Information Extraction from Unstructured Electronic Health Records and Integration into a Data Warehouse

Georg Fette, Maximilian Ertl, Anja Wörner, Peter Kluegl, Stefan Störk, Frank Puppe

Comprehensive Heart Failure Center, 97074 Würzburg
georg.fette@uni-wuerzburg.de
max.ertl@smi.uni-wuerzburg.de

University of Würzburg Computer Science VI
97074 Würzburg
{puppe, pkluegl}@uni-wuerzburg.de

University Hospital Würzburg
Med. Klinik u. Poliklinik I
97074 Würzburg
{stoerk_s, woerner_a1}@klinik.uni-wuerzburg.de

Abstract: For epidemiological research, the usage of standard electronic health records may be regarded a convenient way to obtain large amounts of medical data. Unfortunately, large parts of clinical reports are in written text form and cannot be used for statistical evaluations without appropriate preprocessing. This functionality is one of the main tasks in medical language processing. Here we present an approach to extract information from medical texts and a workflow to integrate this information into a clinical data warehouse. Our technique for information extraction is based on Conditional Random Fields and keyword matching with terminology-based disambiguation. Furthermore, we present an application of our data warehouse in a clinical study.

1 Introduction

Large parts of epidemiological research is based on statistical evaluation of medical data. For these evaluations the source data needs to be available in a structured form (e.g., in an attribute-value representation [DN07]), as statistical evaluation tools (e.g. SPSS) rely on this data structure. A common approach to acquire this structured medical data is to run study-specific test series on a carefully selected subgroup of patients. This process is time- and cost-intensive. Other possible data sources for medical studies are standard clinical reports (laboratory reports, discharge summaries, echocardiographic reports, etc.), which are generated within clinical routine. Archives filled with these reports have to be examined manually, and the desired information has to extracted and processed by, e.g., medical students, which is also an immensely time consuming process. It would be desirable to perform the analysis of those clinical records automatically or at least semi-automatically. Unfortunately, the data structure of this routine data is heterogeneous and ranges from fully structured to completely unstructured plain texts, which is an undesired characteristic for automatic data analysis. Documents from the domain of laboratory reports, for example, consist of attributes from a closed set of attribute types and their respective measurements, and are already in the desired structured form. Echocardiographic reports in our hospital consist of predefined building blocks which

are assembled by the report writers. These semi-automatically generated texts have to be re-transformed into the original information from which the reports were generated. Finally, anamnesis (past and current medical history) and physical examination reports are plain texts, usually without any framing, and therefore very difficult to analyze.
When looking at these examples it becomes evident that it is not trivial to develop a best practice procedure to integrate all medical routine data into one homogeneous system. Every text corpus from any domain has to be processed individually and the best type of information extraction tools and techniques has to be chosen and adopted with respect to the domain's data characteristics.
The goal of our ongoing project is to create a *data warehouse* (DW) in which most of the clinical routine data is integrated in a homogenous way with an easy to use interface. A clinical DW can be used for different purposes:

- The search results can enrich the data basis of an already existing medical study by adding standard routine data from patients which are observed in that study. [DM09]
- The search results can be the sole basis of a medical study in which a medical hypothesis is checked. The hypothesis can be tested on the basis of existing data from patients other than those from the study's original test group. [Bl96]
- The search results can determine sets of patients which appear suitable for being part of a new cohort of study patients, based on selection criteria which include previous reports from these patients. [Ka05b]
- The data from the DW can be explored with data mining techniques to create completely new medical hypotheses which have not yet been discovered because of their complex nature (e.g. a certain symptom only arises by the combined occurrence of multiple, seemingly independent causes). [At09]

The structure of this paper is as follows: In section 2 the underlying DW architecture is explained along with the necessary ETL (extraction, transformation, load) workflow in which the source data for the DW is provided and processed. In section 3 the information extraction process from the text domains is discussed. Section 4 presents the application of the DW in a real life medical study. In section 5 our DW approach is compared to similar systems. Finally, section 6 provides a discussion and conclusion of the presented work.

2 Data Warehouse

2.1 Workflow Architecture

The DW is a system on top of the conventional clinical information system (CIS). The general data model for the DW is depicted in figure 1. Physicians store information that is generated during their daily routine work in the CIS. This data covers a large variety of topics from various domains and can be of arbitrary structure depending on the CIS used. As the data structure of the CIS cannot be altered and as it is impossible to process the data within the CIS itself, the data has to be exported and transformed into an indexable format. All necessary raw data is extracted from the CIS, pseudonymized and saved in an exchange directory. Pseudonymization is done for patient identifiers (PIDs),

case IDs or document IDs, which appear anywhere in the documents. Furthermore, all patient names and addresses are removed. All further processing steps are performed in the exchange directory until the data is integrated into the DW. Medical researchers, as users of the DW, will finally be able to browse the DW via basic database query statements or a dedicated query language developed for the DW.

