regarding documentation in healthcare are discussed below.

Issues documentation time

In addition to providing care to a patient, medical professionals also have administrative responsibilities. Think of noting down the test results, given treatments and the physical complaints of the patient. All the diagnoses and given treatments must be documented in a patient personal file. In the EMR the history of the patient documented. But writing all the patient information takes up to three hours of their time to administrate, this is time they could not help a patient. Within the Netherlands 70% of the care professionals within the sample experience creating and maintaining the EMR as a burden (Hanekamp, 2016). The medical sector, like other sectors, has a shortage of personnel. Because of the lack of medical personnel specialized as doctors, nurses etc. and the increasing stream of patients they are experiencing a higher work pressure. (Bierings, 2017) shows that being a doctor is experienced as one of the most stressful professions. Not only the workload and high work pressure are a problem, also the cost of the health care administration. In 2018, a survey was conducted in the Netherlands about administrative load on care professionals from 21 organizations in three kinds of medical sectors stated that the actual time spending on administrative tasks is 31%. The acceptable time for spending on administration is 17%. When this could be reduced by 14% it would create a savings of 50.000 fte or 2,5 billion euro. This will make a significant impact on the medical shortage, because the estimation on 2025 the medical personnel shortage of 125.000 care professionals (Berenschot and VU, 2018). This corresponds to 110.000 full time work positions per year in the Netherlands alone. There is an increasing trend visible, two years before it was 25% of their time. The envisioned C2R software could have a massive impact:

when there could be a decline of 14% of the administrative workload this could result in a savings of 2,5 billion euros (Hanekamp, 2016).

Declining documentation quality

To keep waiting time to a minimum, medical personnel will type as fast as they can during the consult. This could cause also errors, typos and could encourage abbreviations. Medical abbreviations are used many times in a note, from a set of 169 notes the care provider used 3668 abbreviations with 479 different abbreviations (Sheppard, Weidner, Zakai, Fountain-Polley, & Williams, 2008). The abbreviations are confusing and only half of the abbreviations are understood by another medical professional when reading the notes. Not only the written communication becomes difficult or ambiguous, also speech communication. The doctors use abbreviations and medical jargon to patients for explaining and informing (Walsh & Gurwitz, 2008). The reason to use abbreviations is the high administrative tasks the medical personal point out. Nivel say that two out of three care providers in the Netherlands indicate that their administrative task is too high. Their report presents the numbers that they score above average (Bierings, 2017). The high work pressure affects the documentation quality. It is not possible to execute an analysis on documentation with poor quality.

Problems patient experience

The reason medical professionals want to work in health care, is not because of the administrative tasks. They want to work with people and help them. Medical professionals will give the patients the care and treatment they deserve. By taking the writing time away medical personnel could more focus on the consult, spend more time on helping the patient. The total duration of the consult will decline, less time means cost saving, more patients on a day means more revenue. A doctor that could handle more patients, is also a (partly) solution for the shortage of medical personnel. A number of 90 % experience administrative tasks as pressure (*Berenschot and VU, 2018*). This could result in unsatisfied employees. The writing also affects the waiting time of a patients to see/speak with their doctor. Patient- doctor communication in health care about the patient is essential. Many health care practitioners are concerned about the impact of the EMR on the communication between patient and their doctor. The EMR gathered and report sharable health care information about the patient. The drawback is the doctor's behavior, it has impact on the communication. Non-verbal behavior of the doctor influence if the patient gives certain information or not. Eye contact is import for the patient to feel taken care of by their doctor. The EMR influenced the eye contact between the patient (Rathert, Mittler, Banerjee, & McDaniel, 2017).

The proposed solution

So, there are many reasons to create a solution for the administrative burden for health care professionals. The reason the C2R system is built, is because there are many problems, difficulties and frustrations around the administrative work of medical personnel. The solution is to create an automated medical report system using speech and action recognition technology. This will help reduce their administrative tasks during and after a consultation. The C2R system will make use of a movement by translating the movement to text. The speech of the patient and a doctor will be translated by using speech to text and making use of Bluetooth by using measurements to text.

The input of the Care2Report system is provided in three modalities (see also figure 1)

Video: The movement, treatments and procedures of the medical professional, will be translated into text.

Audio: The designation, recommendation and explanations that are discussed will be turned in to text with speech to text.

Sensor: The measurement that is made by making use of equipment and measurement devices will be transferred with Bluetooth and will be documented (Maas, et al., 2019).

The medical recording is with speech and action recognition converted to text in a transcript. The transcript will be searched for triples, semantic triples ((subject, predicate, object)) (Rohloff, Dean, Emmons, Ryder, & Sumner, 2007). The triples are extracted from the text to the knowledge graph and matched with the ontologies to create matched triples. The ontology will extracted matched triples from the knowledge graph. Rule-based algorithm match the triples and ontology (Antoniou & Harmelen, 2004). From the transcript is natural text generated to fill in the EMR report with relevant sentences (Maas, et al., 2019).

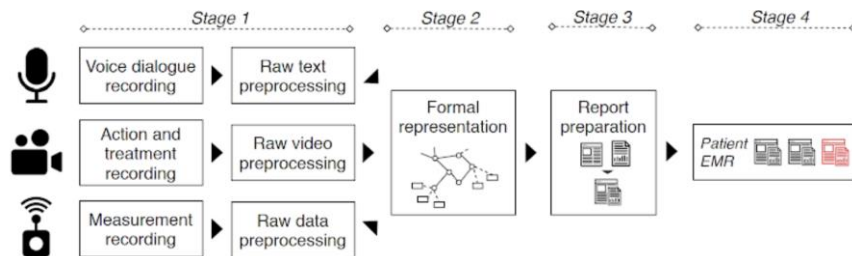


Figure 1 Process overview C2R

The current research is to validate the performance and completeness of C2R system. The validation is done in an integrated testing environment and makes use of created test data. The test data is imported in the C2R system and by the speech and action technology transformed to text and put in a report. The output results are verified with the ground truth that stand for the “golden standard”. The meaning of the golden standard is an EPD report created and checked for errors and also the filled in EPD by the family doctor.

The motivation for this research (Personal interest)

In my personal environment multiple family members are working in the medical domain. This differs between a hospital and doctors’ practice. All the members acknowledged the problem with the amount of writing work after examining a patient. They understand the importance of documenting the problems, findings and treatments, but agree that the administrative load is too high. The results of the wiring during the day, is that they cannot do what they love to do: helping people. The amount of people they want to examine, will increase by reducing the writing. In some of the cases it is possible that a medical staff member does not write clear sentences, but only short words. Only to save time on writing, so that there is more time left for the patients. This makes it harder for other medical staff member to interpret the findings in the patients file.

1.3 Research objective

The aim of the C2R program is to develop a hardware and software platform for automated medical reporting, to remove the writing and typing during and after a medical consultation. The objective of this study is to develop an integrated testing environment for the C2R research project. The solution that will be designed and create is a multimodal database for the multimodal test data that is collected during this research. For the testing data a data management plan will be created. Within this study there are two important objectives: a knowledge- and practical aspect.

Scientific objective

The knowledge objective is the objective that aims to extend the current knowledge.

- To assist the future software C2R in understanding the state of the current literature on data management in the medical domain and design data input infrastructure.

Practical objective

The practical aspect is the aspect for the creation of the multimodal test database for the future C2R software in practice.

- To contribute to the creation of the medical multimodal test data and to the storage of all data.
- To contribute to the automatic importing and extraction of the multimodal data in the database, respectively.

1.4 Research scope

The scope for this research project is within the frame of reference of the C2R project. The scope of C2R is focused on creating a fully functional software solution that will support the (medical) care provider by automatically generating reports of medical consultations that can be uploaded to the EMR.

The project is in the early stages and the first of its kind. In this early stage, it is mostly about investigating what is possible to realize. To create a clear focus and goal for the software, it focuses on one domain. The domain that is chosen for this project is the medical domain. This means that medical data has to be created for the test environment.

The scope of this research is in the domain of data management of multimodal data in the medical domain. The multimodal data types in the medical domain that will be managed are audio, video and (numeric) measurements.

Scope for Yoda

The multimodal database that will be created for the Care2Report program will be done in YODA (Your Data). The YODA program stores all kinds of different digital data, think of raw data, secondary data, complete questionnaires, journals, digital sources, observations and multimodal data like photos, and scans. Especially the multimodal data is important for the current research. YODA makes it possible to store, manage, share and attach metadata on research data. YODA is an integrated digital environment that will let you share data with other researcher, store research data for at least 10 years and makes research data easy to find by making use of meta data. The data is encrypted for safety and stored in two geographically separated places.

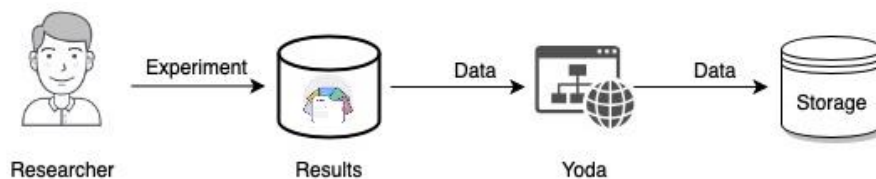


Figure 2 Process storing experiment results on Yoda (Yoda Utrecht University, 2019)

Below a list with the possibilities and limitations that are known in advance within this research scope. To know these in advance it can be taken into account during the research.

Possibilities

- Create a collaboration between a group for the test data in the testing environment
- Safely storing the research data by using encryption.

Limitations

- The database must be made in YODA, because of the used research data.
- The research is done within the university Utrecht, this makes it harder to validate the database solution with for example a test panel or pilot group.
- The gap in practice is that there is not a system that is similar to C2R.

1.5 Research questions

Based on the problem statement and research objectives, the following main research question is defined:

MRQ: “How can research data management for multimodal analysis and testing be organized?”

The Main research question, when it is answered, will tell how it is possible to create a multimodal database for medical research data and testing environment. To answer the main research questions, several research questions need to be answered. The following research questions are defined. Figure 3 shows the complete overview of all the research questions within the research that are defined to answer the MRQ.

RQ1: “What is currently known about research data management in relation to multimodal analysis and testing?”

First a better understanding is required from the definitions as data management, multimodal and corpus and their relations. Also, the current knowledge is studied and interpreted. Literature research gives the opportunity “to build on the shoulders of giants”. The existing research occurs of reinventing the wheel again, also a clear overview is created of the current problems and solution in the investigate domains.

RQ 2: “What are the rules and processes for executing an (simulated/ real) experiment that extracts audio, video and measurements data from medical consults?”

During the experiment there are multiple data types gathered from the participants. The experiment will be executed with fictional patients and doctors and later on also with real patients and doctors. Last mentioned have rules and processes defined in the medical domain for the creation and execution of an experiment and the data.

RQ 3: “How should research data be stored, and organized in a corpus with simulated and real medical consultations?”

Subsequently to the data creation is data storage. For the different data types a multimodal corpus is needed. In the literature study the current possibilities of multimodal data storage are discussed. With this knowledge taken into account, a multimodal corpus is created for the test data to test the C2R software.

RQ4: “What are the rules and processes for testing privacy sensitive data of care provider and patients?”

Where RQ2 is about the rules around data extraction during a medical experiment, RQ4 is about the extraction of data from a multimodal corpus and use it in a testing environment.

The fourth research question will tell how to handle sensitive data form a patient and a care provider. Explain the rules and process that are required for using this kind of data. The data will be used to create a ground truth. The ground truth is the hand-crafted output that will compare to the system output to find the errors in the system (elaborate in chapter 4). The multimodal research data is been used for testing and analyzing the C2R software system. The test data will be passed on by using an API.

RQ 5: “How can real-world experimentation be implemented in the Care2Report program?”

The fifth research question will answer how the simulated experiments are been implement in the Care2Report system. Because the current state of the software is a prototype the recorded medical consultation must be imported as a transcript.

RQ 6: “How valid is the implemented research data management in the Care2Report program?”

The sixth question will explain the validation of the created research data management for the C2R program. The answer to this question will give insight in the manners of validation the Care2report system. The answer explains how valid the research data management is.

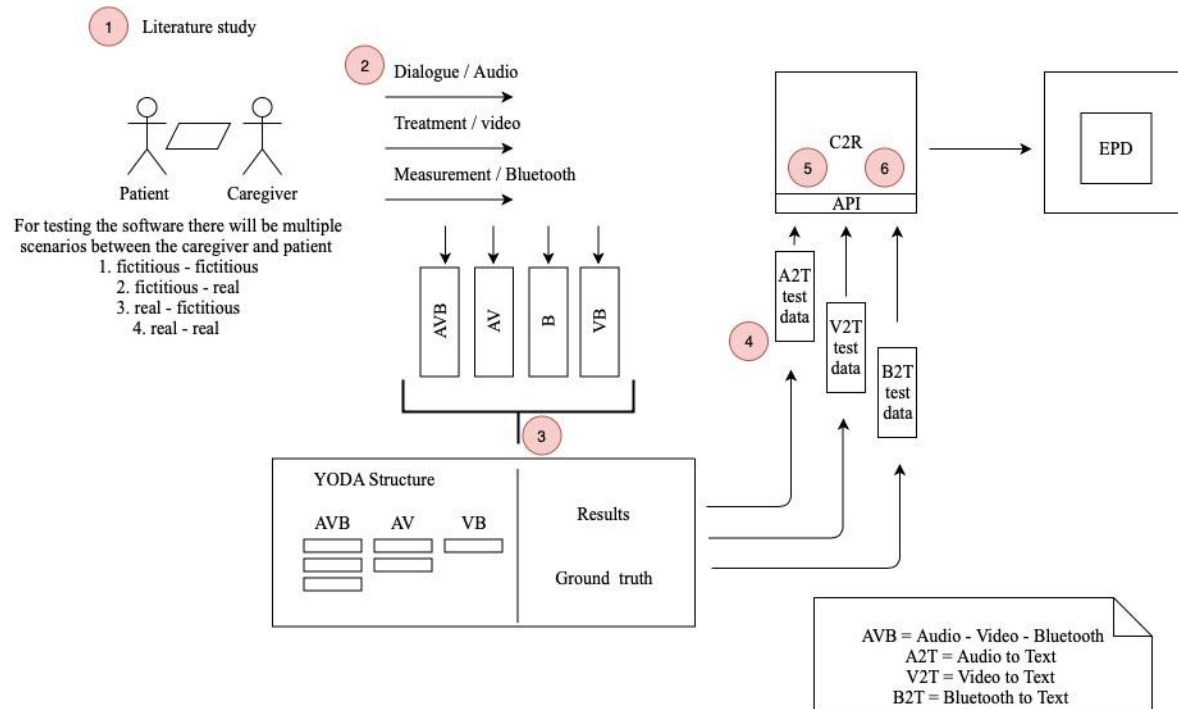


Figure 3 Overview research questions within the research

1.6 Relevance

The relevance of this research is described from both a scientific and the practical perspective.

Scientific contribution

Academic relevance is the value that this thesis will contribute to the specific academic field.

- I. Describe a thorough understanding of the existing knowledge of data management for research data and for large databases. This knowledge will combine different fields that point out the missing knowledge.
- II. Develop and enhance the theoretical base between the fields of data management and the medical domain by designing a research data management for C2R

Practice contribution

The practice relevance is the contribution for organizations and the industry.

- I. Create a guide for the industry to gain and use medical trail multimodal data for future medical software solutions.
- II. Provide the architecture for designing and developing of the database that will be used to store the research data for the testing of the final software product.

1.7 Research outline

The goal of this research is to get an understanding and to get insights in the rules and processes. These are required for processing and storing the multimodal testing data that is used for analysis and testing the created Care2Report system.

The first part is the outline and introduction of this research, this is being discussed in chapter 1 introduction and chapter 2 research method. The second part is the problem investigation in chapter 3 structured literature review. The third part is about the treatment design, this is explained in chapter 4 data and set-up and chapter 5 case studies. The last part is the treatment validation and finalizing this research in chapter 6 testing and analysis, chapter 7 results and discussion and chapter 8 conclusion

Chapter 2 Research method

In this chapter we discuss the research method used in this research. The used research methods are the steps of the Design Cycle by Wieringa (2014). To answer the research questions a combination of desk research and a case study is used. Literature review, interviews and case studies, are also used to perform the research tasks. Tasks such as; problem investigation, designing, developing and validation of the solution.

2.1 Design science – Research framework

The method that is selected and will be used in this master thesis, is Design science by Wieringa (2014). In design science, the design and investigation refer two kinds of problems: the design problems and knowledge questions. *Design problem* makes a change in the real world by creating a solution, this could be a design. The design of an artifact improves the problem within the context. *Knowledge questions* create an answer in the questions about the artifact (design solution) in the context. To find an answer in the knowledge question it is making use of prior knowledge (Wieringa, 2014).

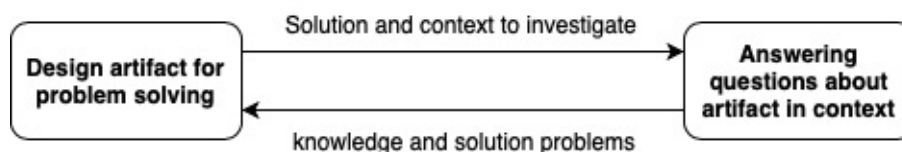


Figure 4 Iteration problem solving in design science research (Wieringa, 2014)

2.2 Design cycle – Research approach

The design cycle consists of an iteration over problem investigation, treatment design, and treatment validation during the complete research (Wieringa, 2014). The three tasks are the design cycle which is a subset of the larger engineering's cycle. Below the five steps of the engineering cycle, the first three are the design cycle. See figure 5 for design cycle in the engineering cycle.

1. **Problem investigation:** This focuses on what must be improved and why it is done to prepare the design of the treatment of the artifact. The current literature is studied to learn more about the problem that would be trailed. The knowledge question is about the artifact in the context.
2. **Treatment design:** Design the artifact that treat the problem. This start with specifying the requirements, identifying the available treatments and designing a new artifact that would treat the problem in context.
3. **Treatment validation:** The treatment design would be validated to see if it treats the problem.
4. **Treatment implementation:** Solve the problem using one of the designed artifacts.
5. **Implementation evaluation:** Evaluation of the treatment's success. This could start of a new iteration of the engineering cycle.

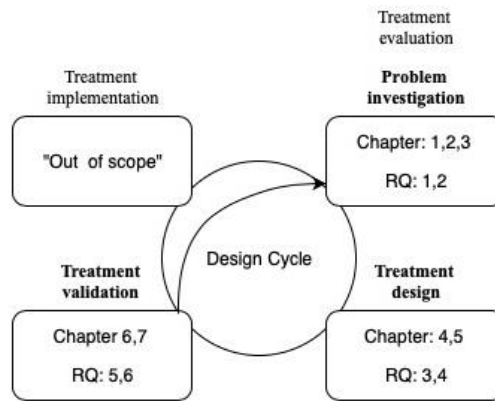


Figure 5 Design cycle (Wieringa, 2014)

The three tasks form the design cycle will now be described in the context of the current research project.

Problem investigation

Problem investigation is getting an understanding of the domain where the research is performed. With the current knowledge of the domain this is a problem that will be solved within this thesis. The problem that must be solved is about how the research data management must be organized for multimodal data.

Treatment design

Treatment design has the aim of designing a database with test data and the corresponding ground truth. The integrated testing environment for this research project must meet the requirements of the created system. Most requirements follow from the rules, processes and laws stated by the literature. In this step the actual database and test data will be developed.

Treatment validation

The testing environment will be evaluated and validated with (bachelor) graduate students. The students are the third-generation team to further develop the system that transforms multimodal input data to medical reports. The database, automatic scripting, multimedia and the ground truth will be validated. The artifacts are also validated by expert reviews. An expert with background knowledge of C2R project will review the solutions.

2.2.1 Literature research protocol

The literature review is used to find answers for research question 1, some insights for research question 2 and research question 4. The literature research protocol has the purpose to give structure to the execution of the literature research. With a proper protocol it gives a more thought out structure to the related work section. The found literature is needed to get insight in the correct knowledge about experiments, storage and testing in the medical domain. It will give some insights in rules, procedures and will gain the answer of the research question 1:

RQ1: "What is currently known about data management in relation to multimodal analysis and testing?"

This research will use literature review of existing relevant academic literature. The scope that is set up is data management and storage of the created test data.

Literature search strategy

The literature search strategy is snowballing literature search. In the literature review parts of literature are combined to form a whole. The parts are found in different papers that address the same topic. Snowballing is a methodology that will review and research the results and current situation from different studies. The reasons for using the literature review strategy are:

- 1) To summarize the current literature both practice and technology;
- 2) To find the gap in the existing literature;
- 3) The opportunities for new research;
- 4) Find the right evidence to support or reject the hypothesis (Budgen & Brereton, 2006 and Kitchenham, 2004).

When individual studies add a contribution to the research it is called primary studies and when a snowballing approach is used it is a secondary study (Kitchenham, 2004). The snowballing process will ensure that the relevant literature is included in the research in a structured way.

Before the literature review can be started there is a review protocol needed. Defining the protocol will result in less bias during the secondary literature study. The review protocol described in the paper (Budgen & Brereton, 2006) for a literature review exists of three steps that are listed below.

- 1) Planning of the literature review;
- 2) Conducting the of literature review;
- 3) Reporting the findings of the literature review.

Planning the review

When starting the search for literature in databases, the first step is to formulate and identify the keywords. With the keywords you search for the papers that are used for the start set in the snowball approach see figure 6. By creating the start set by using the pre-selected keyword, it will prevent that there is a preference for an author. Snowballing search technique tries to create a start set with papers with the high number of citations (Wohlin, 2014).

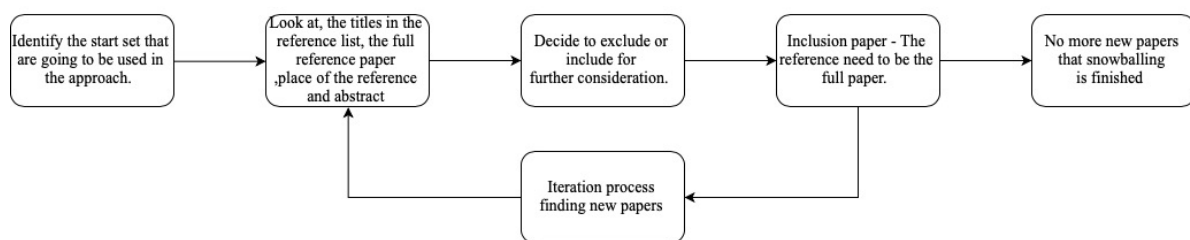


Figure 6 Snowballing approach (Wohlin, 2014)

Below the information about the including and excluding criteria for the start set:

Keywords: Data management, multimodal, multimedia, medical data management, multimodal corpus, research data management, multimodal analysis, testing environment;

Publication years: 2000 to 2011;

Document types: Books, papers, master theses, university sites;

Used search engine: Google Scholar, Google Books, library site Utrecht University. These search engines presenting published papers and books in large numbers and give the most results.

Conducting the review

The literature review is conducted by searching for relevant papers around the different keywords. Within searching there were a couple methods used. The papers that are selected include empirical studies, examples from practice, survey results or definition (Kitchenham et al., 2009). The results of the papers are summarized to create the background in the introduction and the literature review. Webster & Watson (2002) give an understanding of “relevant literature” and how to find it. Within the domain there are leading journals/papers that bring a major contribution. It makes sense to start with the leading journals/papers. Another method used here, is finding the literature backwards. Backward searching is done by using the references where the leading papers is based on, this could be an addition and give in-debt information. Forwards searching is making use of papers that are mentioned in the text, but is not used in the cited papers. (Webster & Watson, 2002).

In this research the most domain leading papers are selected first, by the numbers of citations and ranking. When papers are found that fit the criteria for the selected topic, backwards snowballing is used. Snowballing search uses the sources of the previously selected paper.

Reporting the review

The results of the literature review are important to communicate. The way to do this is to create a short summary of the read paper(s). When using a definition or list from a paper is used, it will be quoted with the proper reference. The reference will always contain; The name of the author(s), title, journal and publication details (Kitchenham, 2004). The papers will be treated, divided and/or discussed per subject or sub chapter.

2.2.2 Semi-structured interview

The semi-structured interview is used in combination with the literature review to find an answer for research question 2. The qualitative interview is used in the qualitative research like the current research. The type of research is about meanings of expert people. Semi-structured interview does not have a complete script. The researcher has some questions prepared, but the conversation is leading (Myers & Newman, 2007).

In addition to the literature review as a method to collect data, Semi-structured interviews are used to gain information from research institutions. Semi-structured interviews have an informal tone and is similar to a conversation. This way we try to find out the experience of the interviewee. It is an informal conversation with a set of questions about what the researcher needs to know. The list of question will cover the main topics that are important (Fylan, 2005).

In this research there will be a semi-structured interview conducted to answer research question 2:

What are the rules and process for executing an (simulated/ real) experiment that extract audio, video and measurements data from medical consults?

The open structure of the interview will help to identify the rules and processes for creating experimental data from patients. It makes it possible for the interviewee to explain their answer on open-end questions and allows a discussion. “They” are institutions that mentioned how they have experienced creating research data out of medical experiments.

2.2.3 Case study

The case study is used to gather data from simulated and real medical consultations to create the test data and the creation of the multimodal corpus. A case study investigates and gathers data from people, groups and institutions. This method is used to answer the third research question. The question about how the research data must be stored and organized in a corpus and how to create it. The purpose of the case study is also to create the data and environment for the complete integrated testing environment.

The case study is creating test data from medical consults between the caregiver and patient in real and simulated situations. Medical consultations follow a general structure: talking about the patient's medical history, physical examination, evaluation and treatment recommendations (Maynard & Heritage, 2005). The test data that will be extracted from the case study are audio, video and measurement data. The case study itself will be held in a complete real or simulated medical setting. The patient is being examined with a medical instrument. This instrument could be a blood pressure sensor, spirometer, temperature sensor, Pulse oximetry or scale sensor. The next step is that the instrument will be applied on the patient and the session will be played based on a script. The session will be recorded, the video, audio and measurement results will be stored. All the data will be used to create an example medical report (ground truth). All the created data from the case study will be used to test the complete Care2Report software product. The case study is also about the storage from the test data and intermediate results within Yoda. The data storage will be described and created for different data sets.

The case study is divided into two separate case studies, the first one is about testing the prototype and storing the test data used within this research and the second case study is about the data plan/ storage for action recognition for a different research.

Overview of the activities within the case study. The first activity is testing of the created prototype with the test data created within this research, the second activity is to create a data management plan for the storage, retrieval and accessibility. Finally, the design and development of the Yoda environment.

2.2.4 Comparison golden standard

A comparison with the golden standard is used to validate the created test data with the system output to answer research question 6: *How can we validate the implemented research data management of the Care2Report program?* The third step in the design cycle is treatment validation. The output of the software will be an automatic generated report of the medical consultations with all the required information. The generated report and intermediate results will be compared to the golden standard. The golden standard is an EMR created by the general practitioner. The difference will point out the flaws in the software that need to be adjusted.

