Experience made using public Cloud infrastructure to analyse clinical patient data

Matthias IHLE[1], Steffen CLAUS, Juliane FLUCK, Zakir KHAN, Pooya MOHEBBI,
Philipp SENGER, Wolfgang ZIEGLER[2], Lena GRIEBEL, Martin SEDLMAYR[3]
Florian BERGER, Julian LAUFER[4], Astros CHATZIASTROS, Johannes DREPPER[5]
[1]*Averbis GmbH, Tennenbacher Straße 11, Freiburg, D-79106, Germany*
Tel: +49 761 203 97690, Fax: +49 761 203 97694,
Email: matthias.ihle@averbis.com
[2]*Fraunhofer Institute SCAI, Schloss Birlinghoven, Sankt Augustin, D-53754, Germany*
Tel: +49 2241 142500, Fax: +49 2241 142460,
Email: {steffen.claus, juliane.fluck, muhammad.zakir.khan, pooya.mohebbi, philipp.senger, wolfgang.ziegler}@scai.fraunhofer.de
[3]*Friedrich-Alexander-University Erlangen-Nuremberg, Krankenhausstr. 12*
Erlangen, D-91054, Germany
Tel: +49 9131 85 – 26720, Fax: +49 9131 85 – 26754
Email: {Lena.Griebl, Martin.Sedlmayr}@imi.med.uni-erlangen.de
[4]*RHÖN-KLINIKUM AG, Am Schlossplatz 1, Bad Neustadt/Saale, D-97616, Germany*
Tel: +49 9771-65-1525, Fax: +49 9771-65-9114
Email: {florian.berger, Julian.Laufer}@rhoen-klinikum-ag.com
[5]*TMF, Charlottenstrasse 42, Berlin, , D-10117, Germany*
Tel: +49 30 220024745, Fax: +49 30 220024799
Email: {astros.chatziastros, johannes.drepper}@tmf-ev.de

Abstract: Patient data describing operations and results of treatment in clinics is stored in clinical information systems as part of the clinical process. Since these documents are unstructured free-texts stored as scanned documents or as documents prepared with a word processing system there was no automated way to find patterns in these documents that indicate significant accumulations of e.g., treatments and side effects. As a result, valuable information hidden in this huge amount of data spread across clinics is just neglected. A solution is the automated processing of unstructured data from different sources with advanced text mining technology. However, processing a large amount of scanned documents in general exceeds the computational power available in clinics. Using Cloud resources on a pay per use basis is a cost-efficient alternative to accomplish this task. We describe the approach of the cloud4health project, its framework for processing anonymized patient data, and its data protection and security developments to turn a Cloud into a trusted Cloud where the data may be processed in compliance with legal requirements[*].

1. Introduction

Patient data describing operations and results of treatment in clinics is stored in clinical information systems as part of the clinical process. These reports are usually unstructured free text documents written by the attending physician, stored, probably partly delivered to the physician, that continues the treatment of the patient once he leaves the hospital. Due to the fact that these documents are unstructured free-texts stored as scanned documents or as documents prepared with a word processing system there was no automated way to find

[*] Part of this work was co-funded by the Federal Ministry of Economics and Technology

patterns in these documents that indicate significant accumulations of e.g., treatments and side effects. While this might be handled manually for diseases with small number of cases in one clinic it is beyond of being feasible doing this across multiple clinics. As a result, valuable information hidden in this huge amount of data spread across clinics is just neglected; information that could help improving treatment and reducing undesired side effects otherwise.

A solution is the automated processing of unstructured data from different sources, e.g. word processor, scans, with advanced text mining technology. However, processing a large amount of scanned documents using Optical Character Recognition (OCR) technology and retrieving relevant information from the resulting documents or from word processer documents in general exceeds the computational power available in clinics. Increasing the computing infrastructure is expensive both with respect to the necessary hardware and the staff to operate it given that in contrast to the other clinical systems these would not be permanently used. The spread of Cloud computing bears a cost-saving approach to process and analyse the patient data without requiring local hardware resources and staff. The German cloud4health project funded by the Federal Ministry of Economics and Technology is aiming to deliver a solution both in terms of software, Cloud infrastructure, text mining workflows and the definition and installation of processes that comply with the legal requirements concerning processing of patient data and data protection.

2. Objectives

The objective of this paper is presenting the experiences made when developing a trusted Cloud infrastructure and the trustworthy processes to analyse the patient data using this Cloud infrastructure. It turned out that the provisioning of Cloud resources and deploying the text mining into this infrastructure is feasible using current Open Source technology stacks for the Cloud management. Also, existing frameworks like UIMA are well suited to implement the text mining workflows and deploy them.

In contrast, the real problems have been (1) making the Cloud secure, so it becomes a resource where patient data processing is allowed by the data protection officers; and (2) making the patient data ready for processing in a public Cloudinfrastructure. We will focus on (2) and describe briefly the processes required in the clinics to anonymize the data, the gateway between clinic and public Cloud and the legal requirements regarding processing of these data in the Cloud.