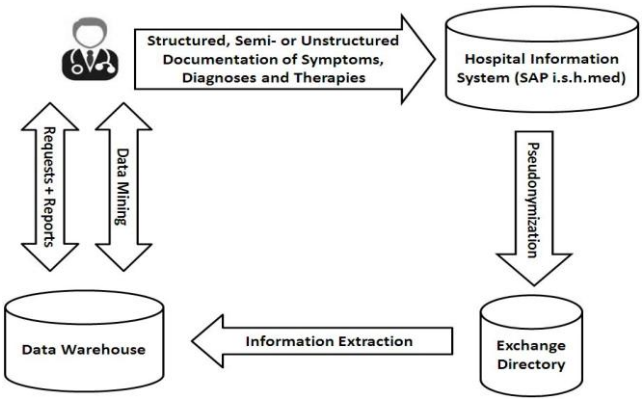


Figure 1: Data warehouse system work flow

2.2 Data Warehouse Model

The DW is realized as an SQL database. Because a detailed description of the database's architecture is beyond the scope of this paper we will only describe the basic data model. The relationships between the data model elements are shown in figure 2. There exist patients with their respective patient IDs (PID) and entry and discharge time-stamp. For every case there exist several report documents. Usually, the reports stem from different domains, but there can be more than one report from the same domain for one case. All documents have a creation time-stamp and contain their case-specific text content if the document is a text document (e.g. laboratory documents are already available in an attribute value form and have no additional text). The documents are linked with their containing information bits which are stored with their intrinsic information value and their corresponding terminology ID (the explanation of the terminology is covered in section 3.3). The access to the DW is addressed in section 3.7.

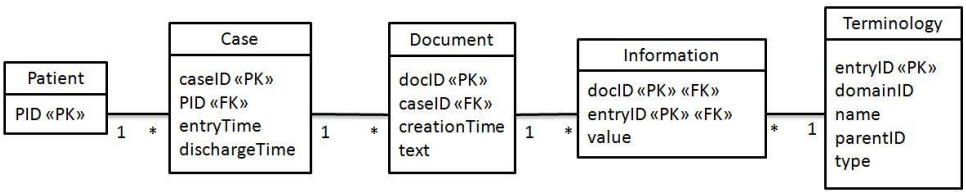


Figure 2: DW as object relational mapping in UML [Ba05] (PK: primary key, FK: foreign key)

3 Data Integration

In this chapter we describe how the medical routine data is integrated into the DW. We first describe the data's general characteristics in section 3.1. Section 3.2. covers general aspects of information extraction related to our problem domains and section 3.3 some specific aspects concerning the information types we want to extract. In section 3.4 we describe the creation of a training corpus for a supervised learning approach and in section 3.5. the application of the supervised learning method on not yet labeled documents. In section 3.6 the automatically labeled data is imported into the DW and in section 3.7. we deal with the possibilities to browse the fully processed data in the DW.

3.1 Data

The whole body of clinical data is subdivided into report *domains*. Each domain represents a certain type of examination which can be carried out on a patient and for which a document is created. In our project, we integrate the different domains one by one into the DW. We do this because all documents from one domain share a common structure and a common terminology, a fact which can be exploited in both the *information extraction* (IE) process itself as well as in domain-specific quality testing phases afterwards.
The different domains we will integrate into the DW originate from different sources and therefore exhibit different document structures. The domains can be divided into *structured* and *unstructured* domains. The data from structured domains is already stored in relational databases or resides in xml-documents, which both can be integrated into the DW quite easily with relatively simple mapping rules.
The unstructured domains consist of text documents. Some of these domains are still quite structured because the texts were semi-automatically created with a predefined text block system from which the physicians have to choose the desired elements to create the final texts. Some of the domains, however, like anamnesis or physical examination reports, are completely unguided free texts; hence, the only common ground are their common domain topics.
A special type of an unstructured domain is the *discharge summary*. Each discharge summary is an aggregation of all important case-related reports. Some of the therein contained sub-domains do not have their own document type in the CIS, so the text blocks from the discharge summaries are those domains' sole source of data. These text blocks can be extracted from the discharge summaries and individually analyzed. This extraction mode is beyond the scope of this report but the approach we used is discussed in [Kl09a].