2.3 Validity & reliability

The research must be transparent and reproducible. To make sure this is possible, a research must be valid and reliable. Validity and reliability communicate the trustworthiness of the findings in this research. To make the research trustworthy, it is based on good evidence. This is gained by the way data is collected, such as semi-structured interviews by the right participants.

Validity says something about the closeness what was measured and compared to the intended measurement. For this research it means that it is important to check the data on quality and being logically. Reliability says something about the results that are gained from the interview and case study. The results could not vary in similar circumstances. This means that the gathered results must perform consistently and must be trustworthy on their accuracy. To do that, the interview results could be reviewed by an independent researcher to agree with the findings after the analyses, based on the transcribed interview (Roberts, Priest, & Traynor, 2006).

To make sure this research is valid and reliable, is chosen to discuss the results and measurements in consultation with another researcher within the C2R program. Before conducting gathering the results is within an C2R meeting discussed for the right approach. When the information is gathered, the results are reviewed by PhD students and the final assessed by the Professor.

2.4 Research execution

This research is executed in eight months and is divided in two separate phases with multiple milestones.

The First phase is the research proposal. This proposal creates an understanding of de gap in the literature and practice. The research questions are central in this research and are filling the founded gap. To execute the research, the research method is presented. Also, in the first phase will the current literature and the background around the topic be studied. The findings are summarized in the literature study. Phase one ends after the first colloquium presentation.

The Second phase contains research approaches, semi-structured interviews and a case study. All these research approaches will be used to gather answers for the different research questions and to design the treatments. The designed treatment will be built and finally validated. The validation is done by integrated testing and an expert opinion.

Chapter 3 Literature on research data management

The starting point of the design cycle is reviewing the relevant literature. This chapter presents the results and problems that are given by previous researches. The literature review gives also insight and clarity in the different subjects and terms that are used within this research around data management and storage.

The aim of literature review is to answer the first research questions with existing literature. Research question 1 that is going to be answered is:

RQ1: “What is currently known about research data management in relation to multimodal analysis and testing?”

The literature review will partly answer the research questions 2 and 4 and be supplemented with the to be conducted interview.

RQ2: “What are the rules and processes for executing an (simulated/ real) experiment that extracts audio, video and measurements data from medical consults?”

Research question 4 will give more insight in the processes in the handling of sensitive data by the literature review.

RQ4: “What are the rules and processes for testing privacy sensitive data of care provider and patient?”

The questions enclosed large terms as (research) data management and multimodal analysis and testing. In section 3.1 are the different data types explained that will be used in this research. Second is the term data management with all the different knowledge areas. As third are the multimodal corpus for the storage of the different data types. As fourth are the policies, processes and rules described for the creation of research data. Finally, the literature description how to set-up the ground truth for software testing.

Reviewing relevant literature and researches is the start of this research. The literature shows the current existing knowledge of the different categories.

3.1 Basic terminology

Within the terminology are the definitions presented for the use of this research. The description ensures understanding when the definitions are used.

3.1.1 Data terminology and definitions

In this research multiple types of data will be used. This chapter will give an introduction to the terms that are used in this research.

From the perspective of information science is , data in the scope of research can be described as “*is collected, observed, or created, for purposes of analysis to produce original research results*” (Surkis & Read, 2015).

The more extended definition stated by the Oxford English Dictionary described data as: – “*The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media*” (Dictionary, sd).

This definition from the ISO standard for data; – “*A re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing*” ((ISO/IEC), 1993).

Data is the representation of artifacts such as text, sound, video or numbers. The different artifacts are captured, stored, and expressed as data (Dama International, 2009). Within this research the definition for data is used from the ISO standard. Data can be qualitative or quantitative, factual or non-factual and numerical, textual or audio-visual. Data is much more than just text or numbers in a spreadsheet, page or document. It could be presented in many forms and in different ways. Some examples are images, videos, software programs and many more (University of Edinburgh Information Services, 2014).

Types of data

Data comes in multiple types and media forms. There are multiple types of data used in this research for data management. There exist many variances of data like research data, master data, multimodal data. Not all types are relevant for this research. The data variants that is focused on this research are; Research data, Scientific data, Multimodal data and Clinical data. Explanation from the different data types is given below:

Research data:

Research data is the data that is used within the research its based on. The data is created by scientific instruments and simulations during the research. The data is organized and stored for data analyses (Gray et al., 2006). The data is the extraction of the work of researchers, it is “*everything that would be needed to reproduce a given scientific output*” (Surkis & Read, 2015). The research data is been collected, stored and used within a research (McGowan & Gibbs, 2009). The research data is primary collected for the research and secondary for third parties (Smith II, 2014).

Multimodal data:

There are multiple channels a person can use for communication. Speech, drawing and expressions are human multimodal. The way of creating an interaction is by making use of multiple channels combined (D’Ulizia, Ferri, & Grifoni, 2010). Multimodality is an interaction of visual information as images, text, voice (speech) communication and also gestures. It is possible to use one or multiple interaction modalities together and support each other (Caschera, Ferri, & Grifoni, 2007). Multimodal data comes through the input of multiple channels, images, videos with visual and audio and sensor measurements. The current research is recording in multiple modalities namely audio, video and sensor modalities. Audio, video and sensor modalities that result from medical measurements e.g. temperature or blood pressure.

Clinical data:

This is data that is gathered during a clinical research, from a participant. It is a medical trial so the data that is gained could be; blood pressure, body weight, oxygen level or medical history records. The clinical data will be defined in a source document with all the first recordings. Examples of source documents are, interviews with the participant, medical records and reports by observation. (Krishnankutty, Naveen Kumar, Moodahadu, & Bellary, 2012).

3.1.2 Multimodal VS Multimedia

The terms multimodal and multimedia are used interchangeably or incorrect in the current literature. Because there is a distinction between the terms a short description of the terms multimedia and multimodal is given for a good understanding that is needed for this thesis.

Defining multimodal

Corpus is Latin for body and that is used a metaphor to describe the collection of set of language and communication data. A digital corpus is a computer-based database that stores sets language related data of one or more modality. An example of a multimodal corpora is a digital set of language and communication data exist of text with figures or pictures, diagrams or a digitalized set of films with their transcription of the conversation (Allwood, 2008).

Multimodal is used as an umbrella term, in many different kinds of context. This could be scientific domains and non-scientific. Communication is not only text or an image that is realized through a single medium. Currently modes are digital, and the medium is now become more a site where multiple modes can be composed. The terms “modes” or “modalities” refer to transferring a meaning. Modes are including “*words, sounds, still and moving images, animation and color*” (Lauer, 2009). Bonacchi and Karpiński (2014) have defined multimodal for multiple fields, only the field useful are selected and used within this section. In the field of arts and design is multimodal defined as “*the incorporation of visual, auditory and verbal stimuli in art objects*” (Bonacchi & Karpiński, 2014). In the field of “*Human-Computer-Interaction (HCI) a multimodal interaction is defined as a form of human-machine interaction using multiple modes of input/output*” (Bonacchi & Karpiński, 2014). Multiple modalities are used in communications input and output. The input could be speech or movement communication, the output displays the visuals and audio.

Multimodal is all about the human communication skills. The types of human interaction could be speech, speech and movement, text and movement etc. Turk (2001) and Caschera et al. (2007) state that multimodal have a focus on communication, it's a combination of different modes or multiple interaction modalities. Multimodal combines visual information like images, text, movies etc. in combination with sound and voices. The multiple modalities are used to provide interaction. Within the communication between individuals or groups there making mostly use of one semiotic channel (with text is it alphabetic) to transfer there meaning. When text have to cross different meanings, multiple semiotic channel is used. The multiple modalities words and visual elements in multimodal text are the integrate words, images, video and sound (Takayoshi & Selfe, 2007). The multi-modalities used to transfer the right meaning are moving and still images, sound, music, color, words and animation. The multi-modalities are distributed by media. (Takayoshi & Selfe, 2007)

Anderson et al., (2007) conducted a survey, the goal of the survey was to get a better understanding of the multimodal composition and what they understand by the term multimodal. The participants were familiar with multimodal theory and so 28 participants considered multimodal as a range that consist of multiple different communicative modes such as audio, video, animation, words, images, etc.

Defining multimedia

Multimedia interpret the information from the communication of the human. The information is the used perceptual and cognitive skills. The media that is used to do this are text, audio and video. Media are the resources for the representation of the modes. The modes; words, sound, human motion etc. are captured in the media audio and video. The word medium is the representation in text or diagram to a human, where multimedia is content in the form of media or text. The name of multimedia suggests its multiple forms of media. It is an integration of words, visual, sound and video.

Multimedia is defined as an integration or combination of multiple communications such as text and sound or video to transfer ideas, educate or entertain (Lauer, 2009 and Mohammed, 2014). An example of a multimedia presentation is video, and audio combined. A medium is a way of transfer a multimedia content and meaning. This media type of context could be an image, video or sound (Mittal & Mittal, 2011).

Multimedia data comes from several perspectives of information, textual description, various data types in sometimes high volume. Stored in a multimedia database is multimedia that is combination of the data types; text, audio, images, graphic objects, animation, video and composite media. This are the data types found in a typical multimedia database. composite multimedia data that is compost from two or more types of data such as video and audio (Adjero & Nwosu, 1997, Mohammed, 2014).

Result

Lauer (2009) describes it correctly “*Modes and media are independent of and interdependent with each other*” (Lauer, 2009). The terms multimedia and multimodal are used sometimes for the same meaning, are close together but mode and media are very different. Below in table 1 is a separation created for the terms multimodal and multimedia.

Table 1 Comparison of the terms multimodal vs multimedia

Multimodal	Multimedia
(human) interaction could be a combination of speech, movement and combines visual information like images, text, movies etc.	Present and combine; Text, audio, video with the help of tools to communicate.
Multimodal is transferring a meaning by using media.	Multimedia is a combination of two or more media types e.g. text, sound and video.
Including words, sounds, still and moving images.	The media data that is stored in databases in the data types; Text, audio, images, video and composite media.

The difference between the modes and media. “*Modes include words, sounds, still and moving images, animation and color*” (Lauer, 2009) and media are the tools and needed resources to make and produce text, such as human voice, computer or database (Lauer, 2009).

The inequality between multimodal and multimedia is multimedia is a logical combination of different modes for example (still and moving) images, text, color and sound etc. multimodal can be acknowledge as the way of information representing “*or the semiotic channels we use to compose a text*”(Lauer, 2009). Multimodal represent just like multimedia the different types of data (media) that would be stored in a multimedia database.

The medium is the representation of the meaning e.g. diagram or text. Where representation stores the semantics of the stored medium e.g. English language (Anastopoulou, Baber, Sharples 2001). The term *modality* is then used to refer to written, speech and images (André, 2000). In communication modality refer to e.g. what see visual, you feel tactile also known as sensory experience and perceptual experience. This is related to the person as individual (Anastopoulou, Baber, Sharples 2001).

Definitions for Care2Report

The Care2Report program is focused on speech recognition, action recognition and sensor measurements. For the Care2Report program multimedia data (could) consist of multiple kind of modalities e.g. speech, movement and measurement data. Multimodal data that will be stored in the multimodal database is speech in audio, movement in video and numbers and descriptions from sensor measurements. When recordings are made the multimodal human speech, movement are stored in the multimedia forms audio (WAV) and video. The multimodal speech and movement are used for the transcription within the (test) software.

Multimodal; Multimodal is the communication, in the form of sound and (human) voice or visual like images, text, movies or sensor measurements etc. or in combination.

In the current research is the multimodal used input to create a transcription of the consultation that will be analyzed.

Multimedia; Multimedia is two or more communications media of content in the form of audio, video or sensor data etc.

In the current research is multimedia e.g. text, audio and video tested and stored in the multimedia database that is going to be designed.

3.2 From multimodal speech and video input to text in C2R

The multimodal input

Documentation of medical data must be done the correct way, annually over 100.000 Americans lose their lives because of errors or poor medical documentation. To prevent this from happening the, to creating system must not wrong interpret the conversation. There are already systems that understand and interpret (human) speech to reduce the time that is spend on documentation, speech recognition for medical reporting is used. Researchers are using this technology to seek ways for documenting data in electronic medical records automatically. The system will improve the documentation quality and the physician productivity without negative impact on the users (Ajami, 2016). By using speech recognition technology will increase the demand in electronic health records. Because without the support of technology the physician has to type the daily consultations(Ajami, 2016). The benefit in contrast with dictation is that speech recognition is a relatively cheap, also it the access is fast to the created report and have an accuracy of 98%. The digital dictation is the predecessor, the physician speaks in the dictated, the recording was sent to a secretary who transcribed the complete record. The physician approves the created report. With speech to text the report is created real-time, the words are appearing on the screen while talking (Luchies, Spruit, & Askari, 2018).

Speech and visual (movement) information will exchange information to the computer so it could be distributed everywhere. The characteristics speech and visual will change the focus on human computer interaction. Speech is the most convenient way for people to communicate. Speech recognition has become more interesting, the reason for that is it allows to keep your hands and eyes free. Therefore, it makes it a more efficient way of working, it can increase the workload. Speech recognition can offer the solution (Ajami, 2016). Using speech modality is to save time, and the aim of the program is in the most efficient way to create a report. The use of multiple modalities for input increase the ease of use. Speech and movement (multimodal) create quicker communication between human and computer. When speech is used for the input it will not only transfer linguistic information but also the intonation of the speaker (Furui et al. 2001). Speech recognition transform medical conversations into text, action recognition captures the movements of the doctor that gives a treatment and the sensor data extract the results of medical examinations (Maas, et al., 2019).

The C2R system

The multimodal data will be obtained from audio in the recorded conversation between the physicians and the patient. The modalities such as speech in audio, treatments in video and measurements in sensor media are the input for the system. The platform makes use of multiple devices such as a camera, microphone and measurement sensors combined. Video that is recorded with the movement and the given treatments, and measurements from the equipment and devices that are used. To make the three modalities to meaningful information for a report they must first be transformed into text (Maas, et al., 2019). As described before the multimodal input consists of audio (speech), video (movement) and sensor (measurements). For the current research is it important to translate this multimodal data to text for further processing. For the interpretation of all the information in C2R to generate a medical report, there is a process of three stages;

1. *Recording.* The medical consultation is recorded and processed into written text. For this is the C2R system using speech and action recognition and measurements analyzer derived from medical instruments See figure 7.
2. *Interpretation.* Semantic technology is used to create a representation of the multimodal input See figure 7.
3. *Report generation.* The semantic representation is used to generate a medical report ready to be uploaded un the EMR and checked by the care provider (Maas, et al., 2019) See figure 7.

Within this research video and action recognition is not taken into consideration by test data creation and testing. The reason is the C2R system does not have action recognition yet. For action recognition, test data will be stored within Yoda.

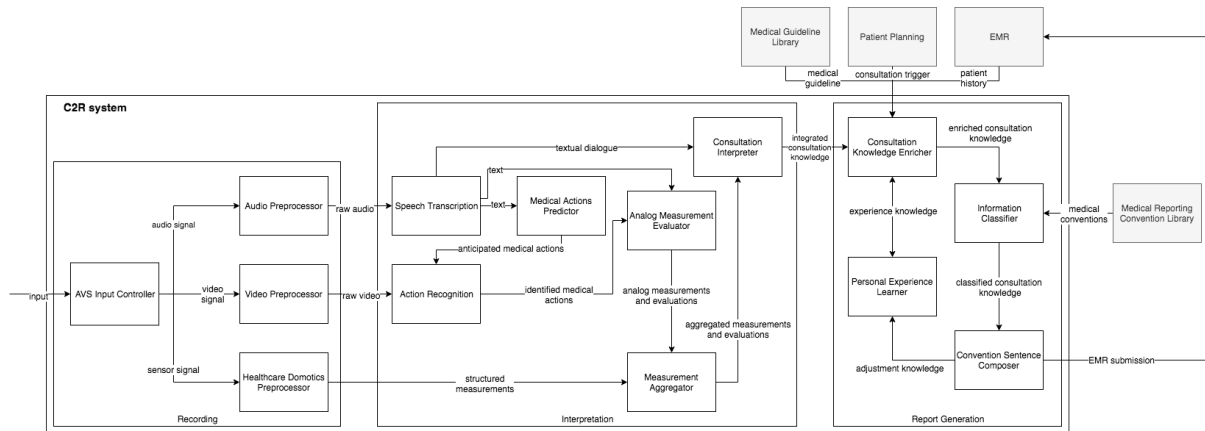


Figure 7 Overview of the Microservice Architecture Care2Report system, based on the functional architecture of (Maas, et al., 2019)

An example of three different multimodal input created to formal representation in figure 8.

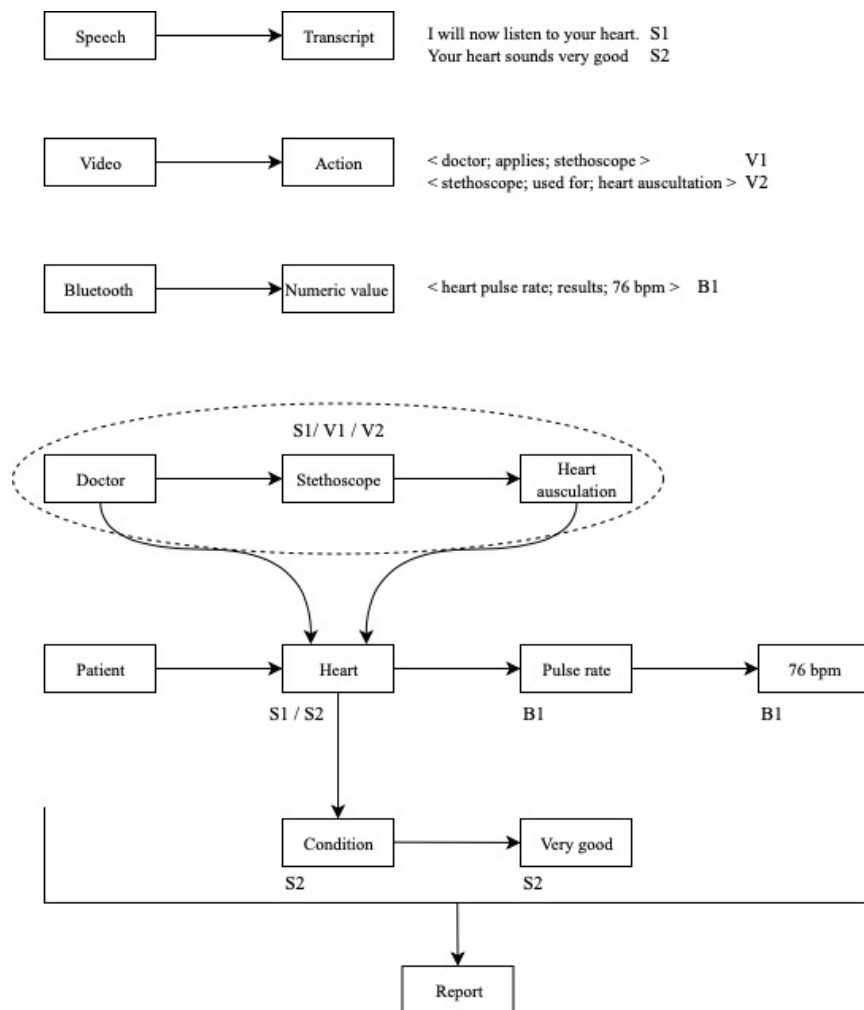


Figure 8 Example of text to formal representation

Below a description of literature about the different multimodalities are transformed to text.

Speech to text

Speech is used between humans as effective communication modality; it is the conversation between the care giver and the patient. When there is a consult or medical treatment is given by a medical professional the conversation will be recorded, and the transcript is created real-time. Speech recognition transform the medical conversations into text. The process from speech to a complete report use recording for the

The next six points are the complete process for automated reporting from speech to formal representation to complete report text. The six points partially reprinted and adjust from (Maas, et al., 2019) part of this research program.

1. *Transcription of speech* (technology Google Cloud Speech-to-Text service are used);
2. *Recognition of concepts and relations*. For the linguistic annotation to extract from the consultation;
3. *Storing and manipulating triples*. Semantic triples ((subject, predicate, object)) are extracted from the text with the knowledge graph;
4. *Building ontologies*. Ontology development is the starting point for the knowledge graph that represents the medical consultation;
5. *Populating ontologies*. A rule-based algorithm to match triples to the ontology;
6. *Generating medical report* in natural language (Maas, et al., 2019).

The six steps of the dialogue summarization pipeline will be clarified by the stages within the Care2Report system.

The recording stage; in the recording stage are multiple modalities: (movement) video, (speech) audio and (measurements) sensor recorded from a medical consultation. For speech a wide available speech recognition is used. The multimodal speech input is turned into a transcript. To give the written text meaning the is happening in the next stage (Maas, et al., 2019).

The Interpretation stage; will transform the transcript into meaningful information. The preprocessing is for the correction of errors in sentences and noise correction. The medical information is enhanced by using information from another modality within the timeline. There is a timeline created with all the modalities for the log of the medical consultation. When the transcript has been removed from possible errors there are semantic triples from extracted. The concept and relations from the text are extracted to create the semantic triples (Maas, et al., 2019). The patient medical graph is a formal representation of a medial consult. This graph presents the human anatomical entities, the found symptoms, medical observation, the diagnoses from the physician and the plan for treatment. To create the ontology, interpretation of the consultation is important. Semantic triples ((subject, predicate, object)) are extracted from the medical conversation and macheted to the ontology with the help of an algorithm (Maas, et al., 2019).

The report generation stage; creates a report within natural language. The fluent sentence is created from a simple fact ((subject, predicate, object)). The sentences are built by making use of several steps. First only the most relevant facts from the ontology are selected called “*content selection*” and the selected facts are placed in order “*Text planning*”. After the facts are transformed in short sentences “*Lexicalisation*”. The short sentences could combine to make longer sentences “*Aggregation*” (Maas, et al., 2019). The sentences are generated into a report, that is imported in the EMR.

Video to formal representation

From video to text start at recording in the Care2Report system. The movements are recorded and put in the preprocessor. Preprocessor is for the correction of errors. The raw video is processed by the action recognition software, this will transcribe the video to text. The action recognition analyzer obtains words from medical descriptions to identify the video stream. This will recognize medical objects from the video. The medical action will be identified and added to the report (Maas, et al., 2019).

Computer recording has advanced in tracing actions of persons and to classify certain actions. It is possible to differentiate objects and identify their attributes. The output of a system that translate actions to text is a semantic representation of the classified actions and (used) objects.

To transform visual content to human language is it important to get a good understanding between the visual content and linguistic information. Rohrbach et al (2013) state that the most approaches that give a summary of video content use rules and templates for the generation of text. The Natural Language Generation (NLG) from video content could be divided in four groups of approaches:

1. Creating a description for a (test) video based on already existing associated text. Use this the text in the most effective way to generate text for the visual content (Rohrbach et al., 2013).
2. Generate text from manually constructed rules or templates. To generate Semantic representation from visual content is by using rules and templates. With the help of this templates text for semantic representation is generated. To describe visual content the object and the corresponding attributes are extracted. For videos the concept hierarchy of actions is defined with different body, hand and head movements. Multiple triples {subject, verb, object, scene} are predicted from a video (Venugopalan et al., 2015). The triples are getting a weighed, the triples with the most weighed are used to generate sentences (Rohrbach et al., 2013).
3. The third approach is to use similar visual content for the extraction and retrieving of text. This is reducing the generation of sentences. The sentences are extracted from a training corpus (Rohrbach et al., 2013)
4. Learning a model to predict the words from semantic representation. Learning a language model to create text from a training corpus. The sentences that are retrieved from the training corpus are based on the object, scene and region recognition (Rohrbach et al., 2013).

Sensor signal to text

The measurement is done with the MySignals – eHealth and Medical IoT kit. The kit contains sensors for measuring for example airflow, blood pressure or temperature by a patient. The sensor signal must be transformed to measurement data for processing in the report. The Analog Measurement Evaluator and the Measurement Aggregator, the interpretation phase is combining multiple modalities. For example, when the doctor measures the patient's temperature and says, "You do not have a fever or increment from your temperature". Without the actual values it does not mean much. The value of 37° is added to the sentence to give it value.

3.3 Multimodal analysis & testing

Analysis is the process of making a complex topic in smaller parts to get a better understanding and detailed examination of the structure. The multimodal analysis in the current research is the process of analyzing the output of the produced results of the Care2Report software system to assess the quality of the created transcript, produced triples and the automatically generated medical reports.

Testing of software is the investigation the quality of the software system. The multimodal testing (data) in the current research is a process of testing the created results from the Care2Report software system.

Multimodal Analysis

Multimodal analysis is about all forms of communication but is mostly concerned on two or more modes of communication (O'Halloran & Smith, 2012). Multimedia analysis is to create valuable insights of the data, situation or activity by using processing. The multimodal data that could be used for the tasks within C2R could be sensory data (audio, video, RFIS) but also non-sensory (internet resources or databases) (Atrey, Hossain, El Saddik, & Kankanhalli, 2010). Multiple modalities are fused together because it increases the accuracy of analysis process.

There are several steps for collecting and analyzing multimodal data. It starts by collecting and logging the gathered data. The face-to-face interaction is recorded in a medical setting. The data is likely captured in a mixture of video and audio recordings, medical measurements and texts used during the interaction. In this research the video and audio recordings are viewed and listened from the medical interaction. The log is a review of what is happened within the recordings (Bezemer and Jewitt, 2010). For this research that will be captured within the metadata, this will describe, tag and label the recordings for easy retrieval. The next step is viewing data is repeatedly done in a multimodal analysis. Multimodal data is played/watched to get insights and understanding of the data. This helps recognizing patterns of routines across different interactions. The third step, a sample is created from all the gathered multimodal data. It is mostly not necessary to analyse all the multimodal data (video, audio). For this reason, a sample is created for the intensive multimodal transcription and analysis. The final step is transcribing and analyzing the speech into writing. Transcribing is the translation of speech to text. Speech is the interpretation of sounds not the accents and voice quality etc. The selected multimodal that is transcribed will be used for further analysis (Bezemer and Jewitt 2010). There are multiple steps for analyze real-time systems as described, there are also multiple techniques for analyzing the behavior that can be classified in three classes.

1. **Simulation:** Execution of the test case, when the system executed properties of the system can be disclosed.
2. **Searching:** By making use of Petri nets to execute attainable analyses.
3. **Theorem proving:** will develop a theory based on logic and the system properties are defined as formulas.(Zhang, Cheung, & Chanson, 1999)

The analysis the analyze different characteristics from multimedia data. This are the overload, lack of quality in textual descriptions and the variety of the data types (Adjeroh & Nwosu, 1997)

Testing multimedia software

Multimedia systems are applications that make use of varied media such as audio, video, images and text mostly combined for presentation. A multimedia system make use of servers and clients that are connected to a network. The general multimedia systems are applications running on several servers connected with a suitable network. The multimedia systems are a collection of different applications running as a system. Because of the different applications and processing the multimedia, are there strict requirements for the systems data delivery, synchronization, streams of data and it must be able to reach

for the stored data and all of this tasks must perform in real-time (Zhang & Cheung, 2002 and Mistic, Chanson & Shing-Chi Cheung, 2002).