It should be noted that the clinical patient data used for analysis are not taken from patients actually under treatment in the clinics but from patients whose clinical treatments already have been completed in the past. To this extent, the benefit for patients is rather in future treatments taking into account findings of the analysis. Due to the legally required anonymisation of the patient data (as described in section 5.2) it would be impossible to identify an individual patient and change its on-going treatment anyway. Using pseudonymisation of the patients' data instead of anonymisation would allow identifying patients in the results of the analysis. However, this approach would require a private Cloud in the clinics to ensure that data stays in the the clinics for analysis.

3. Methodology

After selecting the patient data for a certain question de-identification of the documents anonymize the patient data is a major effort. Providing an automated process for doing this which is accepted by the data protection officers is difficult as the identity of the patient probably can reconstructed from information that are not evidently recognisable for the program as being related to the identity. Not only information on the patient's identity has to be supressed but also information on the attending physicians, the clinic, etc. Once this anonymization has been achieved the data is stored in a dedicated database for transferring the documents into the Cloud for analysis.

The member of the clinic staff authorised analysing the data triggers the startup of the Cloud environment, the virtual machines for the text mining. The number of VMs is depending on the problem size determined by the number of documents, size of documents and type of documents. This allows in production mode to adapt the size of the infrastructure to the needs of the clinic, e.g. fast results at higher costs. The envisaged business model is a pay-per-use approach.

Once the VMs are up UIMA pipelines for analysing the data are deployed into these VMs and started. There may be different instances of pipelines; the suitable one is selected based on the use-case, e.g. which question will be processed based on which type of documents. The number of parallel working pipelines again is depending on problem size.

While the startup of the environment is a synchronous process the following transfer of documents to the UIMA broker is done asynchronously to allow full exploitation of the parallelism of the distributed UIMA pipelines. Also, since processing times vary depending on site and complexity of the documents asynchronous processing allows a constant stream of documents from the transfer database through the secure gateway to the Cloud.

Regarding the data protection it turned out that in Germany with its federal structure no general applicable blueprint for process security can be used which is accepted automatically by the data protection officer in other federal states once in use in one state. Moreover, regarding patient data the data protection officer of the attending clinic is responsible in first place. Thus, two independent levels of data protection officers are in the game: clinic and state government and additional the federal government laws on data protection. As a consequence, negotiating and adapting processes and underlying technology is of utmost importance to get the mandatory consent of the data protection officers for processing patient data. The following section provides a more detailed discussion of the data protection measures taken in the project.

3.1 Data protection

The sensitivity of health information is generally recognized, and health data of individuals associated with hospital treatment is protected in Germany by a complex set of laws, rules, and regulations. First and foremost is the duty for medical confidentiality as regulated by the respective professional law. For hospitals and clinics specific so-called hospital laws apply. While they differ between the federal states, they also regulate occasionally the use and transmission of personal data. Last but not least, non-official bodies have to adhere to German data protection law. Therein a distinction is made between ordinary personal data and special categories of personal data, which are by definition information on racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, health or sex life.

Accordingly, patient data is unequivocally regarded as a special category of personal data. For those categories, German data protection law provides additional cover by means of special regulations for data collection, processing, use, and transfer. For instance, the processing of health data for purposes other than health care is regularly permitted only if informed consent has been obtained from the individual patient, even when the respective data has been pseudonymized beforehand.

Since our use cases depend on retrospective evaluation of clinical data and given the immanent difficulties in obtaining subsequent informed consent from discharged patients, the use of pseudonymized data was not a viable approach in this project. Instead we focus on patient data that has been previously anonymized. Data protection laws are irrelevant in cases exploiting completely anonymized patient data, i.e. personal data that has been altered so that the information concerning personal or material circumstances cannot be attributed to an identified or identifiable natural person.

We make a major effort to effectively anonymize the patient data that has been selected for a particular research question. In a process called deidentification, the identifying text and figures are automatically annotated and subsequently deleted or shifted (see section 5.2 for details). The results are then stored within the hospital in a dedicated transfer database. Due to the sensitive nature of this anonymization stage, the deidentification process can only be triggered by authorized personnel in the hospital. Only after this anonymization procedure, the selected data set will be transferred outside the hospital to be further processed in the Cloud.

