# Ontology-Based Semantic Annotation of Documents in the Context of Patient Identification for Clinical Trials

Peter Geibel[1], Martin Trautwein[4], Hebun Erdur[2], Lothar Zimmermann[1], Stefan Krüger[1], Josef Schepers[1], Kati Jegzentis[3], Frank Müller[1], Christian Hans Nolte[2], Anne Becker[4], Markus Frick[4], Jochen Setz[4], Jan Friedrich Scheitz[2], Serdar Tütüncü[2], Tatiana Usnich[2], Alfred Holzgreve[4], Thorsten Schaaf[1], and Thomas Tolxdorff[1]

[1] Institut of Medical Informatics, Charité - Universitätsmedizin Berlin
[2] Department of Neurology (CBF), Charité - Universitätsmedizin Berlin
[3] Center for Stroke Research Berlin (CSB), Charité - Universitätsmedizin Berlin
[4] Vivantes - Netzwerk für Gesundheit GmbH
{peter.geibel,thomas.tolxdorff}@charite.de, martin.trautwein@vivantes.de

**Abstract.** In this paper, we describe the use of ontologies in the context of a system for identifying patients that are eligible for clinical trials. The main purpose of this clinical research data warehouse (CRDW) is to support patient recruitment based on routine data from the hospital's clinical information system (CIS). In contrast to most other systems for similar purposes, the CRDW also makes use of information present in clinical documents like admission reports, radiological findings and discharge letters. The linguistic analysis recognizes negated and coordinated phrases. It is supported by clinical domain ontologies that enable the identification of main terms and their properties, as well as semantic search with synonyms, hypernyms, and syntactic variants. The CRDW system is currently being tested at hospitals of the *Charité – Universitätsmedizin Berlin* and the *Vivantes – Netzwerk für Gesundheit GmbH*. In the paper, we provide an evaluation of the system based on real world data obtained from the daily routine work of the study assistants.

**Keywords:** Ontologies, RDFS, secondary use of health data, patient recruitment, clinical data warehouse.

## 1 Introduction

In recent years, the secondary use of clinical data has been considered an important topic of research since it enables medical progress based on data that are currently only used for treatment, administrative, and billing purposes. In this paper, we present research on a data warehouse system that is currently being developed in the context of a collaboration between *Charité – Universitätsmedizin Berlin*, the largest German university hospital, *Vivantes – Netzwerk für Gesundheit GmbH*, Germany's largest state-owned healthcare corporation, and a Berlin-based SME software partner.

R. Meersman et al. (Eds.): OTM 2013, LNCS 8185, pp. 719–736, 2013.

The main purpose of the clinical research data warehouse CRDW is to support patient identification for clinical trials based on routine data from the clinical information system (CIS). The CRDW allows finding patients that meet the inclusion and exclusion criteria of clinical trials. These criteria, for instance, correspond to patient data (age, sex), lab values, and coded information on diagnoses (ICD-10 codes) and procedures (OPS codes). Using also unstructured data, i.e., doctor's letters and other clinical documents, allows investigators to formulate more fine-grained criteria compared to using coded information alone. Furthermore, most medications are only documented in admission reports and discharge letters.

Other than our project, relevant efforts include I2B2 [13], EHR4CR (`http://www.ehr4cr.eu`), Cloud4Health (`http://www.cloud4health.de/`), and the KIS REK project [6]. Our system has a strong focus on combining computational linguistics and ontology-based information extraction.

In this paper, we would like to report our experiences in modeling and using ontologies [24], i.e., clinical knowledge bases, which are used by our software for extracting structured information from texts [15,4]. Also, the system is currently being tested at the Department of Neurology of the Charité and at the Clinic of Neurology – Stroke Unit – Center for Epilepsy (Vivantes Humboldt-Klinikum). We will describe qualitative and quantitative evaluation results in the context of these departments.

For the Clinic of Neurology (Charité), we present an evaluation that is based on real patient data and a series of real clinical trials. We compared the predictions of our system to the assessments of the trial team of the Center for Stroke Research Berlin (CSB) in order to obtain estimates of precision, recall, sensitivity, and other quantities. In particular, we will demonstrate that using information from unstructured data improves the performance of the system. This evaluation on real