Problems with multimedia systems could be data that is not arrived timely due to network congestion, transmission errors or lag in the system. Some of the errors could exist of the CPU limitation by applications that run synchronously (Zhang & Cheung, 2002). Because of this kind of problems is testing a multimedia system different from a tradition software. The most challenging of multimedia testing is finding the stretched situations that will escalate the most quickly. In a multimedia system the testing is focused on testing the network and the quality of multimedia presentations (Mistic, Chanson, & Shing-Chi Cheung, 2002).

For the testing of multimedia software systems are there multiple methods and techniques described. It is likely that a multimedia system disfunction when the system is used. The errors could be occurred due to wrong estimation of treads. Testing the system by stressed situations called stress testing approach, it could significant point out the weaknesses within the software. Other types of testing techniques proposed for modeling and timing tests for multimedia. The three approaches are: labeled transition systems, Petri nets, and constraints (Zhang & Cheung, 2002). This are not relevant for this thesis because the multimedia system is not been modeled or test on timing. To test the processing time of the program is to trace the sequence from the starting stage to the final state. During the trance multiple properties are checked. First the intrinsic consistency from the start to the end. Second the extrinsic consistency for example the multiple audio streams (Zhang & Cheung, 2002).

For testing the errors could be described in the categories; functional -, timing- and synchronization errors. The different categories cover different aspect so they could be conducted simultaneously (Mistic et al., 2002).

Software testing of C2R system

The software testing of the Care2Report have the aim to validate the quality of the produced outcome. The outcome diverse from a transcript, triples, generated report and filled in EMR in the Subjective, Objective, Evaluation and Plan (SOEP) standard. The SOEP standard is the manner the EMR is divided so the Dutch family doctors can define and note the important parts of the medical consultation.

To test the software test cases will be created. The test cases consist of one or multiple modalities that will be run by the system. The software of C2R produces multiple outcomes in streams of data. The outcome is verified and compared to the golden standard. It first starts with the raw data, that are videos of a medical consultation that contain multiple modalities.

The first test output is the Transcript. The generated transcript must contain all the exact spoken words in the medical consultation. The transcript is tested if all the right words are there for further processing.

The second test output are the triples.

The triples are also called the unmatched triples. They consist of subject, predicate and object. They are generated by the software Frog (Dutch) and Fred and Ollie (ENG).

The third test output is the matching ontology

The matched triples are checked if the created triples appear in the medical ontology. This is done by Amazon medical.

The fourth test output is text generation

The matched triples are ready to put in an EMR yet, the matched triples are with Ollie produced to intelligible text.

The final test output is the generated report. The generated output will be compared by the golden standard of a medical report form a professional doctor. To get more insight in how the generated report is created are the logs and intermediate results gathered for analysis.

3.4 Data Management

Data management is not only the managing of the data, it is including the data collection, storage and analysis for reportage (Tompkins,2007). The amount of available data grows fast. Data is designed and organized to serve the needs and functions of different applications. Managing the data is a way of handling a complex problem. Data management ensures that the data collection process is organized, understandable and transparent. The data lifecycle is used to help understand the scope of data management (Surkis & Read, 2015). The topics within data management are the 11 topics discussed in 3.3.1 DAMA Framework.

3.4.1 DAMA Framework

Data Management Association (DAMA)-international is the global community for data management. DAMA published “The DAMA guide to the Data management body of knowledge” better known as DAMA-DMBOK. It is a guidebook for the management of procedures, practices, policies and architecture of the data life cycle. DAMA-DMBOK have a framework with the most important areas within data management. The areas contain multiple steps, activities and measurements for control. The framework is used to create an overview of the most important area that is needed to discuss. The Data management and Curation (DMC) framework in section 3.5.1 is used for the research data management. The DAMA framework is used to identify the most important areas within data management. Only the activities within the different areas that are useful for the Care2Report are presented.

Framework

The main areas in data management are described and specified in the framework for data management DAMA-DMBOK2. DAMA-DMBOK Guide framework is a gathering from the processes of the different knowledge areas within data management. The processes are interacting with the data management knowledge areas. The framework shows the industry standard about the management areas, terms that are used and the best practices. “*Data Management is an overarching term that describes the processes used to plan, specify, enable, create, acquire, maintain, use, archive, retrieve, control, and purge data*” (Cupoli, Earley & Henderson,2014). The framework gives a structure and blueprint for the organizing and the needed content. Data management is divided in the different knowledge areas, the 11 areas are captured in the next list and form the DMBOK2 guide (Cupoli, et all, 2014).

There are a lot of knowledge areas that are fork from or have a relation with data management. They together they are a list that must be executed for to fulfill the core area “data governance” (Wiggins, 2012). The topics within data management that will be investigate/explored that are relevant for this research are:

1. *Data governance*
2. *Data architecture*
3. *Data Modeling & Design*
4. *Data storage & operations*
5. *Data Security/ Data privacy*
6. *Data Integration & Interoperability*
7. *Documents & Content*
8. *Reference & Master Data (management)*
9. *Metadata management*
10. *Data quality*
11. *Data warehousing & Business intelligence*

The topics in the list above are reprinted from the DAMA-DMBOK2 guide knowledge area wheel (Cupoli, et all, 2014).

The last two topics from the list *data quality* and *Data Warehousing & Business Intelligence* will not be comprehensively discussed. The test data that will be created by the experiment does not have to accommodate special standards. The analysis part is out of the scope of this research, the data analysis will be carried out after the software is developed. For the analysis is the test data needed that will be created in this research.

The explanation of the reason for choosing the DAMA guide framework instead of another data management framework or substitute areas. Multi-Dimensional Data Management Framework Version 4.0 (MDDMF) is a framework that must be used in combination with the DAMA DMBOK2 framework. The framework is from William Evans from multidimensionalthinkers.com. MDDMF has no academical background or publications (Wiggins, 2012).

Data governance: Data governance is the core part of the DAMA guide framework, it creates the planning, supervision and oversight for data management and use. The planning, oversight and control is over the management and use of the data. Data governance is for the high-level decisions. Data governance activities are used for the planning, guiding and oversight. The goal is to define the data strategies, policies, standards, architecture and procedures. The different steps that are described above are providing deliverables such as: “*Data Policies, Data Standards, Resolved Issues, Data Management Projects and Services, Quality Data and Information and Recognized Data Value*” (Cupoli, et all, 2014). Data management plan will be further explained in 3.4.1 Data Management and Curation (DMC) Framework. In this research this framework and the related activities will be used.

Data architecture: In DAMA- DMBOK is data Architecture about defining data requirements, making designs for the requirements and match the business strategy with the business architecture. It will provide a structure for data and data-related resources (Cupoli, Earley, S.and Henderson 2014). The goal of data architecture is to create the highest data quality, describe the data specification and make a plan for the (long-term) data. To achieving the goals the input next is needed; *Input* are strategies and architecture from different perspective of the business. To zoom in on one import aspect for this example from input, activities and deliverables are data strategies, data needs and data issues. *Activities* are understanding, developing, analyzing, defining and maintaining. It is important to understand de information need in the business as well defining and maintaining the Data Technology- and meta-data architecture. The *Deliverables* are different architecture, specifications and services. For the example are the deliverables test data, test databases, meta data and data access services. In the data architecture is about the data names, explaining the used data definition, the structure of data uses, the rules and regulation about data and the documentation (Dama International, 2009).

Data Modeling & Design: In DAMA-DMBOK 1st edition called *Data Development*. In comparison to data architecture is data modeling & design also has the activities analysis, design and deployment but for the data solution. Data modeling & design is about analyzing and defining data requirements, designing the data structures and solutions and about security and usability of data. The data solution value is the data that is defining the data requirements, data design and solution components (Dama International, 2009).

Data Storage & Operations: In DAMA-DMBOK 1st edition called *Data Operations Management*. Data storage & operations do have two main focus area, the first is database support and the second data technology management. The goals are protecting the data, managing the data lifecycle and the data bases performance. Important parts of data technology are a database service level agreement and make sure there is a back of the data and data recovery. For data storage is to make sure data is stored in a centralized location, and make sure it is available for the software (Dama International, 2009).

Data Security/ Data privacy: Data security is about the policies that are in place for confidentiality, privacy and using data and giving the right people access. The policies for the data security must be must make it possible to cooperate on a daily basis. Data must be accessible, and data be change but it is important to prevent inappropriate access to the data. The best way is to minimal meet the privacy and confidentiality legal regulations. Data need to be clarified in a class of confidentially from General Audiences to Registered Confidential (Dama International, 2009).

Data Integration & Interoperability: new in DMBOK2. Not available in the current DMBOK. The DMBOK2 is not public accessible for free yet.

Documents and content: are focused on the integrity, protecting and access of unstructured data stored in files outside of the relational databases. This is important because around 80% of all the data that is unstructured data and stored outside of the relational databases. The data is found within physical files and electronic files such as text documents, audio files and video recordings. The goal is to store the available data in unstructured way for efficient and effective retrieval. For the retrieval is meta-data used, this comes in different forms. There is making a standard format, existing forms are adopted, or meta-data is self-documenting.

Reference & Master Data management: It is about managing shared data, reduce useless data to improve the data quality. Reference and master data are occurring in data warehousing/ storage. Reference data is a classifier for data to place them in categories. Some types are; Party- Financial- Product- and Location master data. Reference data could be value keys or ID in databases to link categories. Master data gives context about the references. The process is about how it will be created, integrated and used (Dama International, 2009).

Data Warehousing and Business Intelligence Management: is the collection and storage of data for the presentation of data for analysis and making decision based on the data.

Meta-data Management: Meta data is data that will say something about the data itself and such as the object or relationships. The data, objects and their relationships can be modeled in a context diagram. Meta-data is given the data descriptive contact or identification for making managing the data easier. The three methods for segregate meta-data; descriptive meta- data (keywords, terms), structural meta-data (Format (Audio, video etc.), labels), or administrative meta-data (Access rights, relationships). The activities to create sufficient meta-data are first to get a good understanding of the meta-data requirements, then to define the meta-data in a context diagram before developing and implementing the meta data in the environment. This will create quality meta-data and meta-data models (Dama International, 2009).

Data Quality Management: is about improve the data quality and performance. The activities defining requirements, implementation and control are activities for to improve the data quality (Dama International, 2009).

3.5 Research Data Management

Data management is mainly focused on data in large organizations and business, where research data management in large research institutions and universities on the sensitive research data. The main differences will be discussed in this chapter.

Research data management has become more a strategic priority of universities. Research use more computer technology that means they produce more and larger data (sets). All this data must be stored in a way that it is easy to process, access and analyse. The data may or could contain personal information for example of participants, so it must be managed and secured (Cox & Pinfield, 2014). Research data management (RDM) provide support and infrastructure for research data. RDM is needed in research if there is data that is gathered such as images in a folder, treatment timing and given dosage, process data stored in a spreadsheet and analyzed data (Surkis & Read, 2015).

Data management plan (DMP) and research data management (RDM) is important for every research. Where RDM makes it possible to share and re-use data, gives DMP guidance to the researchers for collection, using and storing of raw or processed data. The data that is being shared and reused is diverse sources such as experimental and statistical results, recordings and transcripts (Borgman, 2012). Research data is the data e.g. numeric, text, videos, audio etc. that is used within research. The data is created by research approaches, instruments and studies during the research. The results is data that is captured and stored (Gray et al., 2006).

Research data management is a process of labeling, storing and to make access possible for data in a research project. This is necessary for when a researcher collects image or video data from participants, this could be processed and analyzed. An analysis of the output combines the found information. The combination of multiple different data types combined e.g. measurements with equipment, video and audio makes it more complex to analyze. Data management will organize the different data types for a subject (Surkis & Read, 2015).

The data is stored in files on the computer of the researcher or on (shared) servers within the university. Research data management is data organization through the entire research life cycle. It consists of different activities and processes including the “*creation, storing, security, preservation, retrieval, sharing and reusing*” of the data (Cox & Pinfield, 2014). The data lifecycle is an overview of the multiple stages that are involved in successful research data management. The data life cycle is discussed in chapter 3.4.1 and presented in figure 9.

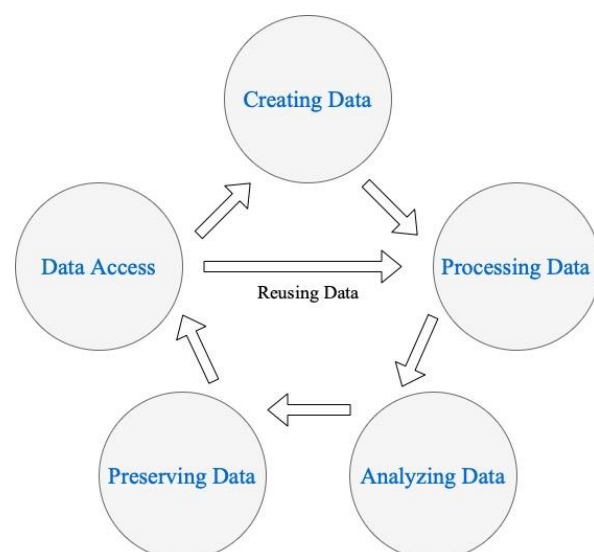


Figure 9 Life cycle for research data is partially reprinted and adjust from (Surkis & Read, 2015).

Research data management is generally within the first three steps (Surkis & Read, 2015). Example with the C2R program:

1. Creating data in the current research are video, audio and measurements that will be collected.
2. Processing data, the data will be extracted from the measurements or the recorded video.
3. The output will be analyzed so the found results can be distributed as academic result.

The three stages require data management to make sure University's and researchers document how they collected and analyze the research data. When the data is described in an understandable way, it could be used by (other) researchers for testing the validity of the original data (Surkis & Read, 2015). When research data is effectively managed, the data lifecycle even begins for data creating. It begins with Data management plan. The Data management plan will describe the data specification, storage and controls. The data management plan is explained in chapter 3.4.1.

3.5.1 DMC Framework

The data management and curation lifecycle

The data management and curation (DMC) wants to improve the current and future use of research data. The research data is data that is collected, used or hold during a research. The meaning of curation is the process or actions for selecting, organizing and preservation of research data (set). The data management framework DMC helps research and give them multiple perspective on the management of store, manage and use research data. For all researchers and scientists, it is hard to manage effectively their research data (Smith,2014). The aim of DMC is to improve data management of future and current research data. Research data is quantitative or qualitative e.g. results, databases, documents, images or audio files.

The four key concepts of data management and curation (DMC) are: Data Management Planning, Data Curation, Digital Curation, Digital Preservation. The DMC framework with the four key concepts are shown in figure 10 (Smith,2014).

Data management plan (DMP): The Entire data lifecycle is the data lifecycle management process for research data. It will create a planning for managing; data types, formats, meta data, standards, privacy etc. (Smith,2014). The data management plan was required in 2011 for research that want to get funding to create a plan for their data sharing and management of their results. The DMP explain the context of the research study and the planning includes:

What is produced during the research such as types of data, the standards and formats are fixed for the data and meta-data, the access and sharing policies for protection and security, policies for distribution and re-use and the planning for storing data and access to other research projects (Smith,2014). The DMP includes the parts that are described below.

Data Curation (level 1): is a data lifecycle process that will provide descriptive- and representative information about the research data through applying metadata (Smith,2014). Data curation ease of the adoption of existing meta data by standard formats for “*description, representation, organization, aggregation, access, discovery, and storage of data*” (Smith,2014).

Digital curation (level 2): a data lifecycle process for the managing and storage of research data. Digital curation facilitates the managing and the storage of existing research data that will be stored within a

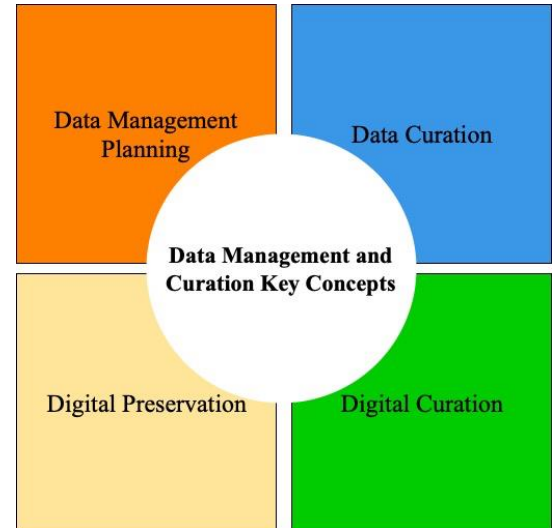


Figure 10 Data Management and Curation (DMC) (Smith,2014).

data repository. It is only effective if the data is stored using the right meta data, guidelines and metrics (Smith,2014).

Digital preservation (level 3): is a data lifecycle process of research data for long-term preservation of the created research data. It focuses is on “*maintaining the authenticity, integrity, and security of curated research data within a standards-based repository*” (Smith,2014).

The four concepts are the stages and processes for data storage, management and preservation within the lifecycle. DMC includes four data lifecycle management processes:

- 1 The organization fulfill the requirements of the policies and data management.
- 2 Data creation (primary and secondary data) and provide minimal description.
- 3 Add metadata value for management and storage of the data lifecycle, to help archive data.
- 4 Organize the data authenticity and executing the technical and strategic actions (Smith,2014).

Data management plan for research data

A Data Management Plan (DMP) provides a road map how to handle data. It is a document that describes how the created data during the research treated during and after the research project. The DMP is a data lifecycle management process of research data what covers all the parts of the data life cycle “*creation, storing, security, preservation, retrieval, sharing and reusing of data*”(Cox & Pinfield, 2014).

Data management planning definition is: “*the planning of policies for the management of data types, formats, metadata, standards, integrity, privacy, protection, confidentiality, security, intellectual property rights, dissemination, reuse/re- distribution, derivatives, archives, preservation, and access*” (Smith,2014).

Data management plans included in the National Science Foundation (NSF). The NSF is a government agency in the United States that supports research and education. In the document of NSF Proposal & Award Policies & Procedures Guide (Chapter II.C.2.j) (NSF, 2011) is the Data Management Plan the objects presented. In the Netherlands is a similar foundation it's Netherlands Organization for Scientific Research named NOW. Both organizations make use of the web-based tool to create a DMP online. The tool is created by the Digital Curation Centre (DCC) (<https://dmponline.dcc.ac.uk/>). The tool will help to create, review and share the data management plan. The DMP is a useful tool for the strategy and plan how the research output is being shared. DMP templates, and best practices for the long-term stewardship and access of data, all which strive to help researchers to manage their data responsibly.

The paper Michener (2015) described 10 rules for the creation of a good data management plan.

Rule 1. Determine the requirements. The requirements for the data management plan from the institution, sponsor or for a proposal the funding program. The research institution mostly develops their own methods for margining the research data. The guidelines from the Health institution are based on the data sharing policy document. For the creation of a DMP is the DMP online recommended. It offers resources for updating the DMP during the research because it's a “living document” that develops during the research (Michener, 2015).

Rule 2. Identify the data that is be collected. The components of the DMP are based on the data types and amount that is gathered during the research. To create a good DMP the collected data must be understanding the will be collected. Therefore, are the types, sources, volume and data file format important to understand the collected data (Michener, 2015).

Rule 3. Define how the data will be organized. There is understanding of the data volume and types that will be collected, the following step is defining how the data is being organized and managed. For managing the data are data tables generated. For small tables can Excel be used, for larger is relational

database management systems (RDBMS) required. For the DMP it is useful to identify on the data types and names (Michener, 2015).

Rule 4: Explain How the Data Will Be Documented. The metadata documentation will define information details about the data, “*details about what, where, when, why, and how the data were collected, processed, and interpreted*” (Michener, 2015). The information must make it possible to find and use the stored data.

Rule 5: Describe How Data Quality Will Be Assured. Data quality refer the process of assess or improve the quality of data. For monitoring of data quality is it possible to use verifications test and double-blind data entry. For error detection is it also possible to use statistical and visuals (Michener, 2015).

Rule 6: Present a Sound Data Storage and Preservation Strategy. It is important to duplicate the data set regularly, it is plausible that papers are getting lost, hard drives crashes and URL are break down. The result is not useful for the research anymore. Data storage and prevention are crucial in a good data management plan. Important is to think how long the data must be stored, protected and the availability after the research. A data store approach is to store it in two geographically different locations example is the desktop and university repository (Michener, 2015).

Rule 7: Define the Project’s Data Policies. Research and institution have a responsibility to the gathered data. The DMP include the policy statements about managing and sharing the data. The policies include data sharing arrangements, the plans for retraining, sharing and the legal and ethical access about the sensitive data about human subjects (Michener, 2015).

Rule 8: Describe How the Data Will Be Disseminated. The research will define the disseminated to others especially if it is outside of the research group. Disseminate is possible in two ways passive and active. The more passive way is mailing the data or post it on a personal website. The active way is publishing the data set in an open repository or submitting the data as an attachment by an article.

Rule 9: Assign Roles and Responsibilities. Within the DMP are the roles and responsibilities defined from the group members. Roles can be data creation, data entry, metadata creation, backup and archive creation and administration (Michener, 2015).

Rule 10: Prepare a Realistic Budget. Create a realistic plan/ budget for data management. Take time and cost of the software, hardware and personnel into account (Michener, 2015).

3.5.2 Research data management in a larger institution

Universities (Higher institution) are trying to work out how to manage and support their research data. The support for infrastructure, storage, sharing and curation. Universities in different countries are developing infrastructure for efficient managing and storing of data in repositories by researchers. The RDM in larger institutions have challenges with the complexity of managing research data. This is because institution, faculties and even departments differ from the way data is created, used and stored (Cox & Pinfield, 2014). When institutions organize their resources in generic form, this could be extended to data management. For the organization of the data is metadata important, to retrieval of curation management.

Universities and institutions store not only there data locally, but they want to store some data at shared sites or other Universities or locations. The aim of the shared location is to use a site to store the data is because of the computing power and could accessing all the needed date to analysis the experiment data. More people have access and it is improving the collaboration because of the possibility to share with the world (Hoschek, Jaen-Martinez, Samar, Stockinger, & Stockinger, 2000).

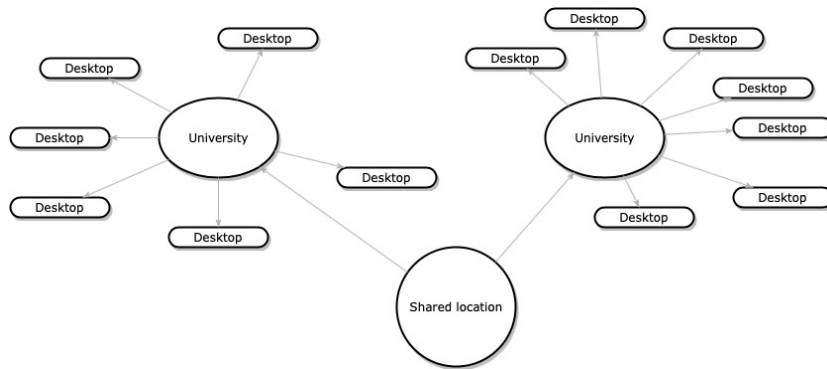


Figure 11 University storage network (Hoschek et al., 2000)

Research data is moving around, in and out databases, it's stored or delivered at a software product. Researchers struggle with unscussesfully store and manage the huge volume of documents and data sets that are result from there research work. There are universities that offer services for better research data magement, but many research have a disorganized manner and lose valuble data for furture research. Data is approximately 80% of the time stored in unstructured formats. Structured formats are databases, files with metadata (Dama International, 2009).

3.6 Storage in a multimodal corpus

More and more multimedia data is stored in types as images, videos and audio. Because of this phenomenon the research interest in efficient and effective storage of multimedia data. When the amount of information that was stored on computers was growing. A process for storage and retrieval is made simpler Database Management System (DBMS) was created. For the storage of multimedia data is the current DBMSs are not sufficient enough (Lu,1999). In this system it is possible to search for documents on letter or number characteristics. For the storage and access of multimodal data the system multimedia database management system (MMDBMS) is used (Lu,1999). A database is a collection of data, where a database management system is a program that connects and related the data within the database.

3.6.1 Traditional vs multimodal database

Because of the broadness of the topic ‘Multimedia database management’ is this research only focusing on the storage, access of multimodal data and architecture for a multimedia database.

Multimedia storage servers, store and provide access to multimedia data e.g. Audio, video, text and images. The multimodal data exist out of multiple media components that has to work together and must be coordinated. The database servers for multimedia data differs from traditional databases servers in the storage capacity, data transfer and real-time obtaining from the multimodal data (Gemmell, Vin, Kandlur, Rangan, & Rowe, 1995).

The multimodal storage servers provide online access to sources as images, videos and scientific research. The connection between the sources and users is a high-speed network. This makes it possible to access the multimodal data and to adjust the data real-time. Multimodal databases serves are significant different from the conventional databases. The difference is because of the two characteristics audio and video (Gemmell et al., 1995).

Multimodal database has other specifications than conventional databases. Not only the storage also retrieval of the multimodal as “stop, pause, record and edit” the data are differing from traditional databases. Multimodal database serves must provide high speed storage and retrieval for large data collections (Gemmell et al., 1995).

The critical components for a multimodal database server are databases that support continuous retrieval and networks that make sure data is delivered on time and synchronous. In paper Subramaya (1999) are important features of traditional and multimedia databases defined. There are quite some differences between the types of database. For the storage of multimodal data is it important to know what functions there are not available in a traditional database. In table 2 the functions for storage of traditional data and multimodal data compared.

Table 2 Salient features of traditional – and multimodal databases (partially reprinted and adjust from “Multimedia databases”, By Subramanya, S.R., 1999, IEEE, Volume: 18, Issue: 5, p.17

Function	Traditional database	Multimodal database
Data acquisition	The data input is data from entry terminals that comes as context out of documents.	Audio, video data from various kind of devices e.g. video cameras and microphones can be input for the database.
Data formats	The data is stored as binary files.	Files are stored in different formats. E.g. for audio: au, wav, midi and for video: MPEG.
Data storage	The data is stored in “uncompressed” form. The data could be stored on multiple disk/locations	Data from video/audio is large in size so they are mostly stored in some compressed form. Data is effectively stored by using storing schemes. (single storage)
Index organization	The index is organized in a suitable data structure.	The index requires multi-dimensional structures e.g. grid files
Query	For data extraction are SQL queries used based on relational calculus.	Queries that are keyword-based. Here for used “query by example” and “query by content”.
Search and retrieval	Searching is based on the key given from the query. The search could be specific or wide.	The query gives a ranked list similar to the query as result rather than a specific result. The user gives feedback to the search engine based on the results.
Transmission	The network transfers the data, this should be fast enough for response on a query. (no real-time requirements).	Real-time requirements, for the quality of services about the performance and synchronization. This for the retrieval of video and audio.
Presentation	The data is primarily text with occasionally some graphs or charts.	The presentation should handle ranked results and different media. This could be audio and video.