This is the current list of domains we have identified for our project:

- disease-related group (DRG) diagnoses
- diagnoses from discharge letters
- laboratory reports
- electrocardiogram reports
- 24-hour Holter ECG reports
- 24-hour blood pressure reports
- bicyle stress test reports

- ergospirometry reports
- 6-minute walk test reports
- lung function reports
- coronary angiography reports
- transthoracic echocardiography reports
- X-ray reports
- endoscopy reports
- sonography reports
- medical treatment reports
- magnet resonance tomography reports
- therapy reports
- anamnesis and physical examination reports

This list is not yet complete but the listed domains were identified as containing important medical information with a large number of available reports. Thus, the benefit of a successful automatic processing pipeline is maximized. For an exhaustive survey on the whole amount of raw data which is available at the University Hospital of Würzburg we refer to [Er11]. In figure 3 some of the domains we already processed are listed with the respective number of documents in the CIS (years 2000 - 2011).

domain name	no.of documents
echocardiography	40,143
ECG	118,078
lung function	14,859
X-ray thorax	9,764
bicycle stress test	9,820

Figure 3: Document counts of selected unstructured domains

3.2 Information Extraction (IE)

With the document sets for each domain at hand we can start extracting information from them. When performing automatic IE from texts the first decision to be made is whether the extraction process should be supervised or unsupervised. In unsupervised IE the algorithm's outcome is defined only by the processed documents themselves with only few possibilities to influence the algorithm´s outcome. In supervised IE it is easier to define the desired output format by training the system with a hand labeled (or differently created) reference standard. The automatic labeling algorithm learns from the provided training data and uses the acquired knowledge to label the remaining documents the way the training data was labeled. For this technique it is necessary that the documents from the training corpus and the remaining documents share the same structure. As in our project all documents from one report domain fulfill this property of homogeneous structure we decided to use the supervised approach. (Although different physicians have individual writing styles, we claim that the overall content of reports from one domain remains the same, so that a supervised learning method will perform well).

3.3 Terminology

Before IE can be performed on an unstructured domain an information type system has to be created. In many IE tasks the type of information to be extracted from texts is a fixed set of types like *Persons*, *Organizations* or *Locations*. In our case the information types are not known in advance, so we first have to define a closed set of information types which exist in the domain. In the following, this set of types will be called *terminology*. The terminology serves in the IE project as a key element. Not only does it provide a definition of types which reside in the domain's texts, it also serves as a basis for structuring the informative content of the domain when the different participants of the project are working together. It incorporates theoretical information on the domain, represents the structure of the domain's report texts and contains technical information used by algorithms during the IE process.
We tried to incorporate existing terminologies like UMLS [Bo04], MeSH [Ro63] or Snomed [CN08] but the gap between the existing terminology's structure and the terminology we needed was too large to efficiently integrate any of the available resources into our desired terminologies.
We decided to organize each terminology as a tree structure with additional type information on each node. A terminology is defined as a set T of nodes $t_i = (id_i, name_i, par_i, type_i)$ with an integer valued node- and parent-ID, a node name, and an information type which can be either *numeric*, *boolean*, *negation indicator*, *single choice*, *single choice value*, or *structure*.

The nodes of a terminology should serve the following purposes:

- information representation: each information node represents a particular specific information. As a consequence, each information type should occur not more than once in each document. It would represent an undefined state if an information would appear with two different values in the same document; e.g., in an ECG report, there should be only one measured heart rate. If there is the same information under two different circumstances (e.g. heart rate before and after physical exertion) there should be two different nodes in the terminology for those two distinct information types.
- structure: the terminology is organized as a tree structure (each node is tagged with its respective parent node in the tree). This way the terminology can group topics, which provides a better readability for human readers. Furthermore, the tree structure can serve as an informative hint for automatic algorithms, as nodes which are closer in the terminology are semantically related.

When multiple domain experts are separately asked to create a terminology for a specific domain from scratch, their results often show marked differences. Furthermore, the ideal terminology's structure should stick to the structure of the domain's texts as its main purpose should be the usage for IE from those texts. Consequently, the terminology creation should be undertaken focussing on the characteristics of the domain's text corpus in a corporate process between all members of the IE project.
Taking all those considerations into account we see that the construction of a proper terminology is a delicate, difficult and particularly time consuming process. Because of

resource restrictions in our project (the time of domain experts is an expensive resource) we tried to incorporate cheaper but unexperienced workers in the overall process as much as possible. As a functioning workflow we developed the following work cycle:

All texts from a particular domain are tokenized and POS-tagged (part of speech, i.e. grammatical word class information). The words are filtered for nouns, adverbs and adjectives and the remaining words counted and sorted by word count. The top 500 are given to a medical expert who organizes the words in a raw terminology tree. This raw product usually consists of about 50 to 100 nodes and takes the domain expert about one to two hours to create. The raw terminology is given to an inexperienced worker who refines it to its final state by analyzing report texts and inserting missing nodes into the tree where it is necessary. Now the refined terminology consists of about 150 to 600 nodes. This refinement process takes about one to two days per domain. For the correction of possible mistakes of the inexperienced worker during the refinement the terminology is again re-checked by a domain expert and corrected when needed. This correction step takes about two to four hours per domain. To improve the terminology's utility for automatic processing the IE engineer enriches the nodes with type information. In figure 4 an exemplary excerpt from the ECG domain can be seen.

entryID	name german	name english	parentID
1	Herzfrequenz	Heartrate	0
2	Sinusrhythmus	Sinus rhythm	1
3	Art	Type	1
4	bradykard	Bradycardia	3
5	tachykard	Tachycardia	3
6	Wert in Schläge/Minute	Beats per minute	1
7	Typen	Types	0
8	Linkstyp	Left type	7
9	Rechtstyp	Right type	7
10	Indifferenztyp	Indifferent type	0
11	Rückbildungsstörungen	Repolarisation disorder	0
12	nein	no	11
..	..	..	..