After anonymization the data records still contain sensitive medical information. In principle, unique combinations of medical information might be used together with other sources of information to detect single individuals. Therefore, a thorough security concept has been developed and implemented to prevent any possibility of a subsequent reidentification. The security concept consists of a register of processing operations and a definition of protection requirements based on the evaluation of the vulnerability of the different components. Among others, it demands encrypted transmission to the cloud-based text mining services and a processing of limited duration without permanent data storage outside the hospital. Furthermore, after anonymization of the data records, the original patient IDs are the same for all records related to the same patient but do no longer allow identifying the individual patient the original IDs. Before sending the data to the Cloud the IDs of the anonymized data records are replaced by random temporal IDs impeding a future tracking and combining of individual data sets of the same patient, outside the hospital during processing in the Cloud (see also section 4.1). It should be noted that the existence of a key or rule within the hospital, which would enable text-mined data to be mapped to individual patients, would render the data as being no longer anonymous, according to predominant view of data protection officials in Germany.

Finally, all aspects of data security are summarized in a dedicated concept of data privacy that has been approved by the local data protection officials. This document serves also as a reference upon future re-examination and represents an essential building-block of the quality management of the project.

4. Technology Description

The technology used for the implementation of the cloud4health solution for text mining on clinical patient data in a trusted Cloud covers an Open Source Cloud middleware together with Java-based web services. In this section we focus on the major challenge when processing data in the Cloud: security and protection identity and data of the patients.

4.1 Security Requirements for Cloud Infrastructures arising from patient data processing

The compute- and memory-intensive processing of patient data is handled by a central text mining Cloud infrastructure. For this purpose, the Fraunhofer institute SCAI provides a Cloud testbed, which serves as a community Cloud for the project partners for developing, testing and researching the text mining services. Beyond these activities, the testbed shall be considered as a best practice implementation for processing patient data in a secure and trustful manner. Given the sensitive nature of personal patient data (which in exceptional cases might even persist despite data anonymisation, see 2.1), high security requirements regarding their confidentiality and integrity have to be met. Besides the German/Federal Data Protection Acts, certain clinic-internal regulations come into play. Within the cloud4health project, clinics are the main data providers for document processing in the Cloud. Hence, technical and organisational security measures not only have to follow the common regulations but also the clinics' specifics for data processing in external infrastructures.

During the first year of the cloud4health project, main requirements for patient data processing in an external Cloud infrastructure have been gathered. They can be outlined as follows: first, a detailed risk analysis has to be performed. This analysis should specify and examine diverse worst-case scenarios regarding breaches of data confidentiality/integrity and has to group the infrastructure components into different categories related to their protection needs and security requirements. Based on this classification, technical security measures should be implemented accordingly. Second, the concrete purpose of data processing not only has to be explicitly defined but also followed appropriately. Thus, security measures have to guarantee that the anonymized personal patient data cannot be used for any other purpose than the predetermined one. Therefore, data must not be accessible by other processes or other users of the Cloud; multi-tenancy and separation of data processing has to be ensured on all levels during the complete lifecycle of data handling (from transfer to processing, storage and deletion). Third, data has to be deleted after a reasonable or an agreed upon period of time. If possible, personal data should not be stored in the Cloud infrastructure at all. Fourth, an incident response process has to be implemented, preferably integrated in an overarching security management concept. Finally, the user of the Cloud infrastructure - in case of the cloud4health project this role is taken by the clinical data provider - has to be able to check and verify the technical and organisational security measures, e.g. through a certification executed by trusted third party experts. Even though there are standards for self-assessment and self-certification, an independent verification of security measures is always preferred.

Summing up, strictly following best-practice guidelines and common personal data processing regulations has not proved to be sufficient within the cloud4health project. In most cases, cloud-specifics haven't been covered in these guidelines yet. As well, federal, national and clinic-specific requirements on personal data processing have to be harmonised and addressed appropriately.

4.2 Secure Gateway from the Clinic to the Cloud

For the communication with Cloud and text mining infrastructure a web-service has been implemented as secure gateway between clinics, trusted Cloud and UIMA text mining. This web-service also replaces the patient id in the original document by a temporal random id before sending the document to the text mining. Details are described in section 5.4.

To encrypt the data during the transport to the Cloud a VPN tunnel is established between the respective interfaces in clinic and Cloud. The entire system is multitenant and guaranties the separation of multiples clinics using text mining services at the same time.

For authentication the mechanism already established in the clinics is used, currently based on username, password like credentials. Based on the authentication the authorisation for accessing services in the Cloud is decided. However, the authentication interface is generic and allows use of certificates issued by a trusted CA as well as attribute-based or role-based authorisation.

