www.kconnect.eu

Deliverable 4.2
Requirements for Medical Record Analysis and Search

Deliverable number	*D4.2*
Dissemination level	*Public*
Delivery date	*21 Aug 2015*
Status	*Final version*
Author(s)	*Agata Rukat. Fredrik Axelsson. Johan Sjöberg, Rebecka Berger, Richard Dobson*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect)

Executive Summary

This document specifies how information processing and search technologies will be used most effectively for KConnect system to deliver value in the end user setting. It also covers technical requirements such as API calls and output formats. It also covers requirements to do with data, including satisfying the ethical requirements for the processing of medical records.

Requirements in this document target two separate uses of the KConnect services, one intended for medical health record analysis within Swedish health care sector and the other for psychiatric health record analysis for patients in the U.K. The KConnect services must be adapted to support the use cases of both applications.

The end KConnect applications will use three main KConnect services/components. An information processing service provided by GATE, index and search service provided by Mimir and a semantic knowledge base service provided by Ontotext's GraphDB™. In addition to the functionality developed in the earlier Khresmoi project, the following component requirements have been identified:

GATE's information processing should be able to identify medical entities in Swedish and English medical records. This include treatments and interventions, health care events, negation of symptoms/medical events and medical practitioners.

Mimir's index and search service should be able to provide medical events sorted by dates.

The Knowledge base should provide information about: relations between symptoms and possible diseases, relations between comorbidities and medication, relations between medication interactions from the literature, adverse drug reactions, incompatibility of medications and symptoms, incompability of treatment and symptoms as well as a larger amount of relations between symptoms and diseases.

Table of Contents

List of Abbreviations

CRIS	Clinical Record Interactive Search
EHR	Electronic Health Record
GATE	General architecture for text engineering
GATE WASP	GATE API
KB	Knowledge base
KCL	Kings College London
NLP	Natural Language Processing
SLAM	The South London and Maudsley NHS Foundation Trust
QUL	Qulturum (Improvement unit at Region Jönköping County)

1 Introduction

Between 2010 and 2014, Khresmoi developed a multilingual multimodal search and access system for biomedical information and documents. The services used in KConnect are based on the services in the Khresmoi system.

This document will describe how to apply the KConnect services to business cases around electronic health record analysis and search provided by a hospital, King's College London (KCL), and a health organization, Qulturum (QUL). The business cases are to make use of the KConnect technology stack to construct search based solutions based on medical health records. This document will describe the requirements to be fulfilled by the KConnect technology stack. Each technology requirement will be presented along the use case requiring it. The document is divided into two parts, describing the methodology and requirements for the two end users (KCL and QUL) separately.

The health records used by the project will be actual records from patients in Sweden and United Kingdom respectively. The records are a combination of structured and unstructured data that is the result of interactions between the patient and the caregiver. Region Jönköping County, the main health care provider in the country of Jönköping Sweden, will provide the Swedish health records, The South London and Maudsley NHS Foundation Trust (SLAM) will provide the English records through KCL.

2 Background

The KConnect project aim to create services based on electronic health records. The services will be constructed with respect to the presence of personal health information. The project mainly use components developed in the earlier Khresmoi project. These components handle processing of text information, providing index and search functionality along with semantic knowledge in the biomedical domain.

2.1 Technology

The KConnect technology stack consists of several components; a semantic search engine (Mimir), a modular text processing pipeline (GATE) and a medical knowledge base (Knowledge Base). A new KConnect application will use these services and via present the KConnect functionality to the end user.

2.1.1 Search technology

Search technology is a fundamental concept of the KConnect technology stack. Medical health records and medical literature alike will be used to cater the need of KConnect users and the search system is integral to leverage that service.

The search engine used is Mimir, which is a product in the GATE ecosystem. In contrast to other search engines it features semantic indexing to combine the power of free-text search with that of structurally annotated data. Information and technologies need to be adapted to the contract provided by Mimir and further to respect the denormalization of information, which is an essential characteristic of any free-text search engine.