Figure 4: Excerpt from terminology of ECG domain

3.4 Training Corpus

When the final terminology is ready, the creation of the training corpus starts. First, we define some formal definitions for the documents and their containing annotations:

A domain consists of a set of documents $D = \{d_1, .., d_n\}$. We call a document $d_i = (text_i, \emptyset)$ without any included annotation information an *unlabeled document*, one with its extracted information $d_i = (text_i, A_i)$ with $A_i = \{a_{i1}, .., a_{ij}\}$ a *labeled document*. The information bits are described by tuples $a_{ij} = (t_{ij}, text_{ij}, offset_{ij})$ which describe the annotated text excerpt from the document's text, its start and end offsets in the text and the corresponding terminology node the annotation is connected with. We take a subset of the documents D_t and manually add the annotation labels to create a training corpus

for an automatic labeling algorithm. In figure 5 an exemplary report from the ECG domain and its corresponding annotations list can be seen.

Indifferenztyp, Sinusrhythmus (92/min.), AV-Block ersten Grades, Linkshypertrophie, regelrechte R-Progression, keine Erregungs-rückbildungsstörungen.

Indifferent type, sinus rhythm (92 bpm), first degree AV block, left heart hypertrophy, regular R-wave progression, no repolarisation disorder

id	covered text	offset
10	Indifferenztyp	0-14
2	Sinusrhythmus	16-29
6	92	32-33
14	AV-Block	40-48
16	ersten Grades	49-62
..	..	..

Figure 5: Example of an ECG report and an excerpt of its annotation set

As a rule of thumb, we took about 100 documents per domain into the training corpus. For the annotation task we advised a student worker to use TextMarker [Kl09b], an comfortable and easy to use annotation editor. The creation of the training corpus by manual annotation took about one to three weeks per domain depending on the average text length of the document texts and the amount of possible labels available in the domain (i.e., number of nodes in the terminology). In figure 6 we show the number of annotated documents in the training corpora for our selected domains and the number of nodes in the respective terminologies (*information* nodes are nodes which are no *structure* nodes).

domain name	no. of documents in training corpus	nodes in terminology	information nodes in terminology
echocardiography	87	649	388
ECG	100	236	69
lung function	84	168	83
X-ray thorax	172	357	158
bicycle stress test	101	273	111

Figure 6: Terminology size and training corpus size for selected domains

3.5. Automatic Labeling

By using the training corpus an arbitrary supervised learning algorithm can be trained and used for automatic labeling. For our algorithm of choice we chose *Conditional Random Fields* (CRF) [LMP01] because it is a state of the art labeling algorithm with freely available and easy to use implementation packages (MALLET [Mc02]). When processing domains with larger document sizes we encountered the problem that the training of the CRF got slower with increasing numbers of information nodes in the terminology (i.e., the time for reaching a satisfactory low average error on the training corpus itself increased). This fact is due to the relationship that the training time of a CRF increases quadratically with the number of state transitions, which are in our case represented by the distinct number of annotated information labels in the texts. The increase in training time prevented us from undertaking the five-fold test for the echocardiography domain, because a single-fold already took about one week of training time. We overcame this drawback by implementing an alternative labeling algorithm on the basis of keyword matching and terminology-based disambiguation [Fe11]. We

performed a five-fold test on the five domains listed in table 1 with both learning algorithms. Figure 7 shows the result table of those tests in training times, precision, recall and f1-scores.

	CRF				keyword matching		
domain name	training time	precision	recall	f1	precision	recall	f1
echocardiography	-	-	-	-	0.99	0.99	0.99
ECG	12 min.	0.98	0.91	0.95	0.98	0.84	0.90
lung function	48 min.	0.94	0.70	0.80	0.83	0.85	0.84
X-ray thorax	5h 44 min.	0.97	0.79	0.87	0.91	0.86	0.88
bicycle stress test	1h 43 min.	0.88	0.70	0,78	0.89	0.72	0.80

Figure 7: Results from information extraction on unstructured domains (5-folds)

As can be seen in the table, the f1 scores for different domains differ remarkably. We explain these differences with the different degrees of semantic and linguistic complexity by which information is embedded in the domain's texts. As the domain of echocardiography consists of semi-automatically created texts our approach shows very good results. The other domains consist of free texts with higher intrinsic syntactic and semantic complexity. We hope to be able to improve the score on these domains by increasing the amount of manualy annotated document or applying further linguistic processing tools.
It is not possible to assign a preferred algorithm as both approaches show superior results depending on the processed domain. It can only be stated that the keyword matching algorithm does not need any training phase and therefore can be applied much better when the CRF's training time gets too long or constant re-training is necessary (because new training documents are created). Thus, a combination of both algorithms can be considered: keyword matching for determining the quality of the current training corpus and CRF for eventually better final results.