The table shows the important features between the different types of databases. The challenges for designing a multimodal database are discussed in the section below.

The function of a multimodal database is similar to a traditional database, but where a traditional database makes use of single tables or record well multimedia database for the access of data single object or composite object (Adjeroh & Nwosu, 1997).

Different ways of storing multimodal data

The databases provide consistency, security and availability of the multimodal data. The database must provide easy adjustments, obtaining and storage of large amounts of data. MMDB gives a “*unified frameworks for storing, processing retrieving, transmitting, and presenting a variety of media types in a wide variety of formats*”(Subramanya, 1999).

There are multiple manners of storing large amounts of multimodal data. Subramanya (1999) describe a hybrid solution where Direct Access Storage Device (DASD) this is used for storage of data that is use regularly and the Optical Jukebox is used as secondary storage. The reasons to make it hybrid is because of DASD has a higher throughput but have higher cost and lower capacity and for the Optical Jukebox is it the other way around.

Different manner of storing means different performance and cost. The pyramid in figure 12 shows in the top main memory with the highest level of performance, highest cost but smallest storage capacity and sustainability (Adjeroh & Nwosu, 1997). The lower you go more performance in access time is decreasing.

The highest level of storage is random access memory, second is a magnetic disk drives that could be connect with online services, third is the optical storage (jukeboxes) online in some situation and finally the lowest level is offline devices or optical disk (Adjeroh & Nwosu, 1997).

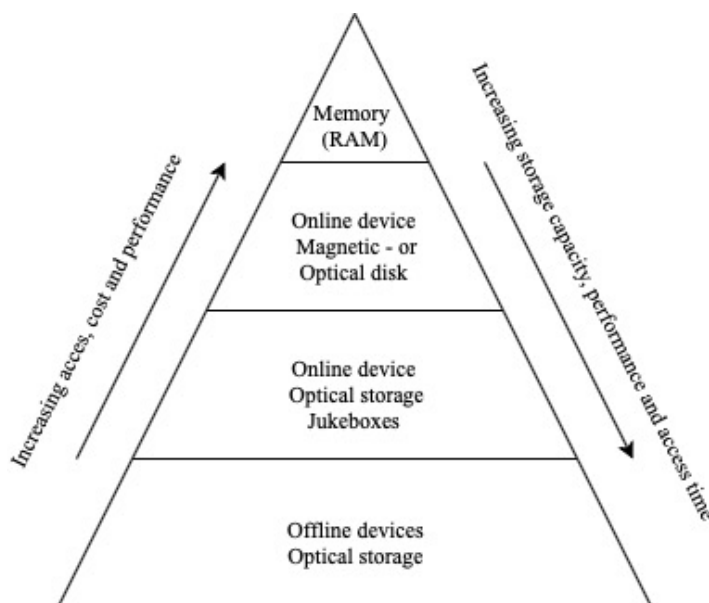


Figure 12 Hierarchically organized storage for multimedia databases (Adjeroh & Nwosu, 1997) and (Mohammed,2014)

3.6.2 Metadata for multimodal storage

Multimodal database has to store more than multimodal data e.g. video, audio. Also, the assigned meta data to describe the multimodal data. The metadata for multimodal data is format data, keyword data and feature data. Below a short description of the three meta data types:

- *Format data* gives the information about what format the data is.
- *Keyword data* is describing the content of the data. The description is generated form the context. From a video it could tell the time, place and persons who participated.
- *Feature data* is content dependent data. The features describe the content. Example for an image this describe the colors, shapes and sizes (Subramanya, 1999).

The metadata helps the queries that select based on the description of the data. A query is mostly keyword-based (Adjeroh & Nwosu, 1997). Metadata is convenient in the steps querying and retrieval. The steps use the terms in for the indexing of the data (Boll, Klas, & Sheth, 1998). Metadata gives a description about the content, condition, quality and tells the location of the data. A video that is stored tells noting more than it is a video data type. Without the metadata it is not possible to tell the underling properties. The metadata could tell the creator of the video, the participants, the quality etc. Metadata helps determine the usefulness of the video for the application. Metadata play an important role in managing of multimedia data (Mohammed, 2014).

Metadata is the most important in the management of multimodal. The reasons why metadata plays a more significant role is because of the match of the query is no longer sufficient for the retrieval of different types of multimedia data. Content-bases processing is a processing technique that is extremely hard and not sufficient in larger datasets. Even when it is possible, it can't be used very often, so it is not efficient. Derived data are the facts that are been gathered by answering questions, this is part of the metadata. The context and semantics are value while working with audio and visual multimedia data (Mohammed, 2014).

Classification in metadata

Metadata is suitable of giving information about the data. Metadata can be classified in content dependent, content independent and hierarchical based.

1a. Content-independent metadata: this is metadata that does not tell something about the content of the datatype. It tells for example the latest modification-date, location or the type of the sensor/medical equipment that is used for the recording. It tells noting about the content, but the metadata is still useful for data retrieval (Burad,2006), (Mohammed, 2014) and (Boll et al., 1998). "*This type of metadata describes the content of document without direct utilization of those contents*" (Burad,2006) and (Boll et al., 1998)

1b. Content-dependent metadata already suggests it, content-dependent metadata is focused on the content of the datatype. This metadata tells about e.g. the size of the data document, and the content within the document. "*This type metadata is based directly on the contents of a document*" (Burad,2006), (Boll et al., 1998).

2. Hierarchical classification gives four types of metadata; descriptive metadata is used to describe the data types, technical metadata that related how the metadata behaves in system functions, administrative is metadata that is used in managing and administration. Metadata for preserving that is related to the maintenance (Mohammed, 2014 and Burad,2006).

3.6.3 Designing issues multimodal database

Design

The first section of this chapter shows that their significance difference between traditional and multimodal databases. To create a multimodal database there are essential aspects that have influence on the design. Subramanya (1999) created an overview of the aspects and their impact. So has the (huge) size of the data, the context, complexity and quality an impact on the database. The impact is made on, things like the storage, retrieval, presentation and queries (Subramanya, 1999). The multimodal database is the core of a multimodal system. The database is a controlled collection of a set of multimodal data in the form of; text, images, objects, video and audio (Adjero & Nwosu, 1997). The designing of a multimodal database contains some challenges and requirements. The requirements that are import for the design of a multimodal database are listed below (Subramanya, 1999).

Design requirements database

Subramanya (1999) created a paper with important requirements for designing a multimodal database. below the main topics that are mention within the paper and are useful for C2R.

- The database must handle different types of storage, in-and output such as video, audio and text;
- The different data will be stored in a various of formats (.text, .mp3, .WAV) the database should handle all those formats;
- The database should be accessible regardless of the operating system is used;
- The media must be findable with user-friendly interface or queries.
- The different media within the database should synchronized different media type for presenting it to the user.

Also, Adjero and Nwosu (1997) described a set of requirements for the designing of a multimodal database. They state that it must have at least the same capabilities as a traditional database. But also, additional requirements like, large storage capacity, query and performance support and data integration, composition, and presentation (Adjero & Nwosu, 1997).

Design issues

The design of a database is dependent on the nature and characteristics of the data. The characteristics of multimodal data that have impact on the database design are: the large sizes of data that is stored and retrieval is hard because data does not own accurate and subjective qualities (Subramanya, 1999). The overload of multimedia and limited descriptions of the data provide retrieval issues. The huge amount of data makes it for users harder to make a good request for the retrieval of the data. Because of the limited text description of the data makes it more convenient for content-based access (Adjero & Nwosu, 1997).

Mohammed (2014) states that, theoretically, it must be able to handle multimodal data in the same manner as data based on numbers or characteristics, but unfortunately, there are some challenges. The problem arises when different data types are extracted from different sources and must be present at the same time. When the problem occurs, users have problem to retrieve their request (Adjero & Nwosu, 1997). The multimedia is captured with different approaches, the processing and storage must be able to handle the different from of capture. Because of the storage of different types could a request not been answered with text. A structure has to be used for the indexing of this type of data.

3.7 Testing environment for software testing

A testing environment is setup to test the software and hardware so test cases can be executed. Testing of the software will help to run careful tests to find bugs within the software before the software is put into use by the user. Unfortunately, not all bugs and errors are found before the software is used, some clients report failure or error. The tester could not eliminate all the bugs on forehand because of multiple reasons. A reason could be the code, or the user operating environment is released but there was not time to test it. To execute the test, must have access to the same hardware, operation system and applications as the user in the testing lap. Also, a reason is the user uses another order, combination or value in their execution than the tester (Whittaker, 2000).

To test software the tester must consider different aspects like the function of the software, the input and the environment in which the software will operate. For testing the effectiveness, the tester must use of fundamental testing techniques, this will test the process rather than a single activity. It gives an understanding how the product is going to be used in its environment.

The tester will simulate here the interaction and input that could be enter the software within the operating environment.

Essential is the simulation of the using behavior and input so as an actual user. This is exactly what makes it difficult since there are different file formats, protocols and a combination of third-party applications (Whittaker, 2000). The process of software testing is running/executing the software to determine if and when it matches the stated requirements for its operating environment. The process of software testing can be subdivided into four phases:

- Phase 1: Modeling the software's environment
- Phase 2: Selecting test scenarios
- Phase 3: Running and evaluating test scenarios
- Phase 4 Measuring of the testing progress (Whittaker, 2000).

Phase 1: Modeling the software's environment

The testers created a simulation of the software and his interface. There are different interfaces that could be used for testing. Below in table 3 the interfaces:

Table 3 The interfaces used by testers for testing software (Whittaker, 2000).

Interfaces	Description
<i>Human interfaces</i>	it is a menu driven interface. For the input is used mouse clicks, keyboard, and input from other devices.
<i>Software interfaces (APIs)</i>	are used by software to access the database or libraries.
<i>File system interfaces</i>	is used by software when external files are read or adjusted.
<i>Communication interfaces</i>	is about direct access to physical devices

When testing the tester must understand the interaction from the user's perspective and also find deviations from usual interaction. For testing the interface there must values be chosen for the input and the sequence of the input. To decide the sequence of the input there a model is generated for formal language. Every input is coded with a symbol from the alphabet to create a model. The model produces a bird's eye view of all the test and the overall picture. The model that is most used is a state diagram that will describe the input for testing the software (Whittaker, 2000).

Phase 2: Selecting test scenarios

There is an endless number of test scenarios and everyone costs time and money. From the many test case scenarios could only a subset been selected. For the determination of what is a complete test set covering code statements and input. This means that at least one time all line of code is executed and applying each external event. These are the criteria testers use to determine the completeness of the

created work. It is not perfectly written code that makes that software works sufficiently. What is important are the execution paths from the code. There is a set created of the input, this represent different scenarios that make the software respond.

There are two possible types of testing software. There is path execution testing and input domain testing. The aim of path execution testing is to select the scenarios that cover at least the source code one's, and that all the statements (such as if,case,while) within the code are executed. This could also be done with the data flow. Test scenarios are covering the data structure. Fault seeding testing is creating an error in the system, write a scenario that would find this error. This scenario will help finding the real errors.

Input domain testing covers the interface and the statistical measurements. For the test a set is selected that test all the input possibilities. This mean all the interface options like menu, button or field are simulated. A path must be created that the typical user would use when using the software (Whittaker, 2000).

Phase 3: Running and evaluating test scenarios

When the tests are selected for the test scenarios are, they convert in testing code. This is to occur errors from manually apply test situation. It is possible to automate the code generation for testing to simulate users. To test is complete automated every input source and output destination of the environment of execution. Within the code are certain data-gathering hook included to gather information about internal variables and object properties etc. The tests will provide information about errors and failures within the system.

There is a comparison of the output after a test has been executed and the expected output that is specified and documented. If the system supposed to present a transcript of an audio file, it is hard to know if the system detected and collected all the right words? That is why the "actual- vs-expected output comparison" still is done by a person (Whittaker, 2000). The tester executes is analyzing the output data compared with the expected output.

For the comparison needs to be a formal specification be described. The specification defines the correct behavior of the system so incorrect (failure) can be found. When a failure has been found and fixed the software has to be re-tested. The reason is that the fix only fixes the reported problem, fail to fix the problem, actually fix the problem and broke something else and did not fix the problem and still broke something else. To Reduce the number of re-test the tester and developer has to be working together to prioritize and minimize the re-tests (Whittaker, 2000).

Phase 4: Measuring testing progress

Testers could not tell what the status of the testing is because testing is not measured in percentage but in found failures, or number of times the application run successfully etc.

But the found failures does not tell anything about the progress or the number of errors is good or bad. Because it is hard to measure the process if testing, are testers answering questions about structural and functional testing completeness.

Functional testing: is testing best on the formal specification. It will test the behavior and is also called black- box testing. To check the functional completeness an answer must found for the next questions; are all the possible manners the software could fail tested? Are all the inputs applied? Are all the possible scenarios that a user would execute tested? (Whittaker, 2000).

Structural testing: the input sources are based on the code structure and its internal data structure. This testing is testing the code and is called white-box testing. To check the structural completeness answer; are all the common error in the system found? Is all the code run through the test? And are all the placed errors found within the system? (Whittaker, 2000).

Chapter 4 Data and set-up

This chapter describes the creation of the testing data. Creating test data is one of the treatments as output from this research. Before the creation is the process, rules and allowed processing of sensitive data gather from Nivel. Because they have a years of field work experience. This is compared and improved with the Bachelor thesis from Domenico Essoussi (2019), also member of the C2R program. After the rules and processes is the experiment setting for the recording of medical consultations are described and explained. When the recording (real and simulated) is completed are the recordings run through the C2R prototype. The output will be analyzed and in chapter 6 be validated.

For the gathering information and answers about the recording of real-world medical consultation, informed consent and when is it allowed to use those recordings within C2R is an interview conducted with the coordinator of the research program Communication in Healthcare at Nivel. The reason is because Nivel record medical consultation for many years and have lot of practical experience. The results of this interview are shown in chapter 4.2.1 and 4.3.1.

Also, more information is needed about the metadata structure and the level of storing. For the gathering the answers is an information-/ collection specialist from the UU interviewed. The reason is to get insight in potential solutions that could be used in the situation of the C2R program. The answers from this interview are shown in the introduction of chapter 4.4. and 4.4.1.

4.1 Requirements Specification

As mentioned in the research method the design cycle of Wieringa is used. The second step is Treatment design where the treatment for created for the found problem(s). The problems we try to solve are the storage of the research data within Yoda, and the stored test data retrieve from Yoda to test the create prototype of Care2Report.

For the treatment are requirements defined. A requirement is a property of the treatment that is wanted by a stakeholder. It's the goal for the treatment that will be designed within this study. All the requirements need a contribution argument to define the reason of the choices (Wieringa, 2014). The requirements will be satisfied by the treatment. Requirements can be classified by priority, urgency or in functional and non-functional requirements.

“A functional requirement is a requirement for desired functions of an artifact. A nonfunctional property, sometimes called a quality property” (Wieringa, 2014).

The functional properties can be measured by executing them in a specified test. Nonfunctional properties are quality, this is measured by using indicators called metrics. This type of indicators are variables that can be measured (Wieringa, 2014).

For this thesis the requirements are divided for two treatments; the storage within Yoda and the test data execution within the prototype. Below are the treatment design requirements (RI) as described in the design cycle (Wieringa, 2014). The treatment requirements contribute to the goal of this research to get understanding and insights in the rules and processes that is required for processing and storing the multimodal testing data that is used for analysis and testing the created Care2Report system.

All the requirements are prioritized with the MoSCoW method, the larger letters are abbreviation for Must have, Should have, Could have and Would have. From right to left it is the need to fulfill the requirements within this research.

Requirements for the data management strategy

RI 1: The data management strategy must explain the data creation process for recreation. [Must have]

RI 2: The data management strategy must describe the creation process of the metadata for the findability and storage of the data. [Must have]

RI 3: The data management strategy should provide a complete data management plan about storage, retrieval and metadata of the research data. [Should have]

Requirements storage within Yoda corpus

RI 4: It must, be possible for multiple file types such as .MP4, .WAV and .MP3 to be stored in Yoda for testing. [Must have]

RI 5: All the recording data folders must contain the required metadata for findability and to make retrieval possible for users. [Must have]

RI 6: It could be possible to automatically transfer logs and intermediate results from C2R to Yoda to improve the gathering of files for analysis. [Could have]

Requirements Test data and execution

RI 7: The test data must exist of simulated and real situation of a medical consultation between a family doctor and a patient. [Must have]

RI 8: The test execution must create a an EMR and possible intermediate results all extracted from audio, transcript or video input. [Must have]

RI 9: The C2R testing should show the error rate of the complete EMR or the intermediate results such as transcript and triples when compared to the “golden standard”. [Should have]

4.2 Policies, codes of conduct and laws

The data creating within this research will recorded video and audio of the interaction between a patient and care provider within a medical consultation. The situations are simulated and real.

The aim of the preparation is to create a guideline for recording persons communication within healthcare. To do this, is to gather the policies, rules and process for recording and processing the videos and audio from care providers and patients in real and simulated situations. Therefore, guidelines are setup for conducting research on healthcare communication. The guideline can be described as a statement or rules to determine an action. The aim of the guideline is to help the process streamline.

To record the most reliable and representative audio and videos of human communication within the healthcare it is important to capture the complete interaction in an inconspicuous way possible. For the execution of the recording is creating your own guidelines fundamental instead of leaning on experience of other researchers (van Dulmen, Humphris, & Eide, 2012). The guidelines can used in practice for quality control and/or checklist of setting up a research within healthcare communication.

The guidelines must cover the relevant steps are;

Recording purpose; within this study is not to record multiple sessions of the same patient to study the patient's improvement or look at the doctor – patient interaction. For this research is it important to capture the full (real and simulated) doctor – patient interaction. Within the paper van Dulmen et al., (2012) an example is show of guidelines for recording health care encounters for research.

Patient recruitment; For the recruitment of participants (patient and family doctor) could be the recruitment procedure opt-in and/or opt-out been used according to van Dulmen et al., (2012).

Opt-in is when participants are agreeing to participate in advance of the research. Participants are invited into the study. Opt-out is recording patients unless they explicit object to the recording and the use of it. There is a hybrid form between those procedures. When there is planted to create series of recordings, participants are invited with the opt-in procedure and in the next sessions is the opt-out procedure used. In both the opt-in as the opt-out is it required to let the participants sign the informed consent before recording can be started. After sign the agreement there must be a reasonable time period to withdraw their participation (van Dulmen et al., 2012).

Care provider and patient **privacy and time investment;** The privacy of the patient and family doctor that participate in the recording must be protected. An example is to make the patient not visible of film them only from behind, for the storage try to make it anonymous, encrypted and coded and observations must sign a confidentiality agreement. For the participants is there the informed consent, it must tell them clearly that the recordings only going to be used within the research and not for public viewing. The informed consent must be drawn up generic so the recordings could be used within a secondary analysis, without the need for a new permission. Because there is a so called “cooling off” period where the participant could withdraw their recordings is it important to tag or give it another identifier so it could easy be removed (van

Dulmen et al., 2012). To keep participants (patient and doctor) motivated about their involvement in the research it is important to keep the time investment as low as possible.

Ecological validity and representativeness; Ecological validity is the degree in which the recordings within the research are equal to the real-life. To make sure the recordings are ecological valid the recording must not disturb, and the healthcare must continue as usual. There are no signs that recording a consultation will change the behavior of the patient or doctor. The decision to use video and/or audio relay on the research question and the focus of the study. Video is used to study non-verbal behavior, audio is used to study the words and emotions within the voice (van Dulmen et al., 2012).

Data observation, storage and use; There is an important decision to make what observation scheme is going to be used. There are the levels macro or micro. The first is when there is a focus on a specific behavior and the second is coding every verbal utterance (van Dulmen et al., 2012). To record in a safe way it is important that the observers have a high reliability and have the know-how to behave during a recording. When the recordings are done it is possible to share them with a third party. If not conflicted with the agreement between the researchers and the participants. The recordings are stored, it is important to think about the accessibility, storage and duration (van Dulmen et al., 2012).

4.2.1 Rules and processes for testing privacy sensitive data

The complete process of recording a medical consultation with real and simulated patients is gathered by conducting an interview the coordinator of the research program Communication in Healthcare at Nivel. See attachment 2 for the used questions.

The process of recording

For the process there is a “playbook” developed that describes the complete process of recording a consultation. The playbook describes what to bring with you, prepare the recording equipment, how to approach the patient and how to handle the data when returned to Nivel. Just like the paper describes above is they use as less people as possible during a recording. Every person more could disturb the natural way of patient – doctor interaction. So, the camera is always unmanned. Related to the privacy most of the time is the patients face not visible recorded. They are recording diagonally from behind or the face is blurred. It is possible to deviate from the standard for example for analyzing the patient face expression. It must be indicated in the informed consent to the patient.

Analyzing, edit and processing the recording

When the informed consents are signed and the recordings are made, the next step is to analyze the recordings. The recordings can be analyzed or processed by software. As long as it is within the informed consent. This document tells the participants what purpose the recordings serves and who it is going to be processed.

Storage of the recordings

The data within Nivel is sensitive real medical consultation recordings. This is during the recording stored on the memory (card) inside the camcorder. The recordings are later transferred to a separate server. This server is only accessible through a pc in a locked room.

The recordings within this research are simulated or with a real doctor. This data does not need this kind of protection. The data have to be anonymized/ pseudonymize where information about the participants reference to a codebook. The codebook is stored in another place and will contain the information.

Solutions for future collaboration

Within the conversation with Sandra van Dulmen are solutions proposed for the storage and gaining of real patient data. The first solution is to store the created real patient data on location by Nivel. The separate server satisfies the requirements for save patient data storage. The disadvantage is that we have to do all the testing on location of Nivel.

The second solution is a portal to the stored recordings by Nivel. They are working on making a set of the recordings available for medical doctors in which they have participated. The idea is behind it is that the doctor could review and judge their conversation with a patient to learn from. To do that Nivel makes a portal for the doctor to login and review from location. Within this process it is possible to make a portal for Care2Report so the recordings could be used for testing the system.

Third collaboration is to record with the help of Nivel in the Sint Maartens clinic. When Nivel is recording they adjust their existing informed consent. This will allow us to use the data and process it in software.

4.2.2 Using sensitivity test data from real patients

Processing sensitive data in automated medical reporting in real-world situation.

The General Data Protection Regulation (GDPR) is a fundamental right of the European Union. It means everyone in the EU have the right of protection of their personal data. Essoussi (2019) researched the seven 7 GDPR principles that are needed for C2R when used in practice to process personal data. The next steps do not apply to the test data.

1 Lawfulness, fairness and transparency

When personal data is going to be processed it needs to be lawfully, fairly and transparent. It is normally unlawful to use and process data that is concerning a patient's health, but healthcare providers invoke on an exception that is parts of healthcare services.

Important to be lawful the participant must give their consent to participate without any pressure or tricks by using ambiguous consent. Its patient must be informed about the reason of the research, where the data is going to use for and his possibly to withdraw his personal data. As described in sub-part above is the main ground of processing data lawful, is that the personal data could not be used other than the agreement. Do not unnecessary processing the health information.

Fairness is that when the patient's data that is processed it is done in a way that patients can reasonably expect. This means in not a harmful or deceitful way.

Transparency about the processing activities is the patients right to be informed and to know about deleting their data. The exact articles are described in the thesis of (Essoussi, 2019).

2 Purpose limitation

As addition on lawfulness and agreement with processing within Nivel is that the care provider must clearly identified the purpose of the documentation and the processing of the data. If the processing differs from the original consent, this is not allowed unless the patient explicitly give consent (Essoussi, 2019).

3 Data minimization

Within the recording of a real patient the gathered data must be adequate, relevant and limited to the purpose. The care provider must define the amount of data that is needed for the C2R system. Also, the data that is processed must be contribute to the purpose that is defined (Essoussi, 2019).

4 Accuracy

This requires that all the data is up to data and accurate. The patients have the right to rectification, that means the data about them is accurate and complete. This is about the data that is stored (Essoussi, 2019).

5 Storage limitation

The data is no longer stored than necessary. The C2R software collect multimodal data to generate an EMR. When the report is generated the data needs no longer to be stored (Essoussi, 2019).

6 Integrity and confidentiality

The patient data must be processed is technical and organisational measures to prevented unlawful processing, accidental loss and destruction or damage. The three most breach causes are theft, access or disclosure by someone unauthorized and IT hacks. Confidentiality requires that only the entities that are authorized can getting access to the sensitive and protected data. The data is available for the persons that are authorised for the processing the data (Essoussi, 2019).

7 Accountability

The healthcare provider is accountable for the for the compliance of the principles above. (Essoussi, 2019).

4.2.3 Rules and processes of using research data within the University

Beside the rules and process about handling sensitive that are gather from Nivel and jurisdiction are also the standard rules and processes within the Utrecht University explored.

Informed consent

These days research project is not only need consent about the gathering of the data from a participant also a consent in the storage and sharing the gather data. The informed consent should mention the likelihood of the data to be shared, published or stored for long-term for the reproducibility. Just like the jurisdiction of Essoussi (2019), Utrecht University (UU) states that for specific use only the data limited to the necessary data.

Storing and preserving data

For the storage of research data time and frustration can be saved by using the proper structure and annotation. What is considered under properly stored data is 1) the storage location. The tool used within this research is Yoda, a data management solution developed by UU. It is suitable for storing large amount of research data. 2) make back-up, versions and store them save. Protect the raw data, different versions and back-up by storing them on a save place, but also keep them separate. 3) structure the names and the folders. Just storing is not enough. The folder must have a structure and the files a logical name. The file name can build form element such as project name & number, research group, type of measurement, subject and date of creation. Keep the elements coded to keep the name short. 4) for understanding and findability of the data is documentation and meta data attached. Documentation is human readable, and metadata is computer readable. Meta data is a form with field that can take (fixed) values. Both can describe subjects of measurements or situations that are obtained. There are multiple goals of meta data with different purposes. Descriptive meta data is used for finding and reusing, it describes the author, tile, abstract, time period etc. For managing the data, the administrative metadata is given like size, access rights about the data. Structural metadata such as version, related project and content tells about the context of the data. 5) Use standardized formats to store the data. The reason behind it is that it can be widely accessed now and in the future. 6) security of the research data is imported, first decide what data needs to be stored. Protection of the data files could be done by using, encrypted data, using legal agreements and destroying or software erasing of data. If the data is processed or stored on a computer than needs the computer system security in the form of anti-virus, passwords and secured wireless network.

Persevering is storing the research data for a longer period about the research has been performed. The Dutch code of conduct requires that research data must be preserved for at least 10 years. The Utrecht University research data framework described in chapter 3.5.3 increase the number to 15 years for medical records based on Wet Geneeskundige Behandelovereenkomst WGBO (article 454) (Francissen, 2004). This sounds a bit conflicted with the GDPR, this states that data about a person may not kept longer than needed for the aim why the data is collected.