2.1.2 Information processing

In order to utilize the full power of search technology data must be structured. When dealing with information as text written in natural language most of the data is unstructured. Text analytics tools can be used to capture knowledge in the data by analyzing the text and derive features from the information in the text. The use cases make use of specific features identified in medical records and medical literature. These features include mention of named entities in the text, such as diseases, diagnoses, symptoms and anatomy. The entities can be linked against complimentary data sets to semantically strengthen the original information.

Text analytics tools works by segmenting the text into smaller chunks and utilizes language specific rules or statistics to extract desired features. Hence most text analytic tools are language specific. The text processing service employed by KConnect is the GATE application. GATE works well in combination with Mimir and excels with low coupling, high modularity and high extensibility.

2.1.3 Semantic knowledge

The semantic search engine makes heavy use of a medical knowledge base to provide semantic relations between medical entities. The knowledge base is an ontology implemented in Ontotext's GraphDB™.

3 Qulturum

This section describes the application of the KConnect system that will be used at the Region Jönköping County by which use cases have been identified in work led by the improvement unit Qulturum.

3.1 Methodology

In order to specify how the services will be used most effectively in the end user setting, use cases were gathered and evaluated. This section will describe the process of finding the final use cases and we will provide some of the use cases as examples.

3.1.1 Use case elicitation

In order to make sure that KConnect meets the needs of end-users, representatives of different healthcare professions who have legitimacy to diagnose and decide treatment were invited to take part in interactive workshops. We also invited care administrators as they are administrating health care records on daily basis. The professionals that took part in the workshops represented five professions i.e. resident and specialized doctors, nurses, physiotherapists, psychologists and care administrators. We asked all of them to identify use cases for KConnect by answering if KConnect could help them in their daily work and if so, how.

The participants were informed about KConnect and the ambition behind the project and were asked to identify use cases for KConnect by answering if KConnect could help them in their daily work and if so, how. The participants wrote several use cases each answering the questions:

1. Who will use the search engine in this case?
2. What is the use for the search engine in this case?
3. Why should they use the search engine in this case?

During the workshop the participants engaged in several discussions and elaborated their thoughts around the use cases.

Apart from the interactive workshops, use cases were gathered from health care professionals participating at the Swedish National Conference in Healthcare ‘’Utvecklingskraft’’ in Jönköping, Sweden. In total, 56 use cases were gathered at the workshops and the conference.

3.1.2 Use cases - out of scope

Some of the use cases contain aspects that could be of interest for the KConnect project – but not of primary interest due to certain circumstances. They could for example describe a function that would both be interesting and of value for the search engine, but not viable in terms of lack of applicable data or requirements of system integration not possible within the project. Some use cases are much specific and only of value for very limited groups of health care professional and other use cases describe minor alterations of major functions already covered by the primary use cases or described in the overall project proposal.

3.1.3 Prioritization of use cases

In order to prioritize the use cases we firstly decided to focus on which target groups that the use case would address. The group “care provider” is for example the biggest target group that KConnect could be of use for, so naturally these cases regarding care providers were ranked high. The following use cases will therefore only target the group care provider.

Secondly we looked at the use cases that would directly impact and improve the care process and increase the patient safety. All use cases are of course a part of improving the situation for the patient, i.e. the end user, but some were more direct than others. We decided to more or less focus on the use cases that were directly aimed at improving care for the patients, i.e. assist care providers in identifying symptoms for complex diagnosing, receive warnings about medication or how they can find expertize to consult within the own organization. Another parameter that was closely looked at was how the KConnect service could diminish the burden of administration for the care providers and the use cases containing requests for this particular theme were ranked high. And as constant quality improvement is essential for every organization, we also took the use cases focused on this theme into close consideration. The use cases were grouped and categorized. The categories that contained most use cases were considered of more interest and ranked high, but the nature of the category and its impact on the care process and the patient were also considered. Some use cases describe processes of a more general character and some more detailed, in these cases the more general were ranked higher, due to their ability to benefit more target groups, but in some cases the more detailed use cases could provide an improvement of a current care situation and therefore be of great value for the care provider and form a unique selling point for KConnect.