3.6 Data Import

The labeled documents have to undergo some final postprocessing steps until they are finally transferred into the DW relational databases. To illustrate the post-processing steps the following rules are exemplified in figure 8, which is the resulting final value table for the example from figure 5. The post-processing rules depend on the type of the terminology nodes by which the information bits are linked:

- Labels belonging to numerical terminology nodes get their covered text stripped from any non-numerical character. Therefore, the value for the node "Wert in Schläge/Minute" ("Beats per minute") with the entryID 6 is 92.
- Labels with single choice values are stored with the ID of their parent's single choice ID and with the value of their choice ID. Thus, the choice "Indifferenztyp" ("Indifferent type") which has the entryID 10 is stored as the value for the choice node "Typen" ("Types") with entryID 7.

- Boolean entries are saved with the value 1. Thus the boolean node "Sinusrhythmus" ("Sinus rythm") which has the entryID 2 is stored with the value 1.
- If in the same document with an annotation for a boolean node exists an annotation for the corresponding negation node, the stored value for the boolean node is 0 instead of 1. Therefore the stored value for the boolean node "Rückbildungsstörungen" ("Repolarisation disorder") with the entryID 11 is 0 because of the existence of the label "nein" ("no").

entryID	value	docID
7	10	1
2	1	1
6	92	1
11	0	1
..	..	..

Figure 8: Excerpt of the information table for the document from figure 5.

3.7 Data Access

The DW can be accessed in multiple ways. The simplest way is by simple SQL queries on the tables of interest. For example, a user may select all information from the documents of a set of case IDs which belong to a certain domain.
Another access mode is the use of a query language that was developed specifically for the DW. For the sake of brevity we will describe only some selected features:
It is possible to restrict the output to a predefined set of desired patient-IDs which are contained in the output set. The query results can be limited to a list of domains, each with a restricted set of terminology IDs. For every desired patient-ID all desired information elements from the different domains are returned. It is possible to nest the domains in the query, so that the documents from the nested domain have a temporal dependency to the documents of the surrounding domain. It is thus possible to ask for, e.g., the laboratory reports of a group of patients and additionally for ECG reports which were taken in a window of one week around the date of particular laboratory reports.

4 Application onto a clinical research question

The study we want to present is a substudy from [An11]. The aim is not to show the clinical results itself but rather describe, in principle, how the DW can be used to enhance the workflow in a standard clinical study. In the mentioned substudy we want to estimate the impact of C3 complement, a measurement which is on request included in the laboratory blood measurements of patients, in relation to a certain set of cardiovascular symptoms. The investigated symptoms are all part of the standard echocardiographic checks and appear in the text blocks of the echo documents. The desired echo attributes are: LVDs, IVSd, LVPWd, LVEF, LADs, Ao-root, AV-Vmax, TR-Vmax, MV-E, MV-A, MV-E/A, IVRT, DT (MV-E), M-Ring-E´, E/E´, sPAP and if

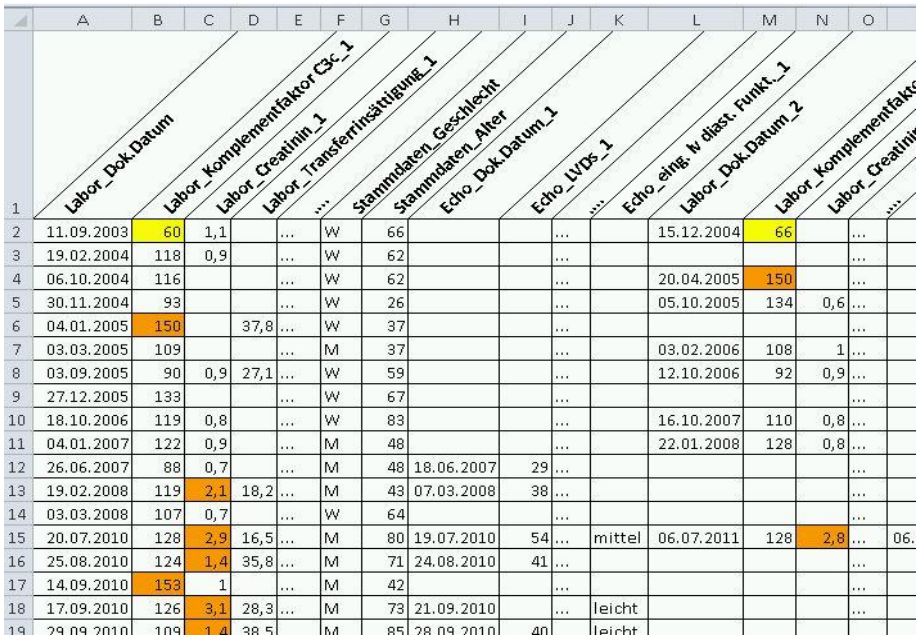
the patient suffered from a limited left ventricular systolic or diastolic pump function. The query requested to the DW was to provide all measurements of patients which had undergone echocardiography and who also had their C3 complement measured in a two-month time window around the echo measurement. From these patients we additionally requested another laboratory report in a time window of at least six and up to eighteen months after their first measurement again containing a C3 complement value. Around this second laboratory measurement we were again interested in echocardiographic examinations in a two-month time window around the second laboratory report. These follow-up reports (combined laboratory and echocardiographic report) were again requested for the time slots of at least 18 to at most 30, at least 30 to at most 42 and at least 42 to at most 54 months.