5. Developments

5.1 The cloud4health Architecture

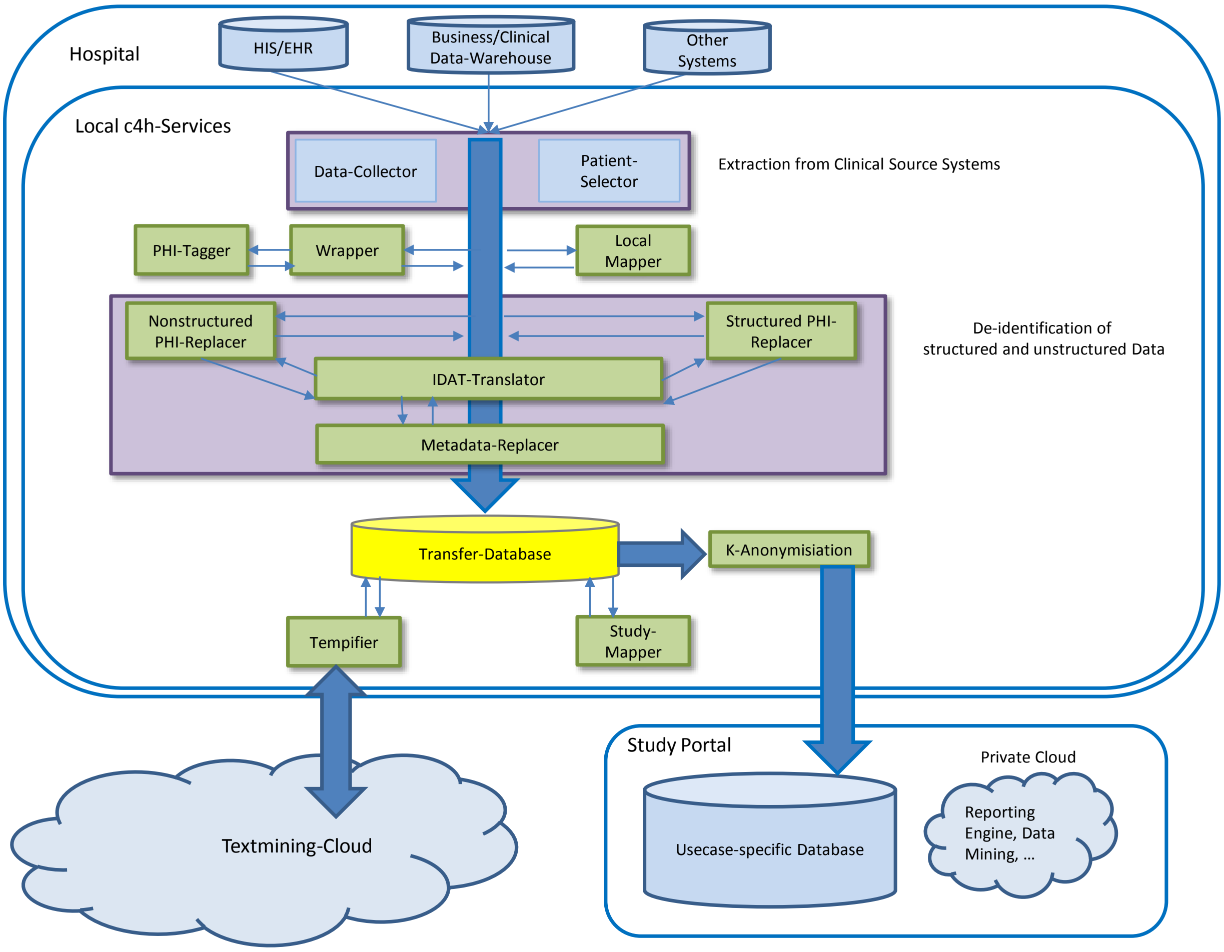


Figure 1: The cloud4health architecture and clinic bus

The architecture realizes a structural framework for cloud4health, consisting of three major parts:

- *Local services* access and de-identify structured and unstructured data from the source systems at each provider's location.
- Within a *text mining Cloud* the information from free text is extracted and returned in a structured format.
- Data from different providers are aggregated in a *central study portal* for further analysis.

The flow of information from the source to the central study portal is as follows:

- Patients to be included into a study are identified in the source systems by the *patient-selector* according to the inclusion criteria (e.g. ICD, age, sex). Then, all required data elements are extracted by the *data-collector*.
- For unstructured text, all identifying attributes (e.g. names, dates, addresses = personal health information (PHI)) in texts are marked with type and place of occurrence (*PHI-tagger*) with the wrapper caching already processed text. The quality of de-identification can be controlled by a human who is also responsible for approving the results.
- For structured data the *local mapper* maps the data onto terminologies (e.g. laboratory values to Logical Observation Identifiers Names and Codes (LOINC)).

- Identifying data can occur in structured and unstructured data as well as in the metadata of the communication packets. This is why all three types of data are considered for de-identification (anonymization or pseudonymization) by dedicated *replacer* components. To ensure uniform replacements, the replacer components contact a central rule engine (*IDAT-translator*).
- After the extraction and de-identification of all data, a first snapshot is generated in the transfer database.

From the transfer database, all unstructured text is fetched for further processing in the text-mining cloud. To increase safety, the *tempifier* generates temporary IDs for all documents to be processed before sending them to an external cloud. The result of the text-mining is returned in a structured way (Operational Data Model ODM-format [8]) and stored in the transfer database.