3.2 Use cases and requirements

This section will describe the use cases identified by Qulturum along with the requirements each component in the technical stack need to fulfil. Each use case is presented with an example, a scenario, a prototypical scenario and the requirements the use case puts on the technology stack.

Given the that the project is still in its beginning it is important to understand that the presented use cases are still in development and may be altered during the project, depending on how the requirements can be realized.

3.2.1 When to use KConnect services

During a medical consultation, the care provider should divide most of the time provided focused on and in interaction with the patient. Regardless when the care provider choose to use information provided by KConnect, before, during or after the patient visit it is of outmost importance that the care provider finds KConnect useful. To be useful it has to be easy to operate, easy to understand and accessible when the care provider needs it. Ideally the long term goal is to completely integrate KConnect technology into the EHR system. If that is not achievable the search engine must be easily accessible in combination with a value offering that is greater than the effort of accessing separate systems.

3.2.2 Use case: Identify diseases

- As a care provider
- In order to quickly identify the risk for diseases such as cancer
- I want the system to provide risk for diseases given the symptoms of a patient

3.2.2.1 Example

A patient has a symptom or numerous symptoms that alone or combined could indicate a certain disease.

3.2.2.2 Scenarios

Scenario 1:

A patient visits the primary care. During the examination the doctor types the patient's symptoms in KConnect; for example weight loss, coughing with blood, thorax pain etc. and the doctor gets a warning that these symptoms combined could indicate that the patient might have cancer.

Scenario 2:

A nurse meets an elderly patient that has lost some weight recently, has no appetite and is medicated with anti-depressant drugs. When the nurse types these symptoms in KConnect she/he should get a warning, for example "risk for malnutrition"! The nurse should also be presented with suggestions for the actions/additional measures to take to prevent it.

Scenario 3:

A nurse at the medical ward/clinic meets an inpatient with symptoms such as: high fever, cough and feeling generally ill. The nurse types these symptoms in KConnect and receives a warning, for example "suspicion of infection" and also suggestions for actions to take.

System should be applicable to all professions.

3.2.2.3 Prototypical scenario

1. The care provider presents the patient's symptoms to the KConnect application
    a. The system provides query suggestion for possible symptoms
    (or automatically extracted from the live record by a KConnect module in the EHR system)
2. A query to the KB to find diseases linked to the all of the symptoms or parts of the symptoms is created by the application
3. Possible diseases/diagnosis are provided by the KB
4. A list of possible diseases are presented to the care provider

3.2.2.4 Requirements on components

GATE text processing

- Identify symptoms in medical records
- Link the identified entities to concepts in the KB

Knowledge base

- Provide information about relations between symptoms and possible diseases

3.2.3 Use case: Present record summary

- As a care provider
- In order to get an overview of all care episodes (diagnosis and treatments) and know which care providers have been interacting with the patient previously
- I want the system to provide a summary of all diseases/treatments/medication and the people who has interacted with the patient

3.2.3.1 Example

A care provider should be able to get a clear and comprehensive overview of the patient's previous diagnosis/treatment etc, in order to be better prepared before meeting the patients alt. be able to get a quick overview during consultation to be able to assist in a more correct diagnosis and treatment.

3.2.3.2 Scenario

A doctor or a nurse meets a patient at the primary care. The patient suffers from back pain. A nurse/doctor should be able to see an overview of all previous care episodes incl. diseases, treatments, medications and care providers in order to get a better understanding of what could be the reason for the back pain. This overview should be as concise as possible so the care provider clearly can see connections with new symptoms to previous diseases/symptoms the patient has.