First has the data that must be preserved be specified. The data that is stored is not only the data for the publication but all (raw) data for the entire C2R program. All the data, documentation and files that will be preserved arranged in a data package. For sharing of data must be done in the rawest version possible. Documentation will help giving a better understanding of the data. A code book will explain the variables in the data. A metadata sheet gives an administrative description such as file format, authors, title and date. When the data is preserved the next step is to make sure the raw data could not be overwritten or deleted by accidentally. In the Yoda solution C2R is using is it possible to store data in a fault. Other solutions are making files read only or to log all versions.

4.3 Data creation steps set-up, process and recording of the medical consultations

In this section the execution of the experiment is described so it is in the future reproducible by others. Explained are the necessary actions, processes and rules for the recording of a medical consultation. First is the setup that is used for the real-world recordings and the steps a consult should follow according to the NHG. Second is the recordings that are created and used within this research, with all their file types. Third, is explaining how the transcript format is used by the C2R system and the errors could arise when the wrong format is entered. Fourth, is an explanation of the RIAS code and where they appear within the SOEP. Two general practitioners have determined which RIAS codes are most important. Finally, an PDD give given as an overview of the process and deliverables of the creation of the recordings.

4.3.1 Recording medical consultation

Positioning camera(s): The positioning of the camera in the different in the real situation and simulated situation. The real situation are two cameras used. The reason is that the main camera records the conversation and the second camera the examination for future action recognition.

The setup: Height: 1,5 m height so the faces and body are within the view of the camera.

Angle: The angle of the camcorder is 90°. The telephone has recorded the examination itself with the aim for action recognition. The angle of the telephone was approximately 45°.

Nivel state that when it is not needed to see the patients face (for example research of facial expressions) film over the shoulder of the patient. The limited space within the family doctor's office makes it not possible to film from behind.

Consideration is made, to film from the side. The conversation is recorded in a useful way for this research and the patient does not share real sensitive information.

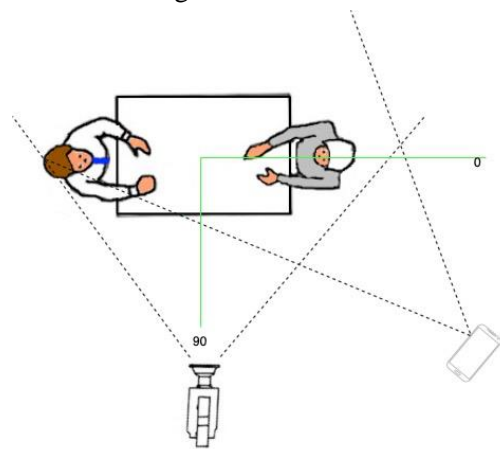


Figure 13 setup recording real situation based on (The University of Texas at Dallas, sd)

The setup of the simulated situation is placed under the condition of Nivel. The patient is recorded from behind with an angle of approximately 45°.

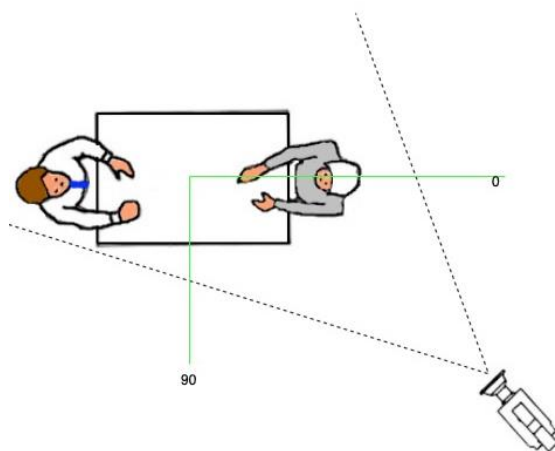


Figure 14 setup recording simulated situation based on (The University of Texas at Dallas, sd)

Prepare consultation

The consultation is focused on ear infection and other ear difficulties. The recordings with the real family doctor are guided by the doctor. The reason is that we could simulate a real medical consultation.

For the simulated recordings was preparation needed. The first step was NHG (het Nederlands Huisartsen Genootschap) to research the (treatment) guidelines. This is a union that create the guidelines for the family doctors. The guidelines contain the possible complains from the patient, where to look for during an examination and options of treatment to recommend. In the guidelines for ears is divided in the next four problem areas:

- Otitis externa
- Acute otitis media (AOM) by children
- Otitis media with effusion by children
- Hearing impairment

The focus of the prototype is Otitis externa. Short description of the illness is as follows:

“External otitis: diffuse inflammation of the skin of the ear canal with pain, itching, discharge, scaling, redness or swelling, possibly in combination with hearing loss” (NHG-Standaarden voor huisarts 2009, 2009)

The guidelines for the family doctor that is recommended to follow during a medical consultation by Otitis externa.

Anamnesis: the patient will explain their medical history. The history contains both similar earlier problem as the current duration of the illness. If the doctor suspects Otitis externa he will ask about:

Earache, itching in the ear, fluid from the ear, hearing loss, duration of the complaints and recent cold. About the history previous (middle) ear complains and ear surgery in the past or hole in the eardrum or eardrum tubes (*NHG-Standaarden voor huisarts, 2009* and Rooijackers-Lemmens et al., 1995). The NHG 2009 is the most recent but not an publishes paper that’s why it is supplemented with an older paper about the NHG.

If it is more frequent or the duration is more than three weeks, ask also:

- Connection with swimming
 - Irritation from ear cleaning
 - Use products (make-up, shampoo etc.)
 - Presence of psoriasis from (seborrheic) eczema.
- (Rooijackers-Lemmens et al., 1995)

The most frequent complain with 40% of the patient is pain within the ear. Less frequent is itching, fluid and feeling clogged (Rooijackers-Lemmens et al., 1995).

Physical examination: The physical examination by a suspected ear infection the family doctor will first examine the “healthy ear” and after the affected ear. Things to look for are the ear shell, scars behind the ear and pain by carefully pulling. The otoscope is used to examine the inside, to look for are; swelling, flaking, redness, otorrhea, vesicle, erosions within the ear canal. On the ear drum shown signs of inflammation and it is not damaged. If the ear drum could not be examined it has to be cleaned. After three weeks the doctor collects a material from the ear to check for bacteria.

Evaluation: are the findings of the physical examination. The presence of ear pain, itching, fluid in combination with swelling, redness or flaking of the ear canal by the physical examination indicated to Otitis externa (Rooijackers-Lemmens et al., 1995).

Guidelines politics: are for the information and treatment of the patient.

Information: inform the patient what is Otitis externa, the cause is not always clear. If the treatment takes more than three weeks there are triggers. Non-drug advice is advice about the reason of the emergence and how. Cleaning: when the ear drum is not visible for examination it must be cleaned. Drug treatment: tell the patient about the use, possible effects of drugs (*NHG-Standaarden voor huisarts 2009, 2009*).

Reduce swelling: the care provider can make the choice to reduce the swelling by using an ear tampon. The doctor will submerge the ear tampon with ear drops that will reduce the swelling. Remove after 24 hours.

Control: there is another appointment necessary if: 1) the symptoms are not disappeared after one week and 2) if the ear tampon does not have any effect or the patient cannot remove it by their self.

Refer to a specialist: if the problem is too complex or urgent the patient can be sent to a specialist when: 1) the symptoms are not cured after five / six weeks, 2) no acceptable results after multiple ear infections. 3) if there are symptoms of otitis externa by elderly or people with reduced health. 4) when the patient has otitis externa with a fever, gets the treatment flucloxacillin and his condition does not improve within 48 hours (*NHG-Standaarden voor huisarts 2009, 2009*).

4.3.2 Creating test data

The multimedia video and audio are recorded from real situations where the family doctor is a doctor in practice and the patient had otitis externa symptoms in the past. The family doctor works in a practice in Amsterdam West for about 28 years. For the recording is executed the informed consent is signed so the recordings are stored safe and can be used by the C2R program. The informed consent form is attached in attachment 3 Simulated scenario are two participants that play the role of family doctor and patient. The consultation is led by the output of the NHG as described above in the *Prepare consultation* and the SOAP format.

Overview of all the created recordings

All the real scenario recordings are following the anamnesis, physical examination, evaluation and guidelines politics. The three real-world recordings exist of two recordings with ear pain and one with shoulder problems. The reason is that the next step within the C2R prototype are shoulder problems.

The simulated recordings are following also the anamnesis, physical examination, evaluation and guidelines politics structure as good as possible. All the recordings are about ear infection. Recording 1,2,3 are about Otitis externa and 4 and 5 about media ear infection.

Table 4 Overview of the recordings

#	Description	File type(s)	Consultation type
1	Otitis externa	.wav .mp3 .text	Real-world
2	Otitis externa	.wav .mp3 .text	Real-world
3	Shoulder problem	wav .mp3 .text	Real-world
4	Otitis externa	.wav .mp3 .text	Simulated
5	Otitis externa	.wav .mp3 .text	Simulated
6	Otitis externa	.wav .mp3 .text	Simulated
7	Otitis externa	.wav .mp3 .text	Simulated
8	Otitis externa	.wav .mp3 .text	Simulated
9	Otitis externa	.text	Nivel
10	Otitis externa	.text	Nivel
11	Otitis externa	.text	Nivel
12	Otitis externa	.text	Nivel
13	Otitis externa	.text	Nivel
14	Otitis externa	.text	Nivel
15	Otitis externa	.text	Nivel
16	Otitis externa	.text	Nivel

Preparing the video transcription

The recordings are created with the least needed participant and people as possible. This means that the most recordings are done in a sequence in one day. Like Nivel proposed are the recordings done without a camera man. When starting the recording button is pressed and walk through the shot to take in the starting position. When the consultation is over that same happened. So, when the recording is finished the walking by in the begin and end are cut for the transcript.

Within the conversation with Sandra van Dulmen (Nivel) is discussed the subject face blurring. In a real doctor-patient consultation is the patient recorded from behind, when the situation occurs that filming from the side is needed the patients face will be blurred. This to protect this privacy and personal data. Because this is a real doctor but a simulated patient. Experienced otitis externa in the past but not during the recording is it not needed to blur the faces.

Create a separate audio file

After the recording are separate audio files created. The audio is used to transcribing the consultation and it will be used as input when testing the C2R prototype. The current prototype accepts only audio or transcript for further processing.

4.3.4 Creating transcript of the recordings

The transcript is a translation of the speech to text. The transcripts in attachment 4 and 5 are placed in the structure that is required by the C2R system see table 5. The structure is selected for a future functionality to make consultation timelines. The time implies when a sentence occurs within the sentence. The abbreviation CP stands for general practitioner and PT for patient.

Table 5 Correct input structure of the transcript in the C2R system

Time (min)	Time (seconds)	Person (speaker)	Sentence
04	31	CP	Please explain why you're here today?
04	45	PT	I have pain in my left ear.
04	53	CP	Can you tell me how it started?

The C2R system reads the transcript like a table (see table 5), the first column is the time in minutes, second time in seconds, third the abbreviation of the person and the final column is the sentence they say within the conversation. If the columns 1 till 3 not filled in the system will fill them up with the first words of the sentence. The system will fill every column with the first value in line independently from the length or significance. The problem will occur that important parts of the sentence will not be used in the transcript and subsequently neither in the triples. In table 6 are incorrect way of creating a transcript examples given with the same sentences as table 5. The first sentence in table 6 shows when the time is entered in one column, and the second and third shows when the time and person are left out completely.

Table 6 Error in interpret the transcript by the C2R system

Time (min)	Time (seconds)	Person (speaker)	Sentence
04:31	CP	Please	explain why you're here today?
I	have	pain	in my left ear.
can	you	tell	me how it started?

The transcript itself is created by hand. The reason is the transcript functionality in the prototype does currently not work. It is only possible to transcribe small fragments of text. The options that have been tried are:

- Full audio file for transcription
- Fragment of the audio file for transcription

Also, beside the C2R system is also looked into other possibilities to create a transcript from the recorded audio.

- Generate transcription with <https://sonix.ai> comparable with Google speech to text. Had a very poor execution with Dutch text.
- Google speech to text online version. The max audio length is 60 seconds and gives an error back.
- Google speech to text API is available but an application must first be built for it

In possession of C2R are also transcriptions from Nivel used in the Bachelor thesis from van der Kooi and Lim (2019). This transcript is placed in the RIAS code. The transcript is recreated and placed in the format that is used by the C2R system.

4.3.5 Finding important words and sentence structure

In this sub-chapter the structure, words and relevant sentences within a medical consultation will be analyzed. Also, the internal word structure within the consultation is visually displayed. For this part is the bachelor thesis from van der Kooi and Lim (2019) used, also part of the C2R research program. Only some of the main points are discussed here.

The global structure and content of medical consultations:

The global structure of a medical consultation according to Roter, (2006) has a certain pattern of; 1) opening/ greeting, 2) talking about history 3) physical examination, 4) counselling/advise 5) closing the conversation. The parts within the pattern can be sub-divided in specification shown in figure 15.

Part of Consultation:	Specification
Opening	Greetings
	Initial Physician probes
History-Taking	Follow up on medical problems
	Personal and family medical history
	Medical background
	Precious treatments
	Other lifestyle/psychosocial concerns
Examination	Physical exam procedures
Counsel	Information-giving or counseling regarding patient's medical condition
	Propose therapeutic regimen
	Lifestyle / Psychosocial information or suggestions
Closing	Introduction of new problems
	Closing of consultation

Figure 15 comparison general consultation structure vs RIAS code made by van der Kooi & Lim (2019) based on paper (Roter, 2006)

Not every care provider keeps strictly this order when conducting a consultation. it is up on the care provider to maintain this structure or switch consultation parts or even combine them. The most consultation parts that are combined is the history-talking and the examination (Andeweg, Terluin, Boendermaker, & Pols, 2002)

SOAP/ SOEP

SOAP is an acronym for Subjective (about the patient history), Objective (the examination of the patient), Assessment (diagnosis) and Plan (treatment steps). In Dutch its more or less the same, assessment is called evaluation. Both have the same meaning; evaluation is also about discussing the diagnosis of the examination. SOAP is used for the fixation of the medical consultation in the EMR. As noted, the general structure is comparable with the SOAP (See table 7).

Table 7 Comparison general structure and SOAP for a medical consultation

SOAP	General structure
Subjective	Opening/ history talk
Objective	Physical examination
Assessment	Counselling/ advise
Plan	Counselling/ advise Closing

Subjective

The subjective portion of the SOAP format contains the information from the patient about the how he is feeling, their concerns, thoughts and problems recorded. This should be a briefly as possible, but the patient's perception of the situation and problems should be clear to an outsider. There are mostly short words or abbreviations used and less quotations of a patient. Unless it is recorded by video or audio. When quotations can be used is when the patient words are suicidal, shifts in their well-being or when their behavior is non-conforming their appearance. It is having also a non-medical use when the patient uses aggressive or abusive language towards the care provider (Cameron & Turtle-Song, 2002).

Subjects that also must be note are when a patient is confusing about place, date, person etc. and when he/she is unable to follow the conversation and looks confused. Also, positive and negative changes from counseling before is important information to note. Example "*the therapy is working. I feel much better over myself now*" – note therapy is working. The aim of subjective is to reflect the patients concerns and problems (Cameron & Turtle-Song, 2002).

Objective

In the objective part of SOAP are the measures or facts such as seen, heard, smelled, counted or measured by the care provider. The different types of objective are observations of the care provider or external written material (reports or consultations by others). When note the objective words as appeared and seemed must be avoid. When not, a final statement can be note it must be justified. By justification is the phrase "evidenced by" useful (Cameron & Turtle-Song, 2002).

When noting observations must personal opinion be avoiding, and professional observations be included. Instead of describing a patient as not- cooperative describe why the patient is not cooperative. The next reader must draw their own conclusion of the consultation (Cameron & Turtle-Song, 2002).

Assessment

The assessment part of the SOAP is a short recap of the findings from the care provider findings and thinking about the situation and problems. The clinical impressions of the consultation without supporting descriptions are used to rule out or rule in a diagnosis. The impressions can help creating a decision tree for the diagnosis. Also gives it an outsider the possibility to see the thinking steps. Before making aa diagnosis the care provider has to ask their self is the data supporting the diagnosis or is all the collected data sufficient for a diagnosis? (Cameron & Turtle-Song, 2002).

Plan

The final sections are the Plan. The plan uses the findings from the consultation (Subjective, objective and assessment) to develop the plan. It consists of the action plan and prognosis. Action plan is the information about the consultation such as data, duration, given instructions, progress of the treatment. The prognosis is a forecast of the possible improvement a patient is going to make based in the diagnosis (Cameron & Turtle-Song, 2002).

RIAS: Ratio of utterances in a medical dialogue

The bachelor thesis from van der Kooi and Lim (2019) describe the relevant and irrelevant sentences within a Dutch medical conversation. The (ir)relevant sentences are marked with the RIAS code. The Rias code is the a commonly used medical interaction analysis system (Ong et al., 1998). The spoken utterances from the patient and provider are labeled in one of the 41 categories. Not all the categories are being discussed only the ones that are marked as relevant. The RIAS method is based on medical conversation for communication analysis. The utterances during the conversation are divided in one of the 41 categories (Ong et al., 1998 and Roter & Larson, 2002). By van der Kooi and Lim (2019) made a comparison between SOAP and RIAS (see figure 16). This shows with RIAS code are used within every step in SOAP. The objective is let empty because the assumption is that there is no conversation within the medical examination.

SOAP	RIAS categories according to the literature
Subjective	Gives information - psychosocial & Asks questions - Psychosocial & Gives information - Lifestyle & Asks questions - Lifestyle & Gives information - Medical condition & Asks questions - medical condition
Objective	-
Assessment/Evaluation	Gives information - Medical information
Plan	Counsels - Lifestyle and Psychosocial & Counsels - Medical condition/Therapeutic regimen

Figure 16 Comparison between SOAP and Rias made by van der Kooi & Lim (2019) based on the paper (Roter, 2006).

Relevant sentences with RIAS code

The sentences that are marked as relevant is done by two family doctors, experts in the medical field. For now, there is no other way to validate the transcripts in a better way. The RIAS codes they think is important are:

Gmed	Gives information - medical condition
Para	Paraphrase/checks for understanding
Qmed	Asks questions - medical condition
Gpsy	Gives information - Psychosocial
Gsoc	Counsels - Social
Cmed	Counsels - Medical condition/Therapeutic regimen
Reas	Reassures, encourages or shows optimism

4.4 Database design in YODA

Introduction of Yoda

Yoda is an abbreviation for Your Data created by the University Utrecht. Yoda is built for the storage, sharing, manage, secure add metadata and publish of (research) data. Yoda is designed to store all kinds of data raw data, secondary data but also experiments (output), questionnaires, digital sources, video, audio etc. The data storage is in collaboration with fellow researchers. The data and metadata can be stored for at least 10 years. Yoda is adding value to reusing, retrieving and protecting the data.

Yoda is built on IRODS an open source data management software. One of the main reasons is it is easy scalable, and it can handle billions of small documents or also a petabyte of data.

IRODS is selected for many reasons, the challenge was to create a storage solution that can be used for collaboration with other researchers and countries. It must have an ease of use for the researchers and must be like a Dropbox on steroids. It must also meet the requirements of the Utrecht University's research data policy of storing the data to a minimum of 10 years after the research. Therefore, are well created data packages and reusability to the data. The FIAR principles help to guidance the data decryptions (Smeele & Westerhof, 2018).

Yoda satisfied the FAIR principle. FAIR stands for Findable, Accessible, Interoperable and Reusable

- Findable: the data must be findable for all the researchers within the research group. To do this Yoda makes it possible to add relevant metadata.
- Accessible: The shared data must be accessible for all the authorized research in the group. Before a research is able to access the data, Yoda provide a reliable access control.
- Interoperable: Yoda used a standardized (machine-readable) metadata schema for the interoperability.
- Reusable: it is possible to share that after publication or closed access. The research group is always in control of the data.

Layout out of Yoda.

The hierarchy of Yoda say something about the levels that are available. The highest level is the Yoda instance, this is an existing or a new to create environment. The Yoda instance we use is <https://science.yoda.uu.nl> (portal) and <https://science.data.uu.nl> (Network drive map) from the category Science. To work and share with different researchers to the same Yoda implementation the subcategories is introduced. The subcategory has a group this is on research project level within a selected category. The group manager can add fellow researchers to the user group. Every participant can add or deleted a folder or the files within them. Below an easy overview of the Yoda distribution.

- Yoda instance
 - Category
 - Subcategory
 - User Group
 - Folder

For the folders it is important to give a logical structure and logical naming convention. The structure for the C2R Yoda storage is described below in *Database design discussed with Otto lange*.

Database design discussed with the information-/ collection specialist from the UU

Because C2R program is generating lots of different data there is a decision made on with level the database has to be divided. The level that is the most convenient is experiment level. Also, the functionality of the data is been separated. So is testing recordings separate from learning videos for action recognition.

Chapter 4.3.6 is about the created data set for testing it C2R prototype. Here is an experiment for the test data is a single or multiple recording(s) of the same medical conditions. An example is ear infection, middle ear infection or shoulder problems.

Chapter 5.4 is about the storage of action recognition data set. The data set of action recognition is divided in annotation and videos. The annotation is used for labeling the frames within the video. The Videos is divided in experiments. An experiment of action recognition could be abdomen, auscultation, percussion, palpation, blood pressure and heart & lung auscultation. An experiment for action recognition is one or multiple recording(s) of the same medical actions.

In chapter 5.4 is the Functional architecture designed for the data extraction will be extracted from the C2R system after testing. That test and intermediate results will be stored in a separated file within Yoda. The reason is that the results must be analyzed for further understanding. The Resulted are being replaced in a storage file that holds all the test results. The testing file is again used for the collection of the output data the C2R prototype is producing.

Figure 17 shows the Overview of the current Yoda database. The real recording transcription contain the format .txt transcript that are used by the C2R system. Because there is no separate file for the output testing results, the results are place with their transcript. Also, in the file “Laura” are the recordings and annotation for action recognition. In Figure 17 an overview of the Yoda files.

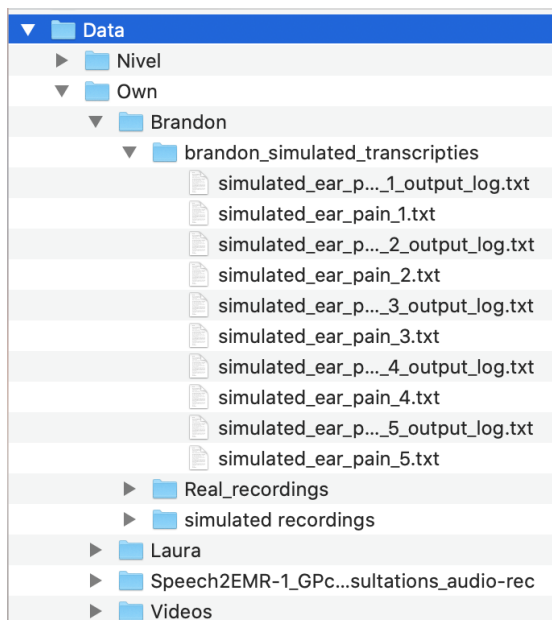


Figure 17 Overview of the current Yoda layout

4.4.1 Metadata structure

A short introduction of metadata handling by IRODS/Yoda. The metadata is registered in an XML file that is attached to the data in the data package. The XML file(s) are compliant to the schemes created by Yoda/ Utrecht University (Smeele & Westerhof, 2018). Yoda have created its adducted metadata schema in cooperation with different departments within the Utrecht University. The standardized schema does contain metadata field for all different departments. The metadata schema consists of two files, XML contains the metadata (values) and the XML Schema Definition (XSD) is for the restrictions such as required fields to fill in, length and data type (Smeele & Westerhof, 2018).

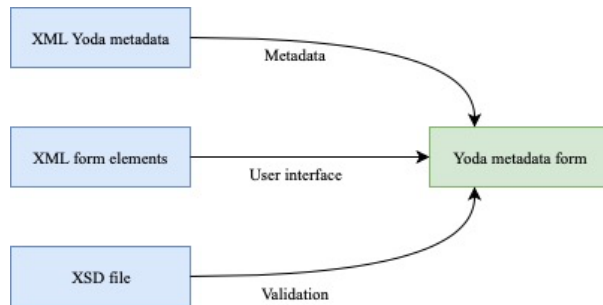


Figure 18 The metadata from construction of Yoda (Smeele & Westerhof, 2018).

Above in figure 18 is an overview of the metadata construction of Yoda. The XML metadata from Yoda contains the specified metadata. XML form elements presents all the information to the web portal metadata form and specifies the required fields and XSD as mention above the restrictions for the validation. Within Yoda is it possible to add metadata to on different levels. It can be assigned to different levels, such as on document level, folder level or group level. The metadata will give more information about the created documents for future research and also it will improve the retrieval of documents. Besides that, it is also a requirement from the subsidy providers to have the created data and the corresponding metadata to be in place. The structure of the metadata within Yoda explained below.

Descriptive: The descriptive part gives the data a descriptive information and a field to describe the data. Also is it possible to maintain version management. An import field is about adding the date of the data gathering and their location.

Administrative: In the administrative part is the retention period state. It is important to know how long the data must me captured. The reason can be added in the field. Possible reason to have store the data for 10 years could be of publication or requirements from the funders.

Rights: The rights will be collected information about the creators and contributors to the data package. Beside that the license of the data is be attached. Finally, the access of the data is determined.

Within Yoda the metadata that is assigned to a folder is automatically taken over by the sub-folder. In the portal (<https://science.yoda.uu.nl>) is the metadata added and appointed to a folder or document. In the network drive map (<https://science.data.uu.nl>) is the metadata stored in an XML (in the future version of Yoda in JSON) file. The file can be copied and been attached to other folders or documents, the metadata is then taken over.

An addition to metadata is the codebook, this is a document that will gives more insight into store data. The metadata is mostly fixed or somehow limited. It is also possible that the data is to sensitive like, names of participants. The (meta)data that is placed in the codebook is keeps separate from the stored recordings and metadata file. The noted information will be interrelated by unique numbers. To create a codebook, it should describe the setup of the research, the dataset and its variables, used units, the sampling method and size and setup of the experiment (Smeele & Westerhof, 2018).

Chosen metadata structure with Otto Lange (metadata specialist UU)

The metadata structure that will be used for the C2R program is discussed with Otto Lange metadata is information-/ collection specialist within the UU. In the C2R program there will be different kinds of data collected. Because of the amount of data is decided to add the metadata on experiment level (folder level). Multiple experiments recordings can be gathered in one experiment (sub) folder. Figure 19 gives an impression of the number of videos within a sub-folder.

In the metadata field is the option to define how the data package is related to other data packages (folders). This is to define if there are part of a larger package.