Before the final export to the central study portal, an additional mapping onto study specific terminologies can be made. Additionally, k-anonymization [13] is important to ensure privacy in aggregated data.

5.2 Anonymization of Patient Data

The central process from the data protection perspective is the anonymization of free text information in health records. Anonymization is thereby defined as modification of personal data in a way that details of personal circumstances can no longer or only with disproportionate effort be attributed to an identified or identifiable natural person.

Since there are no general criteria for such identifying properties in Germany, we have surveyed personally-identifying information occuring in medical documents at German hospitals and have thereby identified 9 characteristic categories of information to be marked during the de-identifying process: name, date, location, contact, division, id, age, biometrics, other). Since it is possible but not always desired to simply delete all attributes – for example it may be required to preserve the intervals between dates by shifting date specifications – the de-identification process has a separate component for replacing marked passages according to the study protocol (the PHI replacer in Figure 1).

To support this de-identification process we have developed the web-based *deid* tool that simplifies and accelerates the annotation and assignment into categories of identifying attributes in textual documents comprehending the PHI tagger, wrapper, and PHI replacer components of Figure 1. Replacement of identified attributes is handled by the IDAT translator web service that is called during document export for all identified attributes and that translates all passages according to the study specific rules.

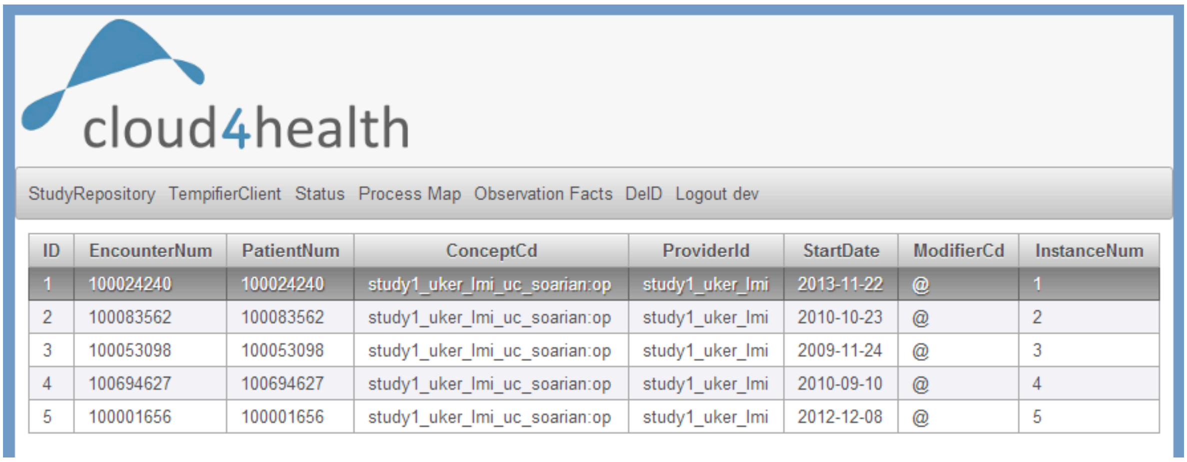
The automatic recognition of personal-identifying data like names, addresses, dates, etc. is a non-trivial task since the de-identification software uses different methods. At first the document is matched with information found in the structured document meta-data as well as in dictionaries for names, locations, etc. Then, pattern recognition methods are used to identify attributes following certain patterns like dates or email addresses. In a last step, sophisticated machine learning methods are used for identifying attributes missed by previous methods. Here, manual corrections are analysed and used to train a model that learns to replicate typical human annotations.

To support the de-identification of large corpora of health records the *deid* tool offers the possibility to either automatically annotate documents in accordance with local data protection officers or to pre-annotate documents that later on will be reviewed by humans in a semi-automatic modus.

5.3 Transfer Database (cloud4health clinical portal)

The transfer database consists of a JDBC compatible database (e.g. MySQL [1] or Oracle [2]), which is managed by a J2EE [3] Web Application running inside of JBoss AS7 [4].

This management interface is used by clinical users and its purpose is to guide them through the different steps of a study and the usage workflow of the different cloud4health services. It implements a strict role based authorization to differ between administration tasks and tasks where original and not anonymized patient data is displayed. This helps to improve data protection, because even clinical administrators do not have direct access to the patient data while they can do configuration on the system. On the other hand can a clinical user, which has authorization to see and work with the patient data, access the original data and work on anonymization of the data.