3.2.3.3 Prototypical scenario

1. The patient's previous records have been previously processed and indexed
    a. The records are crawled/provided on a regular basis
    b. The records are fed into the GATE pipeline where they are enriched with semantic information. Identify entities:
        i. Symptoms (w/ negations)
        ii. Treatments/Interventions
        iii. Diseases (w/ negations)
        iv. Practitioners (in contact with the patient)
    c. The records are indexed in Mimir based on the enriched information
2. The care provider provide info about the patient to the KConnect application (or automatically provided from KConnect module in EHR system)
3. The care provider asks the KConnect application for a summary
4. A Mimir query are automatically designed for retrieving all events related to the patient
5. Mimir provides all events sorted by time and grouped by category of the event
6. The events are presented in chronological order to the care provider in the KConnect application

3.2.3.4 Requirements on components:

GATE text processing

- Identify medical practitioners in the medical records
- Identify health care events
- Identify negation of symptoms/medical events

Mimir

- Get all events/annotations from a specified patient
- Get all practitioners who have interacted with a patient
- Get all practitioners connected to a given treatment/diagnosis
- Sort presented events based on date

3.2.4 API requirements

The following requirements consider the communication between the KConnect components. All components shall work as standard web services. The communication between the services should be standardized, to make it easy to add or replace obsolete services. The services should preferably communicate with RESTful communications, with an open message format standard such as JSON or XML.

3.2.5 Ethical requirements

The services will be constructed with respect to the presence of personal health information. Many of the rules and regulations are regulated in Patientdatalagen[1]. In order to assure that sensitive patient data is not in any danger of being accessed by unauthorized people, the data should be stored locally within Region Jönköping County. To maintain the high privacy standard, the data that accessible via KConnect must be anonymized according to the premises set by Region Jönköping County in consideration of the Swedish Patient Act and the Swedish Personal Data Act.

All EHR data that directly could reveal the identity of a patient, outside of a care giving situation, is in violation of the privacy standard. The Swedish personal number (Social Security Number) is a unique identifier of a patient in Sweden and therefore considered direct data. A patient's forename and surname is also considered sensitive both in combination or provided separately. Any names connected to the identity of the patient that are written in free text within the EHR are also considered sensitive. The same goes for contact details of any kind and information of gender.

If a user of KConnect search information on patients not assigned to them, the information provided by the search engine must be anonymized according to the previous established anonymization standards set in the project.

The care provider is only authorized to access a patient's EHR if the patient is assigned to the care provider. Technically care providers can search information about all patients in the EHR system, but are not allowed to do so if the patient is not assigned to them. Every access to the system is logged.

The care provider is presented with a summarized description of the patient's health status produced by the information provided in the patient's EHR. The care provider is connected to the unit the patient is visiting and the information on the patient previously described by the same unit is easy to access. The patient could authorize the care provider to access all parts of his or hers EHR, but in order to see that information the care provider must access the notes from every other unit that the patient has been in contact with.

3.2.5.1 Audit

Logging can be defined as the registration of activities performed in a system, including any information created, loaded or transferred. Registered tasks and events in the log is used to retrospectively reconstruct and analyze which people and systems did what, with what information and when.

All users are informed that logging is done and in what way, and that there is a follow up to the logging. All logs are regularly audited according to an established routine, which includes; what is logged, how often the logs are reviewed, who undertook the audit, but also what is to be regarded as infringement and how breaches will be handled.

Monitoring of log files is a prerequisite for making health information available. The data security guidelines for Region Jönköping County states that any access to the EHR system should be logged. This includes both login to the system, and all kinds of access to various electronic health records. Every search made via KConnect in the EHR system is therefore to be logged.

3.2.5.2 Authorization

The requirements of how Region Jönköping County controls authorization to data are increasing due to the growing complexity of the care process. The care process of today involves staff from several organizations and challenges of both increasing amounts of data and demands of data access.

Basic requirements for security of data in IT-systems is that the data should be accessible, protected against loss of confidentiality, improper use, unauthorized modification and corruption. Permissions to data are to be governed by the information owner.

In order to perform searches in patient health records via KConnect the system must first verify the user's identity authentication. All users of the IT-systems in the Region Jönköping County have a unique identity following the standard of the Health Services Address (HSA). The user identity is personal. The user identity in combination with passwords or service card (SITHS) will be used to ensure that users of the IT-system are authorized to do so. If the user is authorized to access the system, the login must be logged. Any search made in the EHR system via KConnect is also to be logged. There should be a function that ensures automatic logout of the system after a period of inactivity.