Codebook

The metadata fields are fixed for the reason that all the needed data is filled in. Yoda created the fields are made in cooperation with different departments, so every study has the required boxes. If needed the metadata field can be rearranged or adjusted by the developers from Yoda. A cross reference is made to explain what is executed in every recoding (video/audio/measurement) is a code book created. The codebook will give background information about every recoding individual.

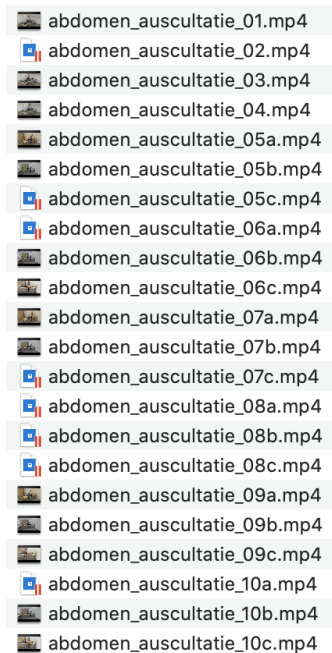


Figure 19 Overview of the videos within a sub-folder

4.4.2 Data management plan - (DCC)

The data management plan is a document where all the created/used documents, data, images, video, audio within the research are described. It exists out of six sections namely; Data collection, Data documentation, Data storage, Data security & Privacy, Data selection, preservation & sharing and Data Management costs and resources. The DMP is a dynamic plan, during the research it can be adjusted, updated and improved. The reason is that not all the questions can be answered in the beginning of the project. Every section contains questions that can be answered, together it creates a proper data management for scientific research.

Research data life cycle

The research data life cycle as shown in figure 9 show the 6 steps that are corresponding with the DMP. The steps from the life cycle begins just like the DMP with the collection and documenting of the created and used data. After that the storage and security of the sensitive data is discussed, where last the preservation, access and sharing is defined. The DMP ensures that the research results, used data is FAIR; findable, accessible, interoperable and reusable.

DMP template

Below are the questions of the Data Management Plan answered that are part this research. The complete DMP on <https://dmponline.dcc.ac.uk/plans/46431/overview> is for the Care2Report program. The DMP template consists of the following sections:

Data Collection

The methodology of how the data is collected or created. The test data is created by recording real and simulated consultation. The consultations are transcript manually as input for the C2R prototype, the output results are stored. Within the DMP online is an overview of all the data and datatypes that are created or gathered. The types are video (mp4), audio (mp3) and intermediate results, logs and system output (.text).

Data Documentation

Description of the metadata that is used to make the create data reusable and interoperable. In chapter 4.1.1 is described what metadata is added. In attachment 6 is an overview of the metadata from with the metadata tags and descriptions that is used for all the data types within the data collections as described above.

- Data item #1 .WAV audio files
- Data item #2 .MP3 audio files
- Data item #3 .MP4 video files
- Data item #4 .text transcript files
- Data item #5 .JPEG

Data Storage

All the data is stored on Yoda on a private category that is only accessible for the research team. There could be a back-up been made in Yoda itself. When (raw) data is placed in the “vault” on original copy is made that could not be removed or adjusted. The location that the data is stored on Yoda is category: science (science.data.uu.nl), in project: research-care2report.

Data Security & Privacy

For handling personal data compliant to the GDPR is the data encrypted. Yoda stores the research data encrypted on multiple geographical locations. Also is within the UU is Boxcryptor available, it is a service for the Utrecht University researchers for encryption of their data. Another manner the data is made more safer is to use pseudonymization and to write personal information in a codebook that refers to the to the actual document. Those are stored separately on separate storage environments. A less technical solution is to sign informed consent forms to protect the desires and demands of the participants to use the gathered data only for what there is agreed to. In Yoda it is possible to give research different access roles. Every role comes with their own rights. Unfortunately, the lowest right is write only, but it is still possible to download from Yoda.

Data Selection, Preservation & Sharing

The data that should be preserved and shared is very divided, from meeting presentation to data used for publication.

Not all that must to be preserved for 10 years. The that is for now stored in the project environment where all data could be adjust, data be removed or added. For longer preservation of the data it is transferred to the vault. Such as old meeting presentations are not a must.

Data Management Costs and Resources

The cost for the resources for the storage of the research data is mostly the cost that is charged per terabyte of data for Yoda. Also, the professionals that must be hired for organizing the data such as; research data management support, metadata specialized and a data manager.

Chapter 5 Case studies

In this chapter are two different case studies executed. The first case study is about the test environment and the testing of the prototype. The second case study about the storage of the used recordings for the training, testing and validation of the machine learning neural network.

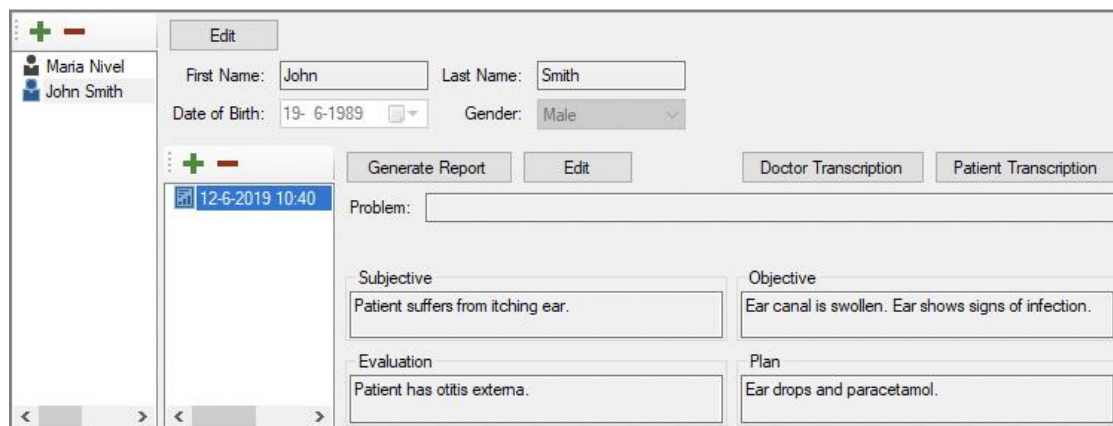
5.1 Execution C2R software with simulated and real medical consultations

According to the literature review chapter 3.6 Testing environment for software testing, the process of testing software is subdivided in the following four sub phases.

5.1.1 Modeling the software environment

For testing different interfaces could be used. The interfaces are Human interface, Software interfaces (API), File system interaction and Communication interface. The interfaces used in this test are the human interface and software interface (API).

In figure 20 is an example of the used user interface of the C2R prototype. To add a transcription of the medical consultation is imported by using the transcription button. In the prototype has did be done two times because in an earlier version the patient text and family doctor text where entered separately. The current prototype the transcript of the complete conversation can be selected two times.



The screenshot displays the C2R prototype's user interface. On the left, a sidebar lists users: Maria Nivel and John Smith. The main area shows a patient record for John Smith, with fields for First Name, Last Name, Date of Birth (19-6-1989), and Gender (Male). Below this, there are buttons for 'Generate Report' and 'Edit'. A 'Problem:' field is present, followed by a table with four sections: Subjective, Objective, Evaluation, and Plan. The Subjective section contains the text 'Patient suffers from itching ear.' The Objective section contains 'Ear canal is swollen. Ear shows signs of infection.' The Evaluation section contains 'Patient has otitis externa.' The Plan section contains 'Ear drops and paracetamol.' There are also buttons for 'Doctor Transcription' and 'Patient Transcription'.

Figure 20 Example of a generated report made by the C2R prototype (reprinted from (Maas, et al., 2019)).

The software interface (API) are tested in the backend of the system. The API are tested while running the transcript through the many micro analyzers. The results of the micro analyzers will be logged and stored. Execution the test and analyzing of the (intermediate) results will be done in phase 3.

5.1.2 Selecting test scenarios

All the code of the software has to be tested. There is not one test that test the complete software, therefore different scenarios are be used. The input set of transcriptions exists of three real medical consultation recordings, eight real consultation transcriptions from Nivel and five simulated medical consultation recordings.

Path execution testing

The line of code is tested because the prototype runs the code in sequence. The lines are tested with real consultation scenarios. The scenarios ask for the same execution of the code lines as a real scenario. What is important is the execution path of the code. For the path execution is a log file extracted that will define the way how the micro analyzers are executed. This will be visualized in the graph figure 21.

Input domain testing

The input domain test will cover the input possibilities of the interface. The buttons and text fields are testing by starting a new consultation with a new SOAP. The interface with the buttons and text fields are shown in figure 20.

Future testing:

It is possible to test the interface but not possible to start testing the backend. Because in the code testing is done within the code itself. When the test scenario can be started from frontend it will be give more results. Also, for now is it possible to test an audio file and a transcript. The modalities video and Bluetooth measurements cannot be used.

5.1.3 Running and evaluation the test scenarios

In this phase are the test scenarios run through the C2R prototype. From the test will be a log be obtained so all the intermediate results can be analyzed. The results of the real consultation produced by the prototype are presented and analyzed.

Recording 1 real situation:

Transcript: The transcript was manually created so there is no transcript created by the C2R prototype to compare.

Translation: The C2R output is analyzed by comparing the translation with the original text. The actual comparison is shown in attachment 7.

The comparison is very good. There were only problems with the “slang” word “daalijk” it is an abbreviation of “zo dadelijk” with means in a moment. The C2R prototype leaves that word out and did not try to make something weird from it. Did didn’t happen with the word “klaren”. It is when divers pinch their nose and blow on it. This word after translation was “cleaning”.

Below the transcription that same sentences are shown but all the you and your are replaced by the patient and doctor. Some examples below:

Table 8 Placing patient or doctor in the transcript

Original translation	Replaced by patient or doctor
How are you?	How's the patient?
didn't know that about you	the doctor didn't know that about you
I can also see that you have had problems with your ear in the past.	The doctor can also see that the patient has had problems with the patient's ear in the past.
I can also give you a nasal spray.	The doctor can also give the patient a nasal spray

Unmatched triples: The triples are created by the micro analyzers Frog, Fred and Ollie optimiser. Where Fred and Ollie use the English translation where Frog use the Dutch transcription.

There where 647 triples generated by the three micro analyzers. The most triples are created by Fred (446), after that comes Ollie (137) and last Frog (67). The most used route is marked green with in figure 21. Examples of triples from the three micro analyzers:

Table 9 Examples of the triples created by the microanalyzers

Micro analyzer	Found triple
Fred	"triple": { "object": "Completely", "predicate": "hasQuality", "subject": "Clear" }
Ollie optimiser	"triple": { "object": "the patient", "predicate": "be has by", "subject": "poor hearing" }
Frog	"triple": { "object": "slecht oké", "predicate": "horen", "subject": "patiënt" }

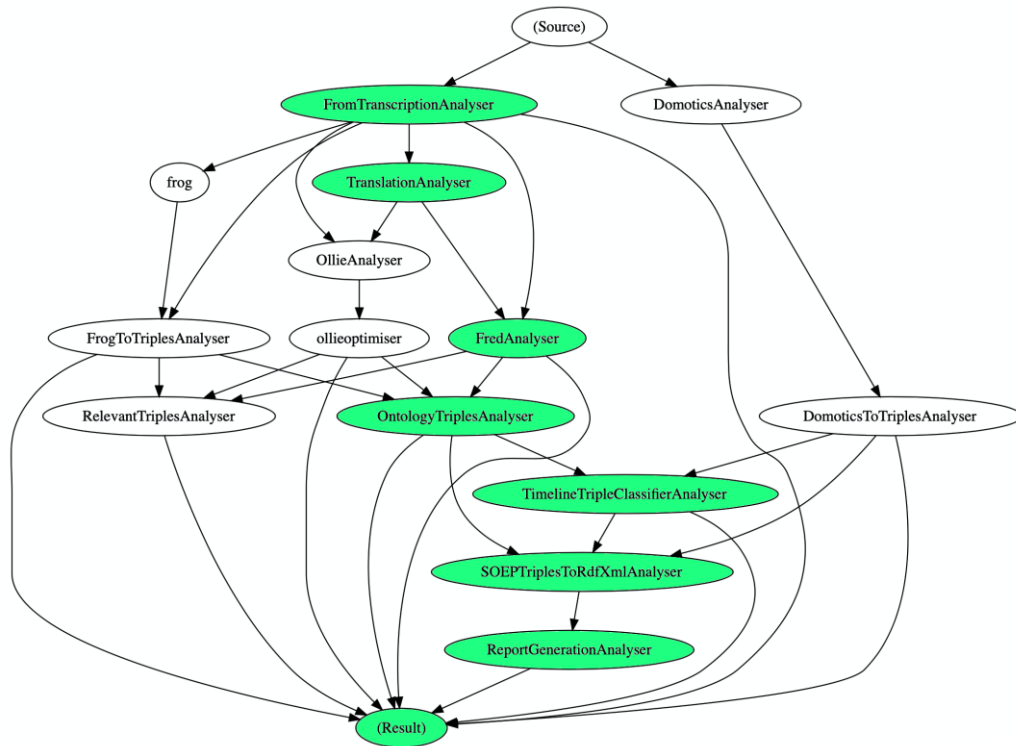


Figure 21 Route of the most used microanalyzers within C2R

Matched triples shown in a chain. clog < subClassOf > event < subClassOf > know < hasQuality > well < hasQuality > eardrum

Generated report / EMR (SOAP): Below are the four SOAP fields that are found by the C2R prototype.

```
Soep S "soepfield": "" }
Soep O "soepfield": "Verstopping zichtbaar aan trommelvlies."}
Soep E "soepfield": "" }
Soep P "soepfield": "" }
```

The actual SOAP report is a very poor result. In attachment 4 is an overview of the complete consultation. The missing words are validated within chapter 6 Validation. The validation will explain the reason why some of the words are not included by the C2R prototype.

Recording 2 real situation:

Transcript: The transcript was manually created so there is no transcript created by the C2R prototype to compare.

Translation: the translation is just like recording 1 good. there are sentences that need to be looked at.

Table 10 Placing patient or doctor in the transcript

Original transcript	Translation from the log
nog steeds wel last van	yes in itself not bothered
ja op zich nergens last van	yes in itself not bothered
dat is goed. Dan wacht ik het rustig af	(Sentences is missing)
vertel wat is de reden dat vandaag afspraak hebt gemaakt	tell us what is the reason that you made an appointment today

The first and second sentences are having both a different meaning but have the same translation.

The first sentence is an answer, if some of the motioned symptoms are decreasing or are still a bother. The second sentence is an answer on the health situation before the all the symptoms started is the health situation was okay.

After testing the translation separately with Google, the phenomenon occurs if you have none punctuation marks, the translation will be different. The punctuation mark comma is the most important in these sentences.

Table 11 Examples of the triples created by the microanalyzers

Original transcript	Translation from the log
Nog, steeds wel last van	Still, it does bother you
Ja, op zich nergens last van	Yes, not bothered by anything
Vertel, wat is de reden dat vandaag afspraak hebt gemaakt	Tell me what is the reason that you made an appointment today

Unmatched triples: The triples are created by the micro analyzers Frog, Fred and Ollie optimiser. Where Fred and Ollie use the English translation where Frog use the Dutch transcription.

There where 954 triples generated by the three micro analyzers. The most triples are created by Fred (446), after that comes Ollie (166) and last Frog (62). The most important/most used route is marked green with in figure 21.

Matched triples via chain: the unmatched triples are compared to the ontology by using the Ontology Triples Analyser. The triples that are matched and are connected by chain are shown in attachment 8.

The found triples are presented as an ontology. An unusual word that is used is 3thing5. The reason that this is used is because the C2R prototype knows there is an “thing” and knows his relations but not how to define the “thing” himself. Therefore, the object is given the name 3thing5. The triples are used to generate the report.

Generated report / EMR (SOAP): The generated report is more advanced than the first situation.

Soep S "soepfield": " *Patiënt heeft last van koorts aan trommelvlies* " }
Soep O "soepfield": "*Ontsteking zichtbaar aan oor*: }
Soep E "soepfield": " *Ontsteking doet middenoorontsteking vermoeden.*" }
Soep P "soepfield": " *Ibuprofen, neusspray en paracetamol.*" }

Recordings simulated situation:

In this section will be the results of the five simulated recordings discussed. The simulated recordings are also used to gain more insight in the C2R prototype. For the simulated recordings are transcripts manually created. This is having the consequence that the performance of the video/ speech to text can't be analyzed. In the first attempt to test the C2R prototype with the simulated transcript was the abbreviation PA swapped with PT. The consequence was that the sentences from the patient in the transcript where not translated and left out in further processing. Below the results per simulated consultation.

Simulated situation 1 see file here: The system separates the transcription from the care provider and the patient. When they are separate, they are translated from Dutch to English. The translation from the care provider and patient are completed and well performed. The created triples: there are 96 triples created there are 11 from Fred, 55 from Ollie and 30 from Frog. There are no triples matched and report generated. See attachment 9

Simulated situation 2: Because of an error within Fred the test has failed multiple times.

Simulated situation 3: Just like simulated situation 1 is also the transcript of both the general practitioner and the patient created. The transcript is translated in English and has been translated fairly well. The created triples: there are 57 triples created there are 2 from Fred, 30 from Ollie and 25 from Frog. There are no matched triples found and no generated report. See attachment 9

Simulated situation 4: Also, this simulated situation 4 started with the translation of the transcript, this was successful executed. The created triples: there are 602 triples created there are from 460 Fred, 98 from Ollie and 44 from Frog. Also here are not matched triples created or a generated report.

Simulated situation 5: The created triples: there are 1035 triples created there are from 845 Fred, 136 from Ollie and 54 from Frog.

The matched triples that are found:

*fever < associatedWith > feverbitincrease < hasQuality > topic < hasQuality > ear
pain < associatedWith > Ithing3 < associatedWith > ear
eardrop < associatedWith > mention < Agent > patient
nasalspray < subclassOf > spray < Agent > patient
paracetamol < associatedWith > have < Agent > patient*

The matched triples are used to create the generated report.

S "Patiënt heeft last van koorts aan oor."
O “ ”
E “ ”
P "Neusspray en paracetamol."

Because the nasal spray and paracetamol are associated with the patient is it not part of the Plan but of the Subjective.

5.1.4 Create a ground truth

The ground truth is the golden standard that is held up against the output of the C2R prototype. For this research there has been looked into multiple options for the creation of the ground truth. Because the simulated recordings did not have a produced EMR, therefore was tried to create an EMR by hand. Because it should be done by a non-medical professional it wouldn't be valid for validation. The golden standard that is used for the validation are the EMR's created by the general practitioner. Also, for the Nivel transcriptions is afterwards by a general practitioner and EMR developed.

5.2 Comparison of the C2R and the real consultation SOAP

The important words and sentences are now marked within the two real recordings. The important sentences are being market in the transcript analyses with the translation. For the three real situation the family practitioner has made an EMR during the consultation. The created EMR will be used as the golden standard. The SOAP will be marked per sentence in the transcript based on the golden standard. The aim is to see if the SOAP output created the C2R is comparable with the results of the family doctor. The best situation is when the SOAP created an output that is the same/similar to that of the family practitioner. Below a comparison of the family practitioners EMR and C2R SOAP output.

Comparison 1

Output EMR Real consultation 1 by general practitioner

```
12.09.2019   C
S   na duik minder goed horen in linker oor, was met een pijnsteek, associeert met een perforatie die hij
    ooit had, ook bij zwemmen toen
O   oren: links en rechts wat littekentjes, rechts normaal TV
    links : ingetrokken, geen perforatie te zien
E   tubastenose
    H73.00 (Tubair catarre/tubastenose)
P   neusspray
    12.09.2019-13.09.2019   FLUTICASON-PROPIONAAT NEUSSPRAY 50UG/DO FL 150DO
    1       2D11 IBN
```

Figure 22 EMR created by the general practitioner consultation 1

Output EMR Real consultation 1 by C2R prototype

```
Soep S "soepfield": "" }
Soep O "soepfield": "Verstopping zichtbaar aan trommelvlies.}"
Soep E "soepfield": "" }
Soep P "soepfield": "" }
```

Comparison

In the C2R output are is the words blockage (verstopping), visible (zichtbaar) and eardrum (trommelviels) used. In the transcript is the is the word clogged (verstopt) is used. Clogged is used in the example of divers that have the same clogged ears and for the description for the treatment to blow through the nose when closed. When there is a connection between the nose and nasal cavity because that is clogged. So, the sentence “clogged eardrum visible” is not what is say in the conversation.

The S, E and P from SOAP are not found at all. Those where however available in the transcript.

Comparison 2

Output EMR Real consultation 2 by C2R prototype

12.09.2019 C
S sinds 5 dagen niet lekker wat verkouden hoesten, na 2 dagen koorts er bij 38,5, en sinds gisteren minder hoesten en meer oorpijn. in het rechter oor heeft al pijnstillers gebruikt en neusspray
O oren: links gb, rechts rood TV++
E OMA rechts
H71.00 (Otitis media acuta/myringitis)
P uitleg over beloop en zelfhulpmiddelen, indien langer bestaand ioverleg iover evt toevoegen AB

Figure 23 EMR created by the general practitioner consultation 2

Output EMR Real consultation 2 by C2R prototype

Soep S "soepfield": " *Patiënt heeft last van koorts aan trommelvlies* " }
Soep O "soepfield": " *Ontsteking zichtbaar aan oor:* }
Soep E "soepfield": " *Ontsteking doet middenoorontsteking vermoeden.* " }
Soep P "soepfield": " *Ibuprofen, neusspray en paracetamol.* " }

Comparison

In Subjective the patient has told that that he suffers from a fever and have pain om is ear so that is correct. The C2R prototype output has missed: having a cold, coughing, pain right ear, 5 days of not feeling well and stuffy nose.

In Objective has found the conclusion of the redness of the eardrum. Inflammation visible on the ear.

In Evaluation of the C2R prototype EMR is the sentence not grammatically correct but the meaning with the general practitioner is correct.

In Plan of the EMR output of the C2R prototype are the treatments correct. What is missing is the antibiotic treatment when the situation not improve.

5.2.1 Explanation of the results

The output of the C2R system comes from multiple variables such as microanalyzers, algorithms and the ontology. For now, it is not possible to unit test the different microanalyzers and algorithms separately. The part that is focused on is the ontology.

In the figure 24 in this paragraph is the ontology presented, the ontology is needed to match the triples. The ontology contains the medical conditions, symptoms and human anatomy. For the knowledge graph is the human anatomy used as starting point. The ontology that is developed for the prototype is Protégé used and modeled as OWL ontology. To extend the current ontology there is SNOMED added as an microanalyzer. SNOMED contains synonyms of global medical and health terms.

The ontology is here analyzed in an attempt to explain why not all words are found in the matched triples. If medical-, anatomical- or treatments terms are motioned in the transcript and they are not incorporate in the ontology, it is not possible for C2R to find them. Also, the other way around, if the terms are incorporate in the ontology the C2R is able to match the triples.

Explanation of the real consultation results

Consultation 1: for this consultation there are 650 found triples analyzed to find words that are either medical, anatomical or a treatment. The found words are placed in the first column of the table in attachment 10. In the second column is verified if the words (or synonym) is available in the ontology. In the third column is indicated that the matched triple / EMR actually found and used the term(s). The fourth and last column present which microanalyzer have found the triple.

The table in attachment 10 have marked the rows with either orange, green or white. When the row is orange this implies that the medical, anatomical- or treatment term is included in the ontology, but the algorithm was not able to match this, green means that term is available in the ontology and successfully found, white means not available in the ontology and could be improved.

Findings:

1. The actions itself that are executed by the general practitioner such as; must look into good ear or the doctor can also see that the patient has had problems with the patient 's ear in the past could not be found by the ontology (see figure 24). The ontology does not say anything about the examination.

2. Also, explanation of a treatment in the transcript such: *as spray a few times a day, then there comes a day whether it is over, that is with 2 days or 4 are* also not found. It is a clear treatment description. It is found in the transcription and recognized by Ollie as a triple. But the algorithm was not able to match this.

3. What is found by the matched triples is clogged and eardrum and that is made into the sentence: Verstopping zichtbaar aan trommelvlies.

"Clog"->"Ear"[label="Patient"]

"they"->"those clogged ears"[label="Then often have"]

"they"->"clogged ears"[label="have"]

The matched triple:

clog < subClassOf > event < subClassOf > know < hasQuality > well < hasQuality > eardrum

4. When the triples are found and available in the ontology, the lack of matching depends on the Algorithm.

Consultation 2: For the second consultation there are 1185 triples analyzed to find the words that are either medical, anatomical or a treatment and placed in the table of consultation 2 in attachment 10. The words are placed in the same manner as in the first consultation.

In attachment 10 the table shows the rows are marked with either orange, green or white. When the row is orange this implies that the medical, anatomical- or treatment term is included in the ontology, but the algorithm was not able to match this, green means that term is available in the ontology and successfully found, white means not available in the ontology and could be improved.

Findings:

1. The correct matched triples is more accurate than in consultation 1. Of the 57 words that are either a medical term, anatomical term or a treatment is 20 words correct found and matched. 7 could be found by the C2R but didn't and 30.
2. What is missing is that within transcription is the instrument that is used and what is measured with the instrument. because of the lack, this is not tried to match during the execution of the C2R system.
3. When the triples are found and available in the ontology, the lack of matching depends on the Algorithm.
4. From the 1185 triples are only 57 found with medical, anatomically or treatment terms in them. When they are found some of them are the same used in another triple. The most are not relevant, see chapter 5.3.1 about the sentences that are marked irrelevant.

Analysis of the results from attachment 10

In attachment 10 the table state the medical, anatomical or treatment terms that occur in the created triples from real-world consultation 1 and 2. All the terms are listed if they are occurring in the ontology, if it is (or not) found and placed in the EMR and found by with microanalyzer. All the results are sorted in anatomical, symptoms, functions, evaluation and treatment.

The terms are marked with green when the terms appear in the ontology and is found by the C2R system, orange when the term appears in the ontology but the C2R system was not able to match them and white when it does not appear in the ontology and matched triples.

To give more insight in the analysis, the colors are divided in Type I and type II errors.

Green = true / positive

Orange = true / negative

White = false / positive

Terms:	Anatomical	Symptom	Functions	Evaluation	Treatment
True / positive	7	4	0	4	4
True / negative	5	7	0	3	4
False / positive	13	18	4	10	3

A quick overview gives the knowledge that the most not found terms are within symptom. The Treatment have the most percentage true / positive of 36,36%. Functions does not have a lot of terms found within the triples, and those that are in a triple could not be matched.