The configuration of the complete DW query request took about five minutes and can be seen in a condensed form in figure 9. The actual computation took another five minutes until the desired report sheet was immediately ready for further statistical processing. A screenshot of the final excel sheet can be seen in figure 10.

Without the DW the desired data had to be picked from the CIS by an assistant researcher, who had to look up all patients possibly considered for the study (find all patients with a C3 complement measurement) and individually match the corresponding echocardiographic reports from all those patient, read them, and manually write the results into a table. Such work easily may consummate several months of work and is more error-prone than the automatic report aggregation. Although the tested f1-value on the echo domain of 0.99 promised a high degree of correctness of the automatically labeled echo documents, the provided values from the study were manually checked for validity, thereby confirming our previous test results.

```
<Anfrage rowIDType="PID">
    <Domain name="Labor">
        <Attribute name="Komplementfaktor C3c" needed="GLOBAL"/>
        <Attribute name="Creatinin" />
            ...
        <Domain name="Stammdaten">
            <Attribute name="Geschlecht" />
            <Attribute name="Alter" />
        </Domain>
        <Domain name="Echo" minMonth="-2" maxMonth="2">
            <Attribute name="LVDs" />
            ...
            <Attribute name="eingeschr. lv syst. Funktion" />
        </Domain>
        <Domain name="Labor" minMonth="6" maxMonth="18" id=”Labor2”>
            <Attribute name="Komplementfaktor C3c" needed=”Labor2”/>
            <Attribute name="Creatinin" />
            <Domain name="Echo" minMonth="-2" maxMonth="2">
                <Attribute name="LVDs" />
                ...
            </Domain>
        </Domain>
        <Domain name="Labor" minDay="1" minMonth="18" maxMonth="30">...</Domain>
        <Domain name="Labor" minDay="1" minMonth="30" maxMonth="42">...</Domain>
        <Domain name="Labor" minDay="1" minMonth="42" maxMonth="54">...</Domain>
    </Domain>
</Anfrage>
```

Figure 9: DW query for a clinical study



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	Labor_Dok.Datum	Labor_Komplementfaktor C3c_1	Labor_Creatinin_1	Labor_Transferrinsättigung_1	....	Stammdaten_Geschlecht	Stammdaten_Alter	Echo_Dok.Datum_1	Echo_LVDs_1	....	Echo_eing. lv diast. Funkt._1	Labor_Dok.Datum_2	Labor_Komplementfakto	Labor_Creatinin	....	
2	11.09.2003	60	1,1		...	W	66			...		15.12.2004	66		...	
3	19.02.2004	118	0,9		...	W	62			...					...	
4	06.10.2004	116			...	W	62			...		20.04.2005	150		...	
5	30.11.2004	93			...	W	26			...		05.10.2005	134	0,6	...	
6	04.01.2005	150		37,8	...	W	37			...					...	
7	03.03.2005	109			...	M	37			...		03.02.2006	108	1	...	
8	03.09.2005	90	0,9	27,1	...	W	59			...		12.10.2006	92	0,9	...	
9	27.12.2005	133			...	W	67			...					...	
10	18.10.2006	119	0,8		...	W	83			...		16.10.2007	110	0,8	...	
11	04.01.2007	122	0,9		...	M	48			...		22.01.2008	128	0,8	...	
12	26.06.2007	88	0,7		...	M	48	18.06.2007	29	...					...	
13	19.02.2008	119	2,1	18,2	...	M	43	07.03.2008	38	...					...	
14	03.03.2008	107	0,7		...	W	64			...					...	
15	20.07.2010	128	2,9	16,5	...	M	80	19.07.2010	54	...	mittel	06.07.2011	128	2,8	...	06.
16	25.08.2010	124	1,4	35,8	...	M	71	24.08.2010	41	...					...	
17	14.09.2010	153	1		...	M	42			...					...	
18	17.09.2010	126	3,1	28,3	...	M	73	21.09.2010		...	leicht				...	
19	29.09.2010	109	1,4	38,5		M	85	28.09.2010	40		leicht					

Figure 10: query result sheet for a clinical study

5 Related Work

Our DW system can be compared to related systems in two ways. First, we can discuss our system's pure DW aspect. Second, we can discuss the aspect of IE from clinical data. The idea of using a DW for accessing clinical data is not new, but there is no comprehensive and easily applicable out-of-the-box solution. Clinical information systems, like any other enterprise information system, are diverse and individual. Even if based on a standard solution there has so to be done much customizing to fit the existing clinical processes and documentation, so that there is no standardization yet to be usable for different hospitals. In order to construct an operational and effective DW it is essential to combine process work, domain expertise and high quality database design [LSH08]. This seems to correspond to concurring literature, where we found either a holistic approach with theoretical conceptualization but lacking practical implementation [Ka05a], [Du90], [Ku07]) or practical implementation focused on one single domain, i.e. lacking generalizability of such concepts [Br09].