3.2.6 Data requirements

The Khresmoi project by which KConnect is based provides a large amount of literature in different languages and quality ensured by the Health on net foundation. In the use cases targeting medical literature it's become apparent that within the different care provider roles there are very specific requirements on both what literature is relevant but also on its origin. For instance the literature overlap between a clinical doctor and a physiotherapist is virtually non-existent. Further in the specific role of a clinical doctor the information sources must not only have a very high credibility rating but the information must also be reviewed by known professionals that can guarantee its accuracy. Similarly, machine translation of literature becomes less attractive due to the reason of accuracy. Apart from only using a narrow selection of sources to cater these use cases the information gathering mechanism must also extract information about the origin of the source such as publish and review date, author, references and details about individuals who has reviewed the information.

The electronic health records have to adhere to the ethical requirements. In particular that means that patient records must be awarded a degree of security that prevents any unauthorized access as well as prevent the information to leave the premise by where it is hosted. To facilitate the anonymization required by some of the selected use cases the electronic health record has to provide not only unstructured text but also a combination of structured and unstructured information. In particular the name of the patient and care provider must be available as structured data.

4 King's college London

This section describes the application of the KConnect system that will be used at the King's College London.

4.1 Methodology

The South London and Maudsley NHS Foundation Trust (SLAM) is the largest provider of mental health services in Europe. The hospital electronic health record (EHR), implemented in 2007, contains records for 250,000 patients in a mixture of structured and over 18 million free text fields.

At the NIHR Biomedical Research Centre for Mental Health and Unit for Dementia at the Institute of Psychiatry, Psychology and Neuroscience, King's College London we have developed unique capacity through the Clinical Record Interactive Search (CRIS) application which allows research use of the anonymized mental health electronic records data, recently exported as a resource to four other UK mental health providers. CRIS has generated substantial, comprehensive and innovative research output since its development and ethics approval in 2008.

Through the CRIS program we have developed a suite of natural language processing information-extraction applications covering a range of hitherto-unrealised constructs such as symptomatology, interventions and outcomes (e.g. adverse drug reactions). However, much of the information within the record is still hidden from the clinician and researcher. This project will sit within this framework of governance and IT infrastructure to provide semantic annotation and semantic search capability across the complete record with integrated biomedical information extracted from the literature knowledgebase.

Such capability will transform the way clinicians and researchers use the EHR and, through our patient led model of 'consent for contact' for example, will facilitate recruitment into trials.

As a result of the CRIS project the governance and infrastructure have already been established and use cases for KConnect were defined well in advance of the KConnect project to a large extent. We provide some of the use cases as examples below:

4.2 Requirements/Mental Health Use cases

This section will describe the use cases identified by KCL along with the requirements each component in the technical stack need to fulfil. Each use case is presented with an example, a scenario, a prototypical scenario and the requirements the use case puts on the technology stack

4.2.1 Use case: Physical symptoms associated with medication

- Detecting/categorising physical symptoms reported in text
- Estimating current medication from text
- Establishing whether the symptom and the medication are linked in research literature.

4.2.1.1 Example

The EHR end-user will receive information when a given physical symptom recorded for/by a patient is consistent with a known side-effect of the medication they are receiving.

4.2.1.2 Scenarios

A patient comes to the secondary care provider, during the examination the doctor enters into KConnect the physical symptoms the patient is also experiencing alongside their mental health related symptoms. The clinician receives a warning that the patient might have developed this symptom in response to medication that they have been earlier prescribed.

4.2.1.3 Prototypical scenario

1. The current CRIS solution provides an anonymized snapshot version of the EHR for research with GATE capability

2. CRIS contains certain reusable components. e.g. binary to text conversion (XHTML), OCR etc.
3. These components will be extended to provide an interface with MIMIR indexing for a near real time view of the patient record for clinical use.