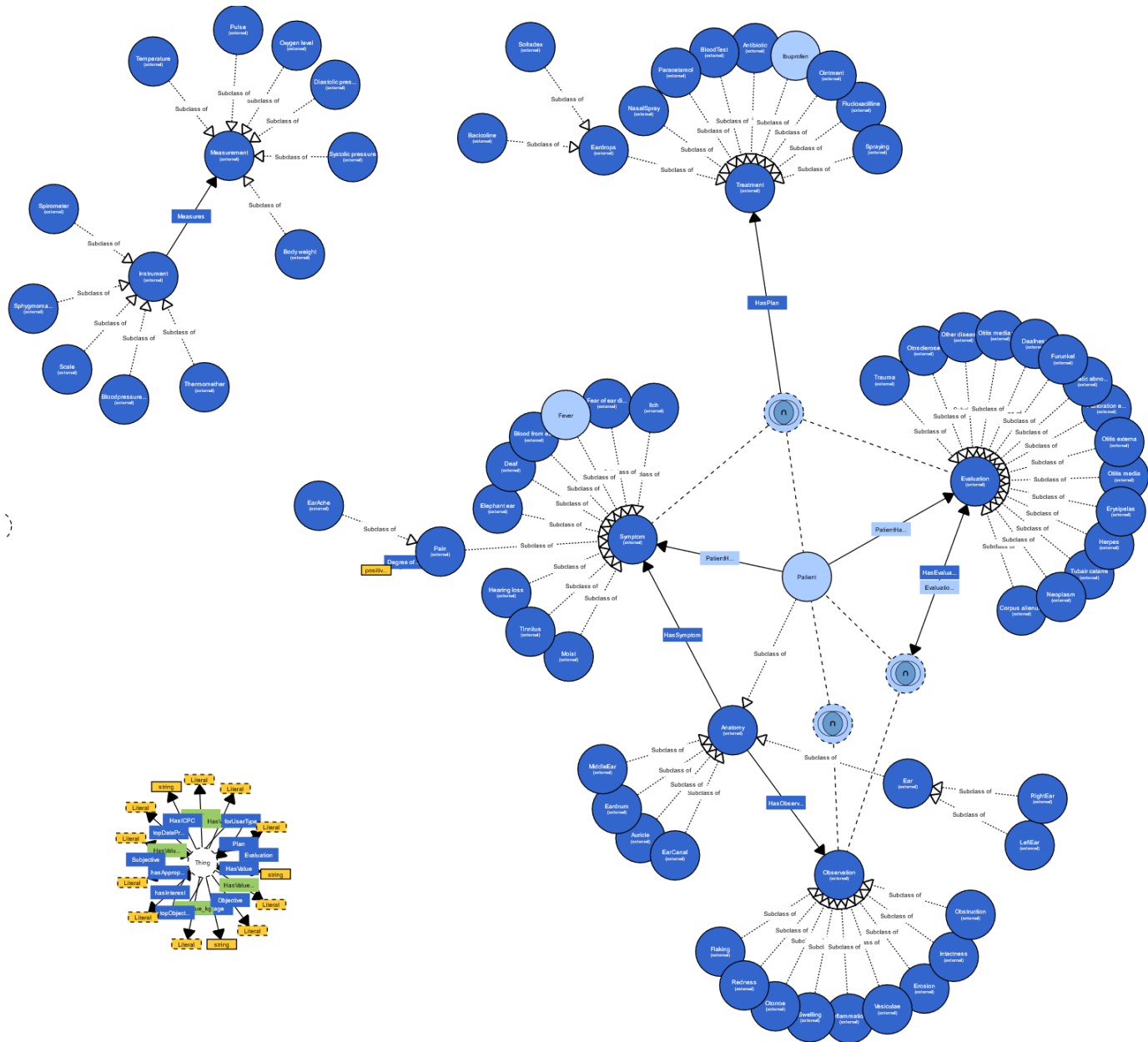


Figure 24 Ontology as used in the current prototype

5.3 Analyzing the sentences and structure used for testing

In this section the sentence from the transcript that is used to generate a blacklist so the prototype can be optimized.

5.3.1 Irrelevant sentences and word blacklist

For the defining of relevant sentences within a medical conversation, more specific the real recorded medical consultation is the created EMR of the general practitioner used. The sentences that contain (medical) words or sentences that corresponds with the EMR that say something about the patient condition or treatment is highlighted in the transcript. All remaining sentences are named as irrelevant sentences, because they don't contribute to the generation of the EMR.

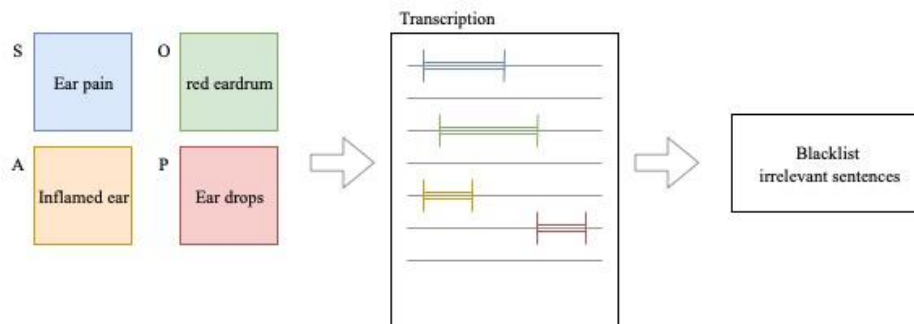


Figure 25 Process for the creation of the blacklist

Not all words and sentences that are not marked must put on the blacklist. The blacklist is made so less useful words can be filtered out of a transcript, so the system does those not take these words into account by creating triples. The aim is to create more accurate triples and later in the process accurate matched triples. The three real consultations are used because they have an EMR created by the general practitioner. Those EMR output can be related back to the corresponding sentence. To give more inside in what sentences are important, they are marked with the RIAS code (see chapter 5.3.2.)

Explanation of the findings of irrelevant words and sentences. In all three of the transcripts is noticeable that greetings, handling explanation during the medical examination and closing of the conversation are not marked as relevant. Because only three transcripts could be used for this, is it recommended to research this more in future research.

The blacklist

Some of the words and sentences are that are noticed not useful to a contribution in the EMR. These words are put in the blacklist below. An overview of the list is shown in attachment 11.

Phrases that are often used and announce an action

Besides finding sentences that are less meaningful, are there some sentences that could be helpful for the C2R program to be. The sentences that are found are used in multiple transcript and are an announcement that the doctor is starting physical examination for ear problems. This could help the next step, action recognition in the C2R prototype.

- ik moet in ieder geval in je oor gaan kijken om even te kunnen zien wat er aan de hand is en dan uuh kunnen we kijken wat er nodig is.
- ik pak even mijn lampje en dan ga ik even kijken
- Ik ga als eerst in je goede oor kijken.

5.3.2. All the important sentences marked in the transcript with RIAS

In chapter 4.3.5 *Finding important words and sentence structure* are the important RIAS codes presented that are marked as important. The RIAS codes they think is important are:

Gmed	Gives information - medical condition
Para	Paraphrase/checks for understanding
Qmed	Asks questions - medical condition
Gpsy	Gives information - Psychosocial
Gsoc	Counsels - Social
Cmed	Counsels - Medical condition/Therapeutic regimen
Reas	Reassures, encourages or shows optimism

The sentences that are highlighted for the blacklist are now also labeled with the RIAS code. The intention is track down which RIAS code is important for the creation of the EMR. This gives more insight in the sentences that are important for creation of an EMR in the C2R system. See appendix 7 for the transcripts + RIAS codes.

Subjective

There are three codes of RIAS important in the subjective part of the medical conversation.

- Asks questions - medical condition (Asks Closed-ended Questions-Medical Condition)
- Gives information - medical condition
- Shows agreement or understanding/ Paraphrase/checks for understanding

Objective

Objective does not have related RIAS codes according to figure 16 *Comparison between SOAP and Rias*. But because the general practitioner does explain the interpretation of the complaints and symptoms is it labeled as with gives medical information.

Assessment

Gives information – Medical condition

Plan

Gives orientation, instructions

Counsels – Medical condition/Therapeutic regimen

Counsels – Lifestyle/Psychosocial

Conclusion: By using the RIAS codes to label the marked the relevant sentences is the RIAS code that is marked important by the general practitioner validated. It is for the C2R program interesting to know what sentences are important so in a future C2R system is could be possible to eliminate the irrelevant RIAS code with their corresponding sentences. In figure 16 by van der Kooi & Lim (2019) shows that the literature does not link no RIAS activity are recoded during the objective. Within the analysis above shows that the general practitioner does speak during the observation.

Just like the general practitioners are in the gives information - medical condition, paraphrase/checks for understanding, ask questions - medical condition and Counsels - medical condition/Therapeutic regimen important and have a frequent occurrence. The RIAS codes: gives information – psychosocial, Counsels – Social and Reassures, encourages or shows optimism are less or not important in the sentences that are needed for the EMR creation.

5.4 Functional design data extraction

Testing is important to determine the performance of a product. The C2R system is tested by using the real scenarios and to compare the SOAP output to those of a general practitioner or the missing words.

There are two important streams of data from C2R to Yoda and vice versa. The transferred test data from Yoda to C2R are the recordings (videos and audio) or transcript. Yoda stores all the test data, among other things. Within Yoda a file is made specialty for test data that needs automatically be transfused to the C2R system. The tester can prepare the file so all the needed data could be used by the C2R system.

When the test is done the C2R system gives (intermediate) results back. These results could be the SOAP output or the logs from different third party microservices. All the results and output are stored in Yoda for further analysis. In figure 28 an overview of the flow of the data within the testing environment.

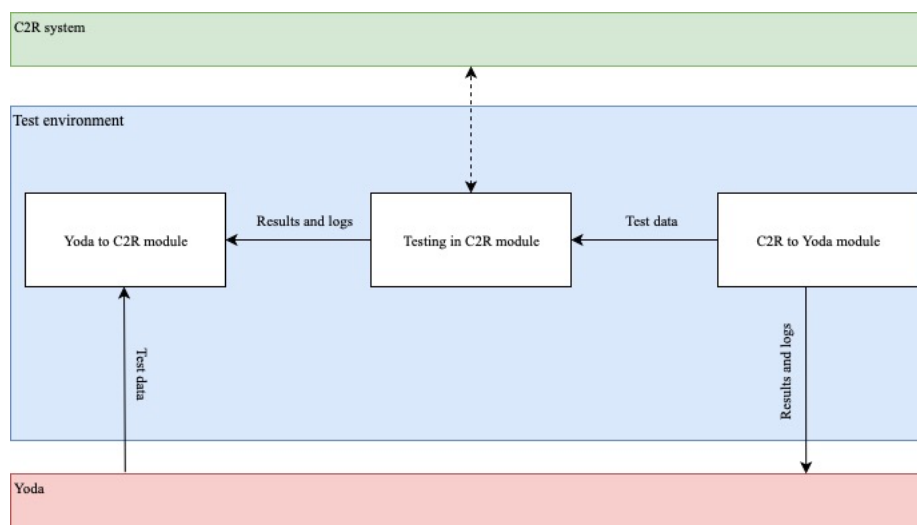


Figure 27 Birds eye view of the functional technical architecture

Epic and User stories for the C# module to improve C2R testing environment.

For the development of the new data entry and gathering are Epic stories and User stories created. They help checking if the to develop solutions meet all requirements. Epic story is a more generic User story. The epic stories and user stories below only apply for the new to realize the solution described in the functional design improvement.

Front end

Epic Story – As a tester I want to add metadata to the testing output, so that the output data can be retrieved more easily.

US: As a tester I want that metadata is been added automatically to the testing output, so that no output is stored undefined.

US: As a tester I want that metadata is been added automatically to the testing output, so that no mix-ups can be made.

Microanalyzers

Epic story – As a tester I want to analyze the output of every analyzer separately, so that the performance and contribution per microanalyzer can be determined.

US: As a tester, I want to test every microanalyzer on their performance, so that it can be used of analysis.

US: As a tester, I want to run every microanalyzer on their own, so that the accessibility could be tested.

US: As a tester, I want to run every microanalyzer separate, so that malfunction can be detected.

C2R transfer to Yoda

Epic story – As a tester, I want to transfer output data from the C2R test environment to Yoda, so that the output can be stored for further analysis.

US: As a tester, I want to transfer the speech transcription in a separate text file to Yoda, so that it can be stored.

US: As a tester, I want to transfer the speech transcription in a separate text file to Yoda, so that it can be used in an analysis.

US: As a tester, I want to transfer the extracted triples from FROG in a separate text file to Yoda, so that it can be stored.

US: As a tester, I want to transfer the extracted triples from FROG in a separate text file to Yoda, so that it can be used in an analysis.

Yoda transfer to C2R

Epic story – As a tester, I want to transfer the selected test data from Yoda to the C2R test environment, so that the C2R system can be tested.

US: As a tester, I want to transfer transcripts from Yoda to C2R, so that the system can be tested.

US: As a tester, I want to transfer a video file from Yoda to C2R, so that the system can be tested.

US: As a tester, I want to transfer audio file from Yoda to C2R, so that the system can be tested.

Report generator

Epic story – As a tester, I want to store the generated report with the other outputs, so that they can analyzed and stored together.

US: As a tester, I want to store the generated report automatically in the test file on Yoda, so that the output is combined.

Functional design improvement of current situation

In the current prototype it is possible to run the multimodal data or a transcript through the system. The outcome varies considerably without clear indication. To get more insight in the results and the processing of the C2R system is it needed to gain more insight information. To get this information the (intermediate) results and logs during the execution of a test with test data is a beginning. In figure 29 is an overview of the collection of the different data that should be gathered. The specifications in figure 29 gives more inside in the type of data, and the kind of data there will be transferred and gathered form the C2R and Yoda for testing.

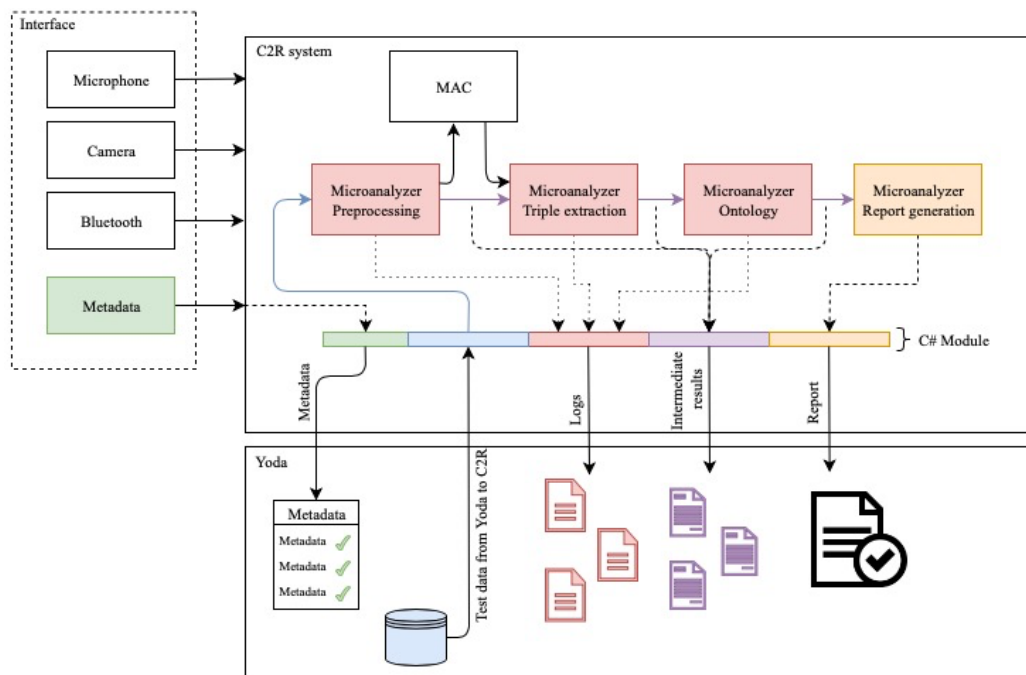


Figure 28 Functional design for extracting of data, results and logs

Here a short explanation of the colors used in figure 29 and the situation to be developed. The colors in the figure are used for the different solutions that are going to be developed to improve the testing and analysis of the output. Green is for the metadata that could be added in the front end and is attached to the output for the storage of the results. Blue is for the test data that is selected and put in the file for testing in Yoda. The test data will automatically be uploaded to the C2R system without interference of the tester. Red is for the log outputs in a .text file of an individual microanalyzer. The output must be stored in the selected test file in Yoda. This gives insight in the performance and participation of the microanalyzers. Purple is for all the intermediate results that is produced as output by an microanalyzer. This gives insight in the results per microanalyzer. Orange visualize the output of the C2R system when all the intermediate results are combined to create a report. Figure 30 a more extensive explanation of all the solutions for the improvement of the C2R system (in attachment 12 a larger overview).

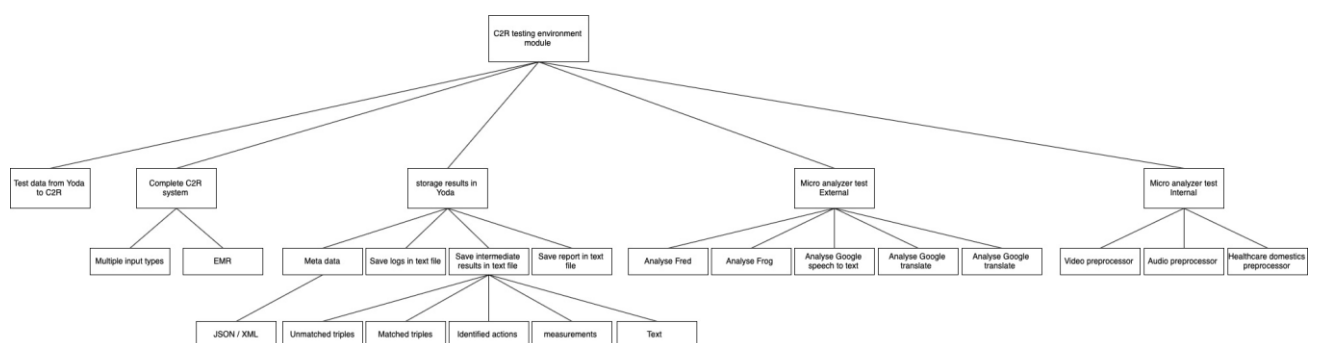


Figure 29 Overview of all the data that must be gathered in the testing environment

Below a more extensive description of the meaning of the colors within figure 29.

Metadata (green in figure 29)

In figure 20 is the front end of the C2R system displayed. It shows some field where metadata could be added like first and last name, data of birth and gender.

In the future version this must be extended to more options and possibilities to added other metadata especially for testing. This will help to the automatization of testing. When the metadata is added in the front end, could it be used for het metadata of the storage of the (intermediate) results and logs of an executed test.

In chapter 4.4.1 Metadata structure is the metadata structure discussed as provided by Yoda. The main head pillars are descriptive, administrative and rights. When the metadata form is filled in, it can easily be taken over for by other files or documents. The attached JSON file can be copied and be reattached to another file or document. This is convenient for more general information about the rights for documents or administrative for documentation.

It is a must to be able to add test specific metadata through the front end. This is more focused on the descriptive fields. The fields that are eligible are; Title, Description, version, collection process, period (dates) and related data packages.

Test data from Yoda to C2R (blue in figure 29)

The multimodal test data such as recorded video's, audio files, transcripts and measurements are stored in Yoda. The complete set of test data must be entered in the C2R system. A save way to do this is to upload the test data automatically to the system. In the current situation is it done using USB sticks and mail to transfer data, files and output. By using Yoda, the transfer of test data would be easier, saver and no data is getting lost.

Microanalyzers participation (red in figure 29)

It is important to store the logs that are produced by every microanalyzer individually. It gives more inside in the microanalyzers that participate to the outcome. It will be possible to visualize the road that has been taken through all the microanalyzers just like figure 17.

Intermediate results from the microanalyzers (purple in figure 29)

The intermediate results will be extracted from the logs that are kept. The results must be filled from a large continuous .text file. The intermediate results will be stored separately. The reason for the storage of these results is to measure the performance per microanalyzer. It also offers the possibility to trace back from the final report back to the triples and match triples.

Final report (orange in figure 29)

With the rest of the results is the final generated report stored in the testing file on Yoda. The final report is needed for the calculation of the accuracy of the C2R prototype. In collaboration with other researchers from the C2R program will the current C2R prototype being improved. The five ways of data retrieval and making data available are developed for the current C2R prototype. Figure 30 gives a more detailed view on the C2R testing environment module presented in figure 29.

The IRODS solution made by C2R

Yoda is built on an opensource IRODS system. To make an easy equation it looks like “Google Drive on steroids” to be able to transfer and manage the data from C2R system to Yoda (and other way around) is iCommands (command for Linux) or webDav (file interface) needed.

The solution that will be built by three research also participating in the C2R program is to make it possible to transfer data from C2R testing environment to Yoda. The solution that they are currently developing is a C# module for IRODS. The module makes it easier to communicate with IRODS for the sharing of data. The module has to features to make and delete folders and share and store data and metadata. The solution is an IRODS implementation as a C# module. There will be a translation of iCommands commands been made for direct access of the C# code. When there is access with the code it could be possible to place documents on the IRODS server (Yoda).

When is the C# module useful for C2R?

The C# module is useful when the create logs can be gathered from and replaced to Yoda (see figure 29). Figure 31 is a simplification of the Functional design. It shows where the C# module will be placed in the C2R solution. It will be a stand-alone module between C2R prototype and Yoda. In figure 29 is the rectangle divided in the colors green, blue, red, purple and yellow a representation of the data streams trough the C# module.

The C# module will be helping in collecting the data and separating the results and logs. The module must offer a solution to separate the continues log text file in individual files. Every file is stored automatically separately on Yoda.

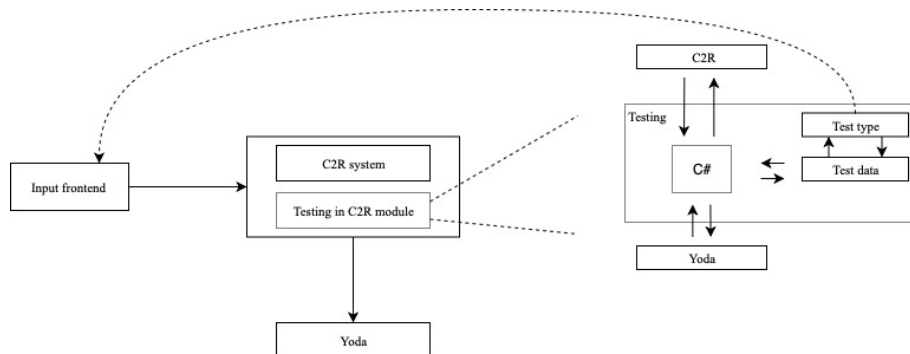


Figure 30 The position of the C# module in the testing environment

Functional design & simplification

The functional design of the C# module with integration to C2R is not been realized yet. Because fellow C2R researchers have a delay of the project and reformulate of the scope. The functional design is created, and the epic-/user stories are defined for a team that will continue and implement the project.

5.5 Storage of training, testing and validation data for action recognition

The C2R program offers research projects for many different researchers from different backgrounds. All the different projects contribute to the C2R program and create different data (types).

One of the research projects is action recognition from medical actions. The actions can be recognized after a large dataset of medical recordings that is created is used for training and learning of the action recognition neural network.

The recordings follow the medical guidelines to make sure the relevant, most occurring actions are presented in the dataset. A medical guideline can consist of one or multiple medical actions. Within the master thesis is looked into what medical actions are the most common by a general practitioner and selected for the creation of the dataset. The recordings within the dataset that is created consist of the following medical actions:

1. Blood pressure measurement
2. Palpation abdomen
3. Heart rate measurement
4. Percussion abdomen
5. Auscultation lungs
6. Auscultation heart
7. Auscultation abdomen

From all the actions are multiple recordings in different combinations. The recordings are with the other research data stored on Yoda. In the file Laura there are three main folders – Annotation – TXT files – Video's.

Annotation of one recording are different medical actions and the same medical actions in sequel. Therefore, the recordings are cut into separate sessions. There is a total of 193 unique session and 453 videos, there are multiple videos recorded with different camera viewpoints of one unique session. After the video has be separated are those annotated using ELAN an online tool for annotating videos.

The aim is to recognize the medical actions. The medical actions form one of the seven medical actions (see above) that is executed are labeled so the neural network could learn the different actions. Beside the actions is also the patient's position (sitting upright, laying down with flat legs or with the knees bent) labeled. The position of the patient tells the network about the medical action. The third annotation is the specific area of the patient's body annotated. The different parts are chest, upper back, abdomen, or arm. Finally, is the distance between the general practitioner and the treated patient annotated.

On Yoda there are two separate folders for the annotation named annotation and TXT files. The folder annotation contains .eaf and .pfsx files. This type of files is readable for the online tool ELAN for the creation and adjustment of the annotation. When it is needed to change the current annotation, these files can be used in combination with the ELAN tool. The TXT folder contains 193 .txt files. Each file has a relation with a unique session. In a file there are three important areas; posture patient, distance to patient and area of investigation. The annotation in the .txt file are human readable and contain the annotation used in the action recognition research.

Video's contains all the recordings that is used for the training, learning and annotating. The recordings are as mentioned before recorded with different camera. This is why a unique recording can be stored with a letter. An example is: abdomen_auscultatie_05a. This recording stores a combination of the medical action auscultation (pressure) on the body part abdomen. It is the fifth recording with camera a. the characters b and c stand for another type of camera. In this case go pro or iPad. All the recordings in the file video are stored in .mp4 format.

Chapter 6 Validation of the generated report

The validation give insight in the quality of the generated reports by the C2R system to compare the output with the golden standard. The validation is measured by the precision, recall and the F1 calculation. The items that are used in the calculations; True positives (TPs) found in the output and the golden standard, False positives (FPs) found in the output but not in de golden standard and False negatives (FNs) found in golden standard not found in the output.

Table 12 presents an overview of the eight real-world consultation form Nivel and the last two columns are the real-world consultation recorded in this research (see chapter 5.2). Because of the restrictions from Nivel is the created EMR not presented in this research. Every consultation (C-x) is divided in the generated report (R-x) and the golden standard (S-x). Every selection of items is placed in the section of the SOEP format in table 12 For the generated reports the number of in TPs/FPs/FNs are also shown, respectively.

Table 12 Number of items included for each section of the SOEP format, with TPs/FPs/FNs for the generated reports.

	C-1		C-2		C-3		C-4		C-5		C-6		C-7		C-8		C-9		C-10	
	R-1	S-1	R-2	S-2	R-3	S-3	R-4	S-4	R-5	S-5	R-6	S-6	R-7	S-7	R-8	S-8	R-9	S-9	R-10	S-10
S	1/0/3	4	1/0/8	9	1/0/4	5	0/0/4	4	0/0/4	4	0/0/9	9	0/0/8	8	0/0/8	8	0/0/3	3	1/0/4	5
O	0/0/2	2	0/1/6	6	2/0/1	3	1/1/8	9	0/0/4	4	0/0/4	4	1/0/2	3	1/0/1	2	1/0/3	4	1/0/1	2
E	0/0/1	1	0/0/2	2	1/0/0	1	1/0/1	2	0/0/2	2	0/0/1	1	0/0/1	1	0/0/1	1	0/0/1	1	1/0/0	1
P	0/2/1	1	3/15	8	1/1/1	2	1/0/2	3	1/0/5	6	0/2/5	5	0/1/2	2	2/0/0	2	0/0/1	1	1/0/0	2

Here a short explanation about the calculation of the quality using Precision, Recall and False positives. The number TPs/FPs/FNs and the golden standard are used for the calculations.