The idea of using IE for clinical data has also already been discussed for decades. The comparison of different medical IE systems is a difficult task because neither the input nor the output data are available because of patient privacy protection reasons. We therefore remain with the description of the system's structural differences:

Most medical IE approaches handle only one specific domain like [CM05], [DP07], [Ma03], [Ju07] [Ra01], [BK02], [Zh07] for the domains of echocardiography, ECG,

radiology, nursing management, disease management, pharmaceutics or drug events. Or they try to combine one domain with another like in [RD08] where the domain of radiology was linked with the domain of pathology.
Most approaches used a commonly accessible terminology standard like UMLS or RadLex as their vocabulary basis for IE. This often leads to a quite poor recall performance because the processed texts share a different vocabulary than the used terminologies.
Only few systems aim at creating an integrated DW in combination with cross domains or domain independent IE. One of these systems is MetaMap (http://metamap.nlm.nih.gov/). As MetaMap's terminology is only based on UMLS it also suffers from poor recall for terms which do not exists in this standard. Systems with an adaptable terminology system are cTAKES [Sa10] and MedLee [Fr00]. MedLee features a high adaptability to new domains. There, as in our approach, different domains were made accessible one by one. In the advanced stages of the MedLee project there was a large variety of clinical domains in the system which could be accessed homogeneously. In contrast to our approach, they integrated the terminology of the different domains in one big terminology, which they enriched with every further domain. As MedLee is not freely available we were unable to further compare it with our system.
A system with a similar type of query language is WAMIS [Do02] that allows posing of queries with a temporal logic structure.
A comprehensive overview on IE and data warehousing in the clinical field can be found at [Me08] and [PG09].

6 Conclusion

We developed a structured approach for the homogeneous integration of different data domains used in clinical routine into a DW. We described the general architecture of the system and the work flow for extracting information from unstructured text domains and their integration into the DW. The IE from the domain of echocardiography reports already shows satisfying f1 score results, thus, the extracted information can already be reliably integrated in clinical studies or applied to clinical research questions. The other domains show promising results but still have to be improved to use them for clinical studies with the same degree of reliability. We compared two IE methods: CRF and keyword matching with terminology-based disambiguation. We evaluated both methods on a selected set of text domains and yielded very encouraging results.
In our future work we will analyze and integrate additional yet unprocessed text domains and improve the performance of our approach on the domains we already worked on.

This work was supported by grants from the Bundesministerium für Bildung und Forschung (BMBF01 EO1004).

References

[An11] CE Angermann, S. Störk, G. Gelbrich, H. Faller, R. Jahns, S. Frantz , M. Loeffler, G. Ertl: Mode of Action and Effects of Standardized Collaborative Disease Management on Mortality and Morbidity in Patients with Systolic Heart Failure: The Interdisciplinary Network for Heart Failure (INH) Study. Competence Network Heart Failure. Circ Heart Fail. 2012 Jan 1;5(1):25-35. Epub 2011 Sep 28. PMID: 21956192

[At09] Atzmueller, M.; Beer, S.; Puppe, F. A Data Warehouse-Based Approach for Quality Management, Evaluation and Analysis of Intelligent Systems using Subgroup Mining: Proc. 22nd International Florida Artificial Intelligence Research Society Conference (FLAIRS), pp. 372-377, AAAI Press, 2009

[Ba05] H. Balzert: Lehrbuch der Objektmodellierung, Analyse und Entwurf mit der UML 2, 2005

[Bl96] Black N.: Why we need observational studies to evaluate the effectiveness of health care. BMJ1996; 312:1215-8.

[BK02] JS Barrett, SP Koprowski Jr.: The epiphany of data warehousing technologies in the pharmaceutical industry. Int J Clin Pharmacol Ther 2002; 40 (3): S3–13

[Bo04] O. Bodenreider: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, Vol. 32, No. suppl 1. (01 January 2004), pp. D267-D270.

[Br09] CJ Brooks, JW Stephen, DE Price, DV Ford, RA Lyons, SL Prior, SC Bain: Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care. Prim Care Diabetes. 2009 Nov;3(4):245-8. Epub 2009 Jul 14.

[CM05] J. Chung, S Murphy: Conceptvalue pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports. AMIA Annu Symp Proc, 2005.