ID	EncounterNum	PatientNum	ConceptCd	ProviderId	StartDate	ModifierCd	InstanceNum
1	100024240	100024240	study1_uker_Imi_uc_soarian:op	study1_uker_Imi	2013-11-22	@	1
2	100083562	100083562	study1_uker_Imi_uc_soarian:op	study1_uker_Imi	2010-10-23	@	2
3	100053098	100053098	study1_uker_Imi_uc_soarian:op	study1_uker_Imi	2009-11-24	@	3
4	100694627	100694627	study1_uker_Imi_uc_soarian:op	study1_uker_Imi	2010-09-10	@	4
5	100001656	100001656	study1_uker_Imi_uc_soarian:op	study1_uker_Imi	2012-12-08	@	5

Figure 2: cloud4health clinical portal

The database schema (which has its origin in I2B2 [5]) can store the documents either anonymous or liked with the real patient id. The latter purpose is to use additional meta-data (accessible with I2B2 or any other data-warehouse system) and compare Cloudresults with already existing data for evaluation or later analysis.

The transfer database is filled with anonymized documents together with text mining relevant information such as study, faculty-, department- and clinical information system of origin.

When all documents for one study are anonymized and ready for the cloud, an authorized user can start the process. The connection to the Cloud is a secure and encrypted Virtual Private Network (VPN) connection through OpenVPN [6]. If the OpenVPN connection is not available, the Web Application starts the OpenVPN client and connects to the cloud. Then the Tempifier Web Service is called to initialize the Cloudrequest. When the Cloudis ready to accept documents, all documents are sent to the Tempifier Web Service, which is done multi-threaded. Each request is synchronous, so the answer, which contains the ODM result, can be saved directly to the database. After all documents are processed, a final Tempifier Web Service to shut down the Cloudis called.

When all results are in the transfer database, an additional job takes all individual ODM [8] results of all processed documents and creates one big ODM, which contains the study meta data along with the clinical data results from the cloud. This ODM is then processed by a Talend Open Studio Job [7] to import the results into the I2B2 database schema.

5.4 Secure Gateway to the Cloud (Tempifier)

The secure gateway has two main responsibilities: (1) it should disburden the clinician from coping with details of the text mining and Cloud infrastructure, and (2) replacing the anonymized patient ID in documents before transferring them to the text mining pipelines in the Cloud by a random temporal ID and later the previously replaced ID in the ODM document returned as result of the text mining.

To achieve (1) the interface to the transfer database includes authentication and description of the types and number of documents to be processed. With this information the Tempifier is able to trigger the JInitiator component, which in turn starts the Cloud environment by connecting to the Cloud management system to start-up the Virtual Machines (VMs) with the required UIMA components and the specific pipelines for processing the text mining on the type of documents as announced by the Tempifier.

When the complete environment is up and running the Tempifier registers the pipelines at the Broker API, which is responsible to pass the documents to the text mining pipelines. As a result of the registration the end point references of the pipelines are returned to the Tempifier.

Once the initialisation process is complete the transfer database starts sending documents to the Tempifier in multiple threads. To accomplish (2) the Tempifier replaces in each document the ID by a random temporal ID and sends the modified document to iTextmining for further distribution. Once the results of the text mining for a document are available in form of an ODM document this document is returned to the Tempifier, which replaced the temporal ID in the ODM document by the previously replaced ID and returns the data back to the transfer database. By using a random temporal ID in the Cloud it is not possible to recombine multiple anonymized documents that belong to the same patient. However, when that ID is restored in the transfer database all ODM documents that belong to the same patient can be stored and analysed together. It should be noted that the ID does not allow identifying the patient because this ID is a result of the anonymization process. If an agreement with the data protection officers can be reached in the course of the project, it would be desirable to pseudonymize the documents instead as patients could benefit from results of the text mining if these results could be mapped to a patient.

When all documents are processed the transfer database indicates the end of the study and the Tempifier continues (1) by requesting the shutdown of the Cloud infrastructure for this specific use case.

5.5 Broker API and Text Mining Pipelines

The cloud4health text mining Cloud provides an environment to process large amounts of clinical data on demand for different use cases. To keep the running costs to a minimum and furthermore ensure that no textual content with possibly sensitive data remains in the Cloudafter processing the VMs are started on demand and shut down after processing by the Tempifier, the clinic-side interface to the cloud.

Currently, the Tempifier still has to estimate the load in before and start a sufficient number of VMs for processing a request. In future releases, an admin service will be responsible for automatic adapting the number of active VMs to the current load.

The cloud4health text mining Cloudconsists of two types of virtual machines (VMs), pipeline VMs processing textual documents in text mining pipelines and broker VMs distributing received input documents among the pipelines.

A broker VM provides the SOAP based iTextmining interface in a tomcat web service for processing text documents in a pipeline. Therefore, the document is passed to an ActiveMQ [10] message broker that routes documents to registered text-processing pipelines according to the specified endpoint name. The result is transferred into the ODM format and returned to the Tempifier. The ODM standard ensures compatibility to clinical systems and allows for easy integration into the clinical context [8]. The text mining infrastructure is based on the UIMA-AS [11] framework from the Apache Software Foundation [12], currently the standard framework for development of text and data mining applications.