Precision is the number of items that occur in the generated report that is considered relevant according to the golden standard. It can be expressed as measure how relevant is the created output.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positives} + \text{False positives}}$$

Recall is the number of relevant items that are included in the generated report divided by the number of items that are in the golden standard. It can be expressed as the completeness of the generated report.

$$\text{Recall} = \frac{\text{True positive}}{\text{True positives} + \text{False negatives}}$$

False positives are the number of items in the generated report that are deemed irrelevant according to the golden standard.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 13 Analysis of relevance and completeness of generated reports

	R-1	R-2	R-3	R-4	R-5	R-6	R-7	R-8	R-9	R-10	μ
Precision	0,333	0,667	0,833	0,75	1	0	0,5	1	1	0,667	0,675
Recall	0,125	0,16	0,455	0,167	0,063	0	0,071	0,231	0,111	0,4	0,178
F1-score	0,182	0,258	0,588	0,273	0,112	0	0,125	0,375	0,2	0,5	0,261
Incorrect Items (FPs)	2	2	1	1	0	2	1	0	0	*0	0,9

*The zero has a mark because the general practitioner used a collective name and not the names individually in the EMR and C2R did.

Chapter 7 Discussion & Conclusion

In this chapter the limitations and the future research is discussed. The limitation is about, if the research is generalizable and what has to be taken into account. The future research describes research possibilities.

7.1 Conclusion

In this research, the research data management process and storage of multimodal data is analyzed in two phases. In the first phase I looked into the literature and how multimodal data should be stored, second is the process of testing the current prototype and the organization of the extracted output data. The research is focused on the main research question “*How can research data management for multimodal analysis and testing be organized?*”. To answer this question it is divided in six research questions.

Research question 1 is stated as follow: “*What is currently known about research data management in relation to multimodal analysis and testing?*”. There is a distinction made between the terms multimodal and multimedia, because the terms are used interchangeably in the literature. There are two definitions created as how the terms will be used in this research and the C2R program. By reading multiple papers and definition statements, two final definitions are determined: **Multimodal**; *Multimodal is the communication in the form of sound and (human) voice or visual like images, text, movies or sensor measurements etc. or in combination.* **Multimedia**; *Multimedia is two or more communications media of content in the form of audio, video or sensor data etc.* The modalities such as speech in audio files, consultation treatments in video recordings and measurements through sensor data are the input for the C2R system. The literature defines the definition of research data. Research data is everything that is produced, used or created during the whole research program. Multimodal and multimedia input and output is also research data from this research. Multimodal analysis focus on all types of communication and the analysis exist out of multiple steps. The first step is the gathering of the multimodal, the multimodal data is analyzed by watching and playing the multimodal data to get insight and understanding of the data. For these data there is a representable sample selected and transcript into text. Multimodal testing of multimedia systems are applications that make use of a variety of media, such as video, audio and text. The main application could run on servers and make use of various micro servers of third parties. To test such systems, it could be in multiple ways but this research focusses on the occurred errors. The errors that occur could tell us about the performance of the system or individual micro analyzers.

The second research question “*What are the rules and processes for executing an (simulated/ real) experiment that extracts audio, video and measurements data from medical consults?*” The multimedia input for the C2R system are videos of the treatment, audio of the conversations between the patient and the general practitioner and measurements of the patient’s condition. In this research the recording guidelines are followed that cover the most relevant steps. It is important to make clear what the purpose is of the recordings. For this research we made use of the opt-in method for patient recruitment. The patient recruitment is finding the right patient. With this method we only record the patient after an agreement which is made before the consultation. For privacy the patient is (mostly) filmed from behind, also the personal data is left out. The recording is recorded in a sequence to reduce the time investment for the participants. By making use of a real-world general practitioner, the ecological validity is equal to real-life situations. The created recordings, transcripts and informed consents are stored in the research folder of C2R on Yoda. The process of recording a medical consultation for testing is received from an interview with coordinator of the research program Communication in Healthcare at Nivel. According to Nivel there is a distinction between real patients and simulated patients. When the data is gathered by using the right defined informed consent there is a drawback in using it to test the prototype. Real-world recordings that used a real patient must be recorded in a 45° angle from behind.

Third, research question 3 *“How should research data be stored and organized in a corpus with simulated and real medical consultations?”*. Data management is not only managing the data, it is also the data collection, storage and analysis. To cover all the important aspects of data management the DAMA-Framework is used. All the individual topics of the framework are covered in chapter 3.4.1. Research data management is a process of labeling, storing and making the data accessible in a research project. The data is stored on computers, shared services or storage platforms. The data management plan for research data provides a roadmap for the activities and how to handle the data. In chapter 3.5.1. I discussed the ten rules for good research data management. Research data is moving around, in and out databases and research are struggle with organizing their huge volume of data. In 80% of the time research data is stored in unstructured formats. Storing in a multimodal corpus, is different than storing it in a traditional database by capacity, data transfer and real-time obtaining of the multimodal data. In chapter 3.6 a comparison between traditional database and multimodal database is made. The real medical consultation that makes use of patient data, could be stored best within Nivel. Because of the strict regulations they are stored in a server that is not connected or accessible by a network. To get access to the server the person has to be physically access the server room.

The fourth research questions: *“What are the rules and processes for testing privacy sensitive data of care provider and patients?”*. In research question 2 is explained how real recordings are recorded, here is explained how the real recordings could be used within the rules. The most important is to arrange the informed consent. This is an agreement to use the data as defined. The data gathered by the research (group) cannot be used in any other manner. When the recording is finished it is transferred from a memory card to a server that is disconnected from the network. So personal data has to be anonymized. The solution for C2R program is to join the new to be build portal. The portal is for doctors to evaluate their own consultation. The portal gives C2R access to more real-world recordings.

The fifth research question: *“How can real-world experimentation be implemented in the Care2Report program?”*. When the informed consent state that the data could be used in this research, used to test the C2R system and have been signed by the participant, recordings can be used for that purpose. During the case study the real-world experiments were implemented through the backend. The reason was that the front end does not give an update of the status of micro analyzers. A future research project will focus on the improvement of the front-end. When the front- and back-end is improved new recordings could be implemented in the C2R prototype through the C# module. The C# module will automate the testing environment by relocate the output, results to Yoda and the testing from Yoda to the testing environment.

The sixth research question is: *“How valid is the implemented research data management in the Care2Report program?”*. To determine the validity of the C2R system, the C2R prototype output is compared with the EMR of a real-world general practitioner. The True positives, False positives and False negatives gives insight in the performance of the C2R system per consult. In table 12 and table 13 the R-9, R-10 are created from the real consultations that are recorded during this research. R-8 performed a less complete report. The S,A,P of the SOAP format is not filled in. The O has mixed up a sentence, but the words that are used match with those of the general practitioner. The other created report R-10 does have all the SOAP fields filled in. The output is even more specific than the golden standard. Where the general practitioner used a collective word, the C2R has the individual treatments.

Finally, the main question *“How can research data management for multimodal analysis and testing be organized?”*. For the data management Yoda can be used to store and secure all the research data. For testing, the functional design can be used to create an automate process. Analysis of the results can be done by comparing the C2R output with the results of a general practitioner.

Reflection on motivation for this research

In the motivation I talked about my personal interest and the experiences of family members whom work in health care. Their experience is that the administrative burden takes a lot of time, time they rather spend with their patients. The C2R prototype has made the first step into automatic generation of an EMR report after a medical consultation. Before the start of this research there were no reports as output created by the C2R prototype, after this research a number of real-world and simulated reports are generated. The results are divided from, no results till a complete filled in report. With the results arising from this study, the current C2R prototype could be improved.

7.2 Limitations

The limitations of this research are discussed according to the four aspects internal validity, external validity, construct validity and conclusion validity of (Wohlin, 2014).

Internal validity is if the treatment or condition is making a difference in the possible outcome. It can be rephrased in the treatment causes the outcome. To answer this and explain the confidence of the findings within this research. The confidence of the results that express the quality of the generated report is high. The output does not change if it is executed multiple times. The output is compared against the EMR of the general practitioner.

The completeness of the results for the irrelevant words/ sentences can be questionable. The words/ sentences are marked irrelevant when they did not contribute to the EMR created by the general practitioner. The words/ sentences that are irrelevant can be relevant when creating a report for another illness.

external validity is about the generalization of a research. There are some threats for generalization for this research. The whole research is focused on the C2R system which is for now the only one that exists. For the topics data management and recording set-up can be generalized for other researches whether or not in another context. The functional design, data analysis and the created results are C2R specific.

construct validity determines how well a test has measured what it is supposed to measure. The test measures the generated reports from the C2R against the EMR of a general practitioner. When another general practitioner fills in the EMR with different use of words the measurement will stay the same, but the results could change. So, the test that is used measures what it is supposed to. What is lacking, is testing of the internal performance of the microanalyzers. That's why they are included in the future work section.

Conclusion validity determines the validity of the relationship between the treatment and the outcome. There is no clear threat perceived within the validation of the C2R prototype and the results. All of the results are validated with multiple researches within the C2R program. All the possible threats are discussed before it is added in this report. All the analyzes are performed by one research, so there is no risk of a variation in the interpretation of the results.

7.3 Future research

In the future work I will discuss the topics that need some more research. The topics that need more research are within the C2R program and to create improvements. The topics are presented below;

Irrelevant sentences and words filtering could help the C2R system in finding the most relevant sentences and words. The selection of the relevant sentences shortens the sentences that is used of creating the triples. The sentences and words can be placed in a so-called blacklist. To create a relevant blacklist, more consultations has to be compared to select the not required words. Also, more type of medical conditions and treatments gives better insight in the irrelevant sentences and words.

An additional idea is to give the words and sentences a weight in percentage. Where 100% stands for always leaving out and for example 20% is not always required to be left out. The threshold value can be set on which sentences and words to filter out.

Improvement the algorithm for triple matching. Currently there are many tripples not matched despite the availability in the ontology. One of the reasons is that matching could not been done over multiple lines. When medical terms are available in a triple and the following triple contains a related (medical) term it cannot be related.

Improving the ontology will help to find more medical, anatomy, treatment, observation, evaluation and symptoms. When the most common terms are added to the ontology, then also their synonyms can be added.

New test environment (automatically) would help the number of tests that can be runned in a short period of time. The new testing environment should be accessible from a browser so it can be accessed from different geographical locations. Also, the server must be arranged so the test data could be import data from Yoda and external.

A **portal has to be set up with Nivel** to get more real-world recordings. The consolation with real patient data must be stored at Nivel because of privacy law. To use this data a portal has to be set up to get access to real-world medical consultations. Nivel has agreed to cooperate too providing access for the C2R program.

Not all the bad triples can be assigned to the C2R system. The different microanalyzers create less useful triples. It's important to **unit test the microanalyzers** individually to find out their performance. Unit testing is already possible in the C2R prototype.

7.4 Important findings

The research questions are answer, there are more findings made in this research. Here some findings that can be used for the future research.

In the current prototype is it not possible to match the triples over multiple lines. When the “my ear hurts” is caught in a triple and “it is also leaking fluid” is the next triple, the prototype could not understand that the ear has two symptoms.

In the micro analyzer Fred gives multiple time errors back. The reason is not known, it could be a word combination, length of a sentences or punctuations. Because it is a third party, it is not possible to find the problem.

As seen before in chapter 5.2, consultations received better results in the matched triples. The logs show that especially Frog responded well to larger sentences.

References

- (ISO/IEC). (1993). ISO/IEC 2382-1:1993 Information Technology -- Vocabulary -- Part 1: Fundamental Terms. (ISO).
- Adjeroh, D. A., & Nwosu, K. C. (1997). Multimedia database management - Requirements and issues. *IEEE Multimedia*. <https://doi.org/10.1109/93.621580>
- Ajami, S. (2016). Use of speech-to-text technology for documentation by healthcare providers. *National Medical Journal of India*.
- Allwood, J. (2008). Multimodal Corpora. *Corpus Linguistics. An International Handbook*. <https://doi.org/10.1007/978-3-642-04793-0>
- Anastopoulou, S., Baber, C., & Sharples, M. (2001). Multimedia and multimodal systems: commonalities and differences. In *5th Human Centred Technology Postgraduate Workshop*, University of Sussex.
- Anderson, N. R., Lee, E. S., Brockenbrough, J. S., Minie, M. E., Fuller, S., Brinkley, J., & Tarczy Hornoch, P. (2007). Issues in Biomedical Research Data Management and Analysis: Needs and Barriers. *Journal of the American Medical Informatics Association*. <https://doi.org/10.1197/jamia.M2114>
- Andeweg, M. E., Terluin, M., Boendermaker, P. M., & Pols, J. (2002). De structuur van een consult een kwalitatieve benadering Wat vinden huisartsen zelf? *Tijdschrift Voor Medisch Onderwijs*. <https://doi.org/10.1007/bf03056549>
- André, E. (2000). The generation of multimedia presentations. In *Handbook of natural language processing*.
- Antoniou, G., & Harmelen, F. van. (2004). *Semantic Web Primer*. The MIT Press.
- Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*. <https://doi.org/10.1007/s00530-010-0182-0>
- Berenschot and VU (2018). Enquête administratieve belasting langdurige zorg – editie 2018. https://www.waardigheidentrots.nl/wp-content/uploads/2018/11/Enquete_administratieve_belasting_langdurige_zorg_2018.pdf
- Bezemer, J., & Jewitt, C. (2010). Multimodal analysis: Key issues. *Research methods in linguistics*, 180-197.
- Bierings, H. (2017, April). Beroep en werkdruk in Nederland. Retrieved from Beroep en werkdruk: <https://www.cbs.nl/nl-nl/achtergrond/2017/14/beroep-en-werkdruk>.
- Bishoff, C., & Johnston, L. (2015). Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University. *Journal of Librarianship and Scholarly Communication*. <https://doi.org/10.7710/2162-3309.1231>
- Boll, S., Klas, W., & Sheth, A. P. (1998). Overview on using metadata to manage multimedia data. In *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*.
- Bonacchi, S., & Karpiński, M. (2014). Remarks about the use of the term “multimodality.” *Journal of Multimodal Communication Studies*.
- Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. <https://doi.org/10.1002/asi.22634>
- Budgen, D., & Brereton, P. (2006). Performing systematic literature reviews in software engineering. In *Proceeding of the 28th international conference on Software engineering - ICSE '06*. <https://doi.org/10.1145/1134285.1134500>
- Burad, A. (2006), “Multimedia Databases”, *Computer Science and Engineering*, Indian Institute of Technology, April 6, 2006.
- Bushong, V. (n.d.). *Multimedia Database Management Systems (MMDBMS): An overview of design, retrieval methods, and applications*. Evangel University.
- Cameron, S., & Turtle-Song, I. (2002). Learning to write case notes using the SOAP format. *Journal of Counseling and Development*. <https://doi.org/10.1002/j.1556-6678.2002.tb00193.x>
- Caschera, M. C., Ferri, F., & Grifoni, P. (2007). Multimodal interaction systems: information and time features. *International Journal of Web and Grid Services*. <https://doi.org/10.1504/ijwgs.2007.012638>

- Cox, A. M., & Pinfield, S. (2014). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science*.
<https://doi.org/10.1177/0961000613492542>
- Cupoli, P., Earley, S., & Henderson, D. (2014). DAMA-DMBOK2 Framework. DAMA International.
- D'Ulizia, A., Ferri, F., & Grifoni, P. (2010). Generating multimodal grammars for multimodal dialogue processing. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*. <https://doi.org/10.1109/TSMCA.2010.2041227>
- Dama International. (2009). *The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide)*. Bradley Beach, NJ-USA.
- Dictionary, O. E. (n.d.). Lexico by Oxford. Retrieved April 2019, from Oxford English Dictionary:
<https://www.lexico.com/en/definition/data>
- Duineveld, B., Kole, H. M., Van Werven, H., & Njoo, K. H. (2019). NHG-Richtlijn Adequate dossiervorming met het elektronisch patiëntdossier (ADEPD). Nederlands Huisartsen Genootschap.
- Essoussi, D. (2019). *Privacy considerations in automated medical reporting*. Utrecht: Utrecht University.
- Fienberg, S. E., Margaret, M. E., & Miron, S. L. (1985). *Sharing Research Data*. In *Sharing Research Data*. Washington, D.C.: NATIONAL ACADEMY PRESS.
- Francissen, P. W. H. M. (2004). Wet op de geneeskundige behandelingsovereenkomst (WGBO). In *Cliënt en Medezeggenschap in de zorg*. https://doi.org/10.1007/978-90-313-8646-8_38
- Franks, E. W. (1967). The MADAM System: data management with a small computer (No. SDC-SP 2944/000/00). SYSTEM DEVELOPMENT CORP SANTA MONICA CA.
- Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y., & Tamura, S. (2001, May). Ubiquitous speech processing. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) (Vol. 1, pp. 13-16)*. IEEE.
- Fylan, F. (2005). Semi-structured interviewing. In *A Handbook of Research Methods for Clinical and Health Psychology Semi-structured interviewing*.
<https://doi.org/10.1093/med:psych/9780198527565.001.0001>
- Gemmell, D. J., Vin, H. M., Kandlur, D. D., Rangan, P. V., & Rowe, L. A. (1995). *Multimedia Storage Servers: A Tutorial*. Computer. <https://doi.org/10.1109/2.384117>
- Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2006). Scientific data management in the coming decade. *ACM SIGMOD Record*.
<https://doi.org/10.1145/1107499.1107503>
- Hanekamp, M. (2016, 07 13). Administratieve taken langdurige zorg kosten jaarlijks € 5 miljard. Retrieved from Berenschot: <https://www.berenschot.nl/actueel/2016/juli/administratievetaken/>
- Hoschek, W., Jaen-Martinez, J., Samar, A., Stockinger, H., & Stockinger, K. (2000). Data Management in an International Data Grid Project. https://doi.org/10.1007/3-540-44444-0_8
- Katz, B., Lin, J. J., & Felshin, S. (2002). The START Multimedia Information System: Current Technology and Future Directions. In *International Workshop on Multimedia Information Systems*.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical Report TR/SE-0401, Keele University, and Technical Report 0400011T.1, NICTA.
- Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*. <https://doi.org/10.1016/j.infsof.2008.09.009>
- Krishnankutty, B., Naveen Kumar, B., Moodahadu, L., & Bellary, S. (2012). Data management in clinical research: An overview. *Indian Journal of Pharmacology*. <https://doi.org/10.4103/02537613.93842>
- Lauer, C. (2009). Contending with Terms: "Multimodal" and "Multimedia" in the Academic and Public Spheres. *Computers and Composition*. <https://doi.org/10.1016/j.compcom.2009.09.001>
- Lu, G. (1999). *Multimedia Database Management Systems*. Boston, MA: Artech House, Inc.
- Luchies, E., Spruit, M., & Askari, M. (2018). Speech technology in Dutch health care: A qualitative study. In *HEALTHINF 2018 - 11th International Conference on Health Informatics, Proceedings; Part of 11th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2018*.

- Maas, L., Geurtsen, M., Nouwt, F., Schouten, S., van de Water, R., van Dulmen, S., . . . Brinkkemper, S. (2019). The Care2Report System: Automated Medical Reporting as an Integrated Solution to Reduce Administrative Burden in Healthcare. HICSS20.
- Maynard, D. W., & Heritage, J. (2005). Conversation analysis, doctor-patient interaction and medical communication. *Medical Education*. <https://doi.org/10.1111/j.1365-2929.2005.02111.x>
- McGowan, T. & Gibbs, T. A. (2009) Southampton Data Survey: Our Experiences & Lessons Learned [unpublished].
- Michener, W. K. (2015). Ten Simple Rules for Creating a Good Data Management Plan. *PLoS Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1004525>
- Misic, V. B., Chanson, S. T., & Shing-Chi Cheung. (2002). Towards a framework for testing distributed multimedia software systems. <https://doi.org/10.1109/pdse.1998.668160>
- Mittal, S., & Mittal, A. (2011). Versatile question answering systems: seeing in synthesis. *International Journal of Intelligent Information and Database Systems*. <https://doi.org/10.1504/ijiids.2011.038968>
- Mohammed, O.B. (2014). Application of Multimedia Technology in Database and IT Service Management.
- Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*. <https://doi.org/10.1016/j.infoandorg.2006.11.001>
- National Science Foundation. (2014). Chapter II, Section C.2.j: Special information and supplementary documentation. Grant proposal guide. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp
- NHG-Standaarden voor de huisarts 2009. (2009). NHG-Standaarden voor de huisarts 2009. <https://doi.org/10.1007/978-90-313-6614-9>
- O'Halloran, K. L., & Smith, B. A. (2012). Multimodal Text Analysis. In *The Encyclopedia of Applied Linguistics*. <https://doi.org/10.1002/9781405198431.wbeal0817>
- Ong, L. M. L., Visser, M. R. M., Kruijver, I. P. M., Bensing, J. M., Van Den Brink-Muinen, A., Stouthard, J. M. L., . . . De Haes, J. C. J. M. (1998). The Roter Interaction Analysis System (RIAS) in oncological consultations: Psychometric properties. *Psycho-Oncology*. [https://doi.org/10.1002/\(SICI\)1099-1611\(1998090\)7:5<387::AID-PON316>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1099-1611(1998090)7:5<387::AID-PON316>3.0.CO;2-G)
- Perrier, L., & Barnes, L. (2018). Developing Research Data Management Services and Support for Researchers: A Mixed Methods Study. *Partnership: The Canadian Journal of Library and Information Practice and Research*. <https://doi.org/10.21083/partnership.v13i1.4115>
- Rathert, C., Mittler, J. N., Banerjee, S., & McDaniel, J. (2017). Patient-centered communication in the era of electronic health records: What does the evidence say? *Patient Education and Counseling*. <https://doi.org/10.1016/j.pec.2016.07.031>
- Roberts, P., Priest, H., & Traynor, M. (2006). Reliability and validity in research. *Nursing Standard*. <https://doi.org/10.7748/ns2006.07.20.44.41.c6560>
- Rohloff, K., Dean, M., Emmons, I., Ryder, D., & Sumner, J. (2007). An Evaluation of Triple-Store Technologies for Large Data Stores. In *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*. https://doi.org/10.1007/978-3-540-76890-6_38
- Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., & Schiele, B. (2013). Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*. <https://doi.org/10.1109/ICCV.2013.61>
- Rooijackers-Lemmens, E., Van Wijngaarden, J. J., Opstelten, W., Broen, A., Romeijnders, A. C. M., & Geijer, R. M. M. (1995). NHG-STANDAARD OTITIS EXTERNA. *Huisarts En Wetenschap*. https://doi.org/10.1007/978-90-313-8279-8_13
- Roter, D. (2006). *THE ROTER METHOD OF INTERACTION PROCESS ANALYSIS*. Baltimore, Maryland: The Johns Hopkins University.
- Sheppard, J. E., Weidner, L. C. E., Zakai, S., Fountain-Polley, S., & Williams, J. (2008). Ambiguous abbreviations: An audit of abbreviations in paediatric note keeping. *Archives of Disease in Childhood*. <https://doi.org/10.1136/adc.2007.128132>
- Smeele, T., & Westerhof, L. (2018). Using iRODS to manage, share and publish research data: Yoda. Durham, North Carolina, USA: iRODS UGM 2018 June 5-7.
- Smith, P. L. (2014). Exploring the data management and curation (DMC) practices of scientists in research labs within a research university. Dissertation.

- Subramanya, S. R. (1999). Multimedia databases. *IEEE Potentials*. <https://doi.org/10.1109/45.807273>
- Surkis, A., & Read, K. (2015). Research data management. *Journal of the Medical Library Association : JMLA*. <https://doi.org/10.3163/1536-5050.103.3.011>
- Takayoshi, P., & Selfe, C. L. (2007). Thinking about multimodality. *Multimodal Composition. Resources for Teachers*.
- The University of Texas at Dallas. (n.d.). Elements of Cinematography: Camera. Retrieved from school or arts, technology, and emerging communication: <https://www.utdallas.edu/atec/midori/Handouts/camera.htm>
- Tompkins, Anne. (2007). Data Management in Clinical Trials. Ch7. *Principles and Practice of Clinical Research*. 2nd Edition. John Gallin Frederick Ognibene .
- Turk, M. (2001). Perceptual User Interfaces. In *Frontiers of Human-Centered Computing, Online Communities and Virtual Environments*. https://doi.org/10.1007/978-1-4471-0259-5_4
- University of Edinburgh. (2014, Sep). About the Research Data Service. Retrieved March 2019, from Information Services Research Data Service: <https://www.ed.ac.uk/information-services/research-support/research-data-service/about-the-research-data-service>
- University of Edinburgh Information Services. (2014, sep). Research Data Service. Retrieved March 2019, from Information service: <https://www.ed.ac.uk/information-services/research-support/research-data-service>
- van der Kooij, J., & Lim, M. (2019). Automatically identifying the relevant information from the patient-care provider dialogue for documentation in the patient's Electronic Medical Record. Utrecht: University of Utrecht.
- van Dulmen, S., Humphris, G., & Eide, H. (2012). Towards a guideline for person-centered research in clinical communication; lessons learned from three countries. *International Journal of Person Centered Medicine*. <https://doi.org/10.5750/ijpcm.v2i1.89>
- Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence - Video to text. In *Proceedings of the IEEE International Conference on Computer Vision*. <https://doi.org/10.1109/ICCV.2015.515>
- Walsh, K. E., & Gurwitz, J. H. (2008). Medical abbreviations: Writing little and communicating less. *Archives of Disease in Childhood*. <https://doi.org/10.1136/adc.2008.141473>
- Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*. <https://doi.org/10.12101/12213>
- Whittaker, J. A. (2000). What is software testing? And why is it so hard? *IEEE Software*. <https://doi.org/10.1109/52.819971>
- Wieringa, R. J. (2014). Design science methodology: For information systems and software engineering. *Design Science Methodology: For Information Systems and Software Engineering*. <https://doi.org/10.1007/978-3-662-43839-8>
- Wiggins, B. (2012). Chapter 7 Data management. In *Effective Document and Data Management : Unlocking Corporate Content*. Routledge.
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. <https://doi.org/10.1145/2601248.2601268>
- Yoda Utrecht University. (2019, juli 17). Introduction to Yoda. Retrieved from Yoda: <https://yoda.sites.uu.nl/home/introduction-to-yoda-2/>
- Zhang, J., Cheung, S.-C., & Chanson, S. T. (1999). Stress Testing of Distributed Multimedia Software Systems.
- Zhang, J., & Cheung, S. C. (2002). Automated test case generation for the stress testing of multimedia systems. *Software - Practice and Experience*. <https://doi.org/10.1002/spe.487>

Appendix