[CN08] R. Cornet, K. Nicolette: Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making 8: S2. PMID 19007439

[DP07] J. C. Denny, J. F. Peterson: Identifying qt prolongation from ecg impressions using natural language processing and negation detection. In MedInfo, 1283–1288, 2007

[Do02] W. Dorda, W. Gall, G. Duftschmid: Clinical data retrieval: 25 years of temporal query management at the University of Vienna Medical School.InfMed 2002; 41: 89–97.

[DN07] V. Dinu, P. Nadkarni: Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. Int Journal of Medical Informatics 76 (11–12)

[DM09] M. Dugas, S. Amler, M. Lange, J. Gerß, B. Breil, W. Köpcke: Estimation of Patient Accrual Rates in Clinical Trials Based on Routine Data from Hospital Information Systems. Methods Inf Med. 2009;48(3):263-6.

[Du09] M. Dugas, B. Breil, V. Thiemann, J. Lechtenbörger, G. Vossen: Single Source Informationssysteme - Nutzung von Routinedaten für die klinische Forschung. 54. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. Düsseldorf: German Medical Science GMS Publishing House

[Er11] M. Ertl: Erfassung von klinischen Untersuchungsdaten und Transfer in ein Data Warehouse. master thesis, Würzburg 2011.

[Fe11] G. Fette, P. Kluegl, M. Ertl, S. Stoerk, F. Puppe: Information Extraction from Echocardiography Records. In Proc. LWA 2011

[Fr00] C. Friedman: A broad-coverage natural language processing system. In AMIA, pages 270–274, 2000.

[Ju07] K. Junttila, R. Meretoja, A. Seppälä, E Tolppanen, T. Nikkola, L. Silvennoinen: Data warehouse approach to nursing management. J Nurs Manag 2007; 15 (2): 155–161.
[Ka05a] C. Katzer, K. Weismüller, D. Brammen, R. Röhrig, G. Hempelman, T. Chakraborty: Data-Warehouse aus klinischen und genomischen Daten zur Entwicklung kombinatorischer Scoring-Systeme, 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds). Düsseldorf, Köln: German Medical Science; 2005.
[Ka05b] J. Kamal, K. Pasuparthi, P. Rogers, J. Buskirk, H. Mekhjian.: Using an information warehouse to screen patients for clinical trials: a prototype. AMIA Annu Symp Proc. 2005:1004.
[Kl09a] P. Kluegl, M. Atzmueller, and F. Puppe.: TextMarker: A tool for rule-based information extraction. In Proc. UIMA, 2009 Conference of the GSCL, 2009.
[Kl09b] P. Klügl, M. Atzmueller, P. Puppe: Meta-level Information Extraction.. In: Mertsching, B.; Hund, M. & Aziz, M. Z. (Hrsg.): KI. Springer, 2009 (Lecture Notes in Computer Science 5803)
[Ku07] R. Kush, L. Alschuler, R. Ruggeri, S. Cassells, N. Gupta, L. Bain, K. Claise, M. Shah, M. Nahm. 2007. Implementing Single Source: the STARBRITE proof-of-concept study J. Am. Med. Inform. Assoc. 14: 662-673.
[LMP01] J. Lafferty, A. McCallum, F. Pereira: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. , Proc. 18th International Conf. on Machine Learning 2001
[LSH08] JA Lyman, K. Scully, JH Harrison Jr.: The development of health care data warehouses to support data mining. Clin Lab Med. 2008 Mar;28(1):55-71, vi.
[Ma03] B. Mamlin, D. Heinze, C.McDonald: Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc, 2003
[Mc02] A. McCallum: MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. 2002.
[Me08] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle: Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. IMIA Yearbook 2008: Access to Health Information, 2008:128–144
[PG09] HU Prokosch, T. Ganslandt: Perspectives for medical informatics - reusing the electronic medical record for clinical research. Methods of Information in Medicine 48/1(2009): 38-44
[Ra01] Ramick DC: Data warehousing in disease management programs. J Healthc Inf Manag. 2001; 15 (2): 99–105
[RD08] DL Rubin, TS Desser: A data warehouse for integrating radiologic and pathologic data. J Am Coll Radiol 2008; 5 (3): 210-217
[Ro63] FB Rogers : Medical subject headings. Bull Med Libr Assoc 51: 114–116. ISSN 0025-7338. PMC 197951. PMID 13982385
[Sa10] G. Savova, J. Masanz, P. Ogren, J. Zheng, S. Sohn, K. Kipper-Schuler, and C. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. JAMIA, 17(5):507–513, 2010
[Zh07] Q. Zhang, Y. Matsumura, T. Teratani, S. Yoshimoto, T. Mineno, K. Nakagawa, M. Nagahama, S. Kuwata, H. Takeda: The application of an institutional clinical data warehouse to the assessment of adverse drug reactions (ADRs). Evaluation of aminoglycoside and cephalosporin associated nephrotoxicity. Methods Inf Med 2007; 46 (5): 516–522