A pipeline VM provides the SOAP based iPipeline interface with that text mining pipelines may be deployed, configured and registered to the broker with an unique endpoint name. Such a pipeline consists of several analysis units, each of them responsible for a specific task like i.e. finding the synonyms of a terminology. All units are executed consecutively and the analysis results of all previous units are at disposal during execution. A pipeline receives, processes and returns documents in form of UIMA CAS objects, which provide an interface for text mining tasks like annotating or accessing the type system. CAS objects are propagated through all analysis units and contain the results and annotations of each analysis unit. Each VM may deploy several text mining pipelines in parallel for one or several endpoints. Pipelines with identic endpoint name must also share the same configuration.

The text mining pipelines consists of a number of natural language processing tools. We do not describe the whole pipeline in detail but give a short overview of the main information extraction steps. Essentially two subsequent processing steps are integrated in each pipeline: the annotation of entities, a process called named entity recognition (NER) and the identification of relationships between two or more entities. These two steps are discussed in more detail below.

5.5.1 Named entity recognition

NER units are implemented to find the relevant terminology for a certain use case. Different clinical terminology resources such as ICD or OPS terminology are used and adapted for the specific use case. Additional primarily use-case specific and manual developed and curated terminologies are developed to cover the most relevant aspects of the surveyed domain. In currently implemented pipelines the annotation of entities was done using the NER system ProMiner [9] or the ConceptMapper developed by the Averbis GmbH. These methods can reliably identify the biomedical entities of different terminologies like diseases, drugs, doses, and human anatomy. This can be done by regular expression and/or by using specific terminologies.

5.5.2 Relation extraction

The subsequent identification of relevant and use-case specific relationships between the identified entities is a subsequent step of the pipeline. One simple example is to find a drug, a dosage, the administration duration, and the relation between those entities.

A mixture of rule, business logic, and machine learning based approaches developed by the Averbis GmbH and Fraunhofer SCAI are used to build analysis engines (AE) for relation extraction. The output of the pipeline is a collection of ODM objects representing the found entities and their relations with each other. Furthermore, the relation extracting components can attach a confidence value to each identified relationship. This value helps the integration of different annotations for the same relation.

5.5.3 Confidence based integration strategy of multiple relations

By using the assigned confidence values of each relationship, it is possible to define a simple and efficient integration strategy to create a single relation as an output. This is needed when one relationship was identified by more than one rule or one pipeline. One approach is to use a simple majority-voting schema with at least two different states:

1. If the majority of relations belong to the same class, then this class is chosen.
2. If there is no majority, the summed confidence values of each class acts as a decision value. By doing this, the relation class with the highest confidence values are chosen.

This ensures that, if there is no majority, the relations with the highest confidence values are used to generate a unique output of the complete pipeline.

5.6 Implementation problems experienced

In the clinic several existing and new components has to be plugged together along a clinical bus. The anonymiszation tool was has been developed from scratch to make sure it meets the requirements of the clinical data protection officers. The interaction between the Tempifier and the Cloud infrastructure for automatically starting up the necessary Cloud infrastructure and the required Virtual Machines is complex since it is depending on a number of factors: the amount of documents, the type of documents, the clinic from which the documents originate as well as the department within the clinic and the focus of the study.

While in the initial prototype described in this paper these dependencies were handled on a per case bases manually, for the productive version of the next phase we are preparing a mapping tool that allows the Tempifier to automatically determine the right environment for a n-tupel of factors.

In the next phase we will also work on a robust solution to encrypt and sign the documents exchanged between Tempifier and iTextmining. i.e. documents received from the transfer database and result files received from iTextmining. The problem to be solved here is the lack of communication channels leaving the Cloud since all communication of components in the Cloud must be initiated from the clinic. Thus, a CA for providing the necessary certificates cannot be reached from components in the Cloud. For security reasons the CA cannot be operated in the Cloud either

6. Results

The overall result of the cloud4health project achieved in its first year is the prototype implementation of a process for analysing clinical patient data in the Cloud. The results affect both information processing in hospitals and trustworthy processing of anonymized patient data in a Cloud.

In detail, a number of results have been achieved in the two areas the project is focussing on: data selection and customisation in clinics and analysing these data with text mining methodologies in the Cloud.

In the clinics a software framework has been implemented that supports the entire process from patient data selection to secure transfer of the anonymized patient data to the text mining processes in the Cloud. The text mining results are stored in a database for further analysis. Also, starting up the Cloud environment and deploying the necessary text mining VMs is part of this process.

The UIMA text mining pipelines have been set-up for the first use cases with patient data to be analysed. Virtual machines for dynamical deployment in the Cloud infrastructure have been created containing the resulting UIMA components.

7. Business Benefits

The project is focussing on four use-cases:

- hip joint endoprostheses: looking into the possibility to support the evolving German endoprostheses register, performing retrospective studies on the effect of different implants and technologies used for implanting
- pharmacovigilance: identification of so far most often unknown undesirable side-effects of medicaments, substantiation or invalidation of suspicion regarding side-effects of medicaments through analysis of patient data
- plausibility: plausibility check of treatments with pharmaceuticals
- pathology: facilitating of multicentre studies (including international ones), support for the conversion of unstructured pathology information systems into structured pathology information systems, fully documented probes as an important economic factor

Business benefits resulting from the outcome of the project can be identified for several target groups:

Clinics can improve treatment of their patients and their processes by identifying patterns in their data that could not easily be found otherwise due to the nature of data and data processing in clinics. Moreover, clinics get access to high performance computing for analysing their data without being forced to buy and operate such infrastructure at a high price. Instead, the clinics can access these resources following a pay-per-use model.

Cloud providers that dedicate part of their resources to become a trusted Cloud may attract new customers that until now refrained from using external Cloud resources for reasons of security and data protection.

Pharmaceutical companies can get a more systematic insight into so far unknown side effects of medicaments.

Finally, a new **commercial service** offering will emerge on the basis of the results of the project providing trustworthy infrastructure and processes for analysis of medical data according to the requirements of their customers.

-

8. Conclusions

During the first 12 months the cloud4health project achieved to implement the mechanisms to dynamically start the required Cloud infrastructure upon a case study request of a clinic and deploy the text mining pipelines into this Cloud. At the same time the project is in contact with the data protection officers of the clinics and the state government about the process to get a mandatory agreement for making the anonymized data available for processing in the trusted cloud.

It turned out that the provisioning of Cloud resources and deploying the text mining into this infrastructure is feasible using current Open Source technology stacks for the Cloud management. Also, existing frameworks like UIMA are well suited to implement the text mining workflows and deploy them.

In contrast, the real problems have been (1) making the Cloud secure, so it becomes a resource where patient data processing is allowed by the data protection officers; and (2) making the patient data ready for processing in a public Cloudinfrastructure. We will focus on (2) and describe briefly the processes required in the clinics to anonymize the data, the gateway between clinic and public Cloud and the legal requirements regarding processing of these data in the Cloud. Successful set-up of processes and infrastructure for processing patient data requires a close and often iterative cooperation with the data protection officers of the clinics. Otherwise, even the anonymized the data may not leave the clinics to be analysed in a Cloud infrastructure.

During the second year of the project the technology for end-to-end encryption of anonymized patient data will be implemented. As a result patient data will only be unencrypted in the memory of the computing resources during processing by the UIMA pipeline. Equally, the results of the analysis are encrypted in the Cloud immediately after the results have been produced by the UIMA pipeline. The results are double protected against fraud: by the encryption in the Cloud and while they are sent through the VPN tunnel from the Cloud to the clinics. As a result, the IPR of the clinic on the results of the text mining is protected both in the Cloud and during transport.

Towards the end of the project members of the project consortium plan to establish and offer a commercial service to analyse clinical patient data for e.g. clinics, manufacturers of medical devices, health insurance companies and pharmaceutical companies.

References

[1] MySQL open source database. Website. Online at http://www.mysql.com/, visited 22 April 2013.

[2] Oracle Database. Website. Online at http://www.oracle.com/index.html, visited 22 April 2013.

[3] Java Platform, Enterprise Edition (Java EE). Website. Online at http://www.oracle.com/technetwork/java/javaee/, visited 22 April 2013.

[4] JBoss – open source Java application server. Website. Online at http://www.jboss.org/jbossas, visited 22 March 2013.

[5] i2b2 - Informatics for Integrating Biology and the Bedside. Website.Online at https://www.i2b2.org/, visited 22 March 2013.

[6] OpenVPN – Virtual Private Network technology. Website. Available at http://openvpn.net/, visited 22 March 2013.

[7] Talend – data integration products. Website. Online at http://de.talend.com/products/data-integration, visited 22 March 2013.

[8] CDISC operational data model. Website. Online at http://www.cdisc.org/odm, visited 22 April 2013.

[9] Hanisch, D., Fundel, K., Mevissen, H. T., Zimmer R., and Fluck, J. ProMiner: Rule based protein and gene entity recognition. BMC Bioinformatics, 6 (Suppl 1):S14, 2005

[10] The ActiveMQ messaging andintegration pattern server. Website. Online at http://activemq.apache.org/, visited 22.April 2013.

[11] Apache UIMA project. Website. Online at http://uima.apache.org/, visited 22 April 2013.

[12] Apache Software Foundation. Website. Online at http://www.apache.org/, visited 22 April 2013.

[13] Aris Gkoulalas-Divanis and Grigorios Loukides. 2012. Anonymization of Electronic Medical Records to Support Clinical Analysis. Springer Series:Briefs in Electrical and Computer Engineering, Springer Publishing Company.