4.2.1.4 Requirements on components

GATE text processing

- Medication and symptom apps

Mimir

- Indexing of medication and physical symptoms

Knowledge base

- Provide information about relations between symptoms and medication

4.2.2 Use case: Comorbidities affecting choice of medication

- Estimating current medication and comorbidity from text
- Consolidating and presenting relevant literature on treatment options with the comorbidity in question

4.2.2.1 Example

The EHR end-user will receive information where comorbidity potentially affects medication choice. Examples include hepatic dysfunction, renal failure and epilepsy.

4.2.2.2 Scenarios

A patient comes to the secondary mental health care provider, during the examination the doctor writes in KConnect the proposed medication. The clinician receives a warning that the proposed medication choice may be affected by the comorbidity burden of the patient.

4.2.2.3 Prototypical scenario

1. The current CRIS solution provides an anonymized snapshot version of the EHR for research with GATE capability
2. CRIS contains certain reusable components. e.g. binary to text conversion (XHTML), OCR etc.
3. These components will be extended to provide an interface with additional MIMIR indexing for a near real time view of the patient record for clinical use.

4.2.2.4 Requirements on components

GATE text processing

- Medication and Comorbidity apps

Mimir

- Indexing of medication and comorbidities

Knowledge base

- Provide information about relations between comorbidities and medication

4.2.3 Use case: Identifying potential for drug-drug interactions

4.2.3.1 Use case

- Estimating current medication from text
- Consolidating and presenting relevant literature on treatment interactions in question.

4.2.3.2 Example

The EHR end-user will receive information identifying potential drug-drug interactions based on current medication.

4.2.3.3 Scenarios

A patient comes to the secondary mental health care provider, during the examination the doctor is provided with an alert if there is evidence for a potential interaction between the drugs that the patient is currently receiving or between a drug that the clinician enters into the system with a drug that they are already being prescribed.

4.2.3.4 Prototypical scenario

1. The current CRIS solution provides a pseudonymised snapshot version of the EHR for research with GATE capability
2. CRIS contains certain reusable components. e.g. binary to text conversion (XHTML), OCR etc.
3. These components will be extended to provide an interface with additional MIMIR indexing for a near real time view of the patient record for clinical use.

4.2.3.5 Requirements on components

GATE text processing

- Medication apps

Mimir

- Indexing of medication

Knowledge base

- Provide information about relations between medication interactions from the literature

4.2.4 Common use case functionality

As well as the system estimating these entities from the text, the end users will also be able to enter the information themselves.

The end user will be able to find relevant guidance on a given issue above (by manually entering the conditions in an EHR field [or in an external field that populates the EHR]).

The end user will be able to correct any information obtained from NLP that they believe to be wrong. Any corrections will be made available to the system, allowing active learning to take place.

An end user will be able to enter information in an external field not linked to the EHR in order to carry out hypothetical querying (e.g. to find out in general what antidepressant to prescribe in people with renal failure).

4.2.5 API Requirements

The component communication should follow the guidelines defined by GATE WASP.

4.2.6 Ethical requirement

The ethical requirements are regulated in SLAM, through their Clinical Record Interactive Search (CRIS) application [2].

5 Conclusion

According to the use cases, the identified users of KConnect see the search engine functioning as a complement and assistance to all major decision making processes, such as prevention, diagnosing, treatment and follow-up. KConnect could be of great help in the administrative burden and therefore give the care provider extra time to care for the patient. KConnect could function as a reliable source for spread of skills and new knowledge and provide assurance for new professionals or professionals in training.

The KConnect application could impact the healthcare sector by contributing to more time efficient patient visits; ensuring the care provider with more time with the patient instead of spending it on administration; secure knowledge spread and confidence in decision making.

6 References

[1] Patientdatalag (2008 :355) http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/Patientdatalag-2008355_sfs-2008-355/

[2] Fernandes, Andrea C., et al. "Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records." BMC medical informatics and decision making 13.1 (2013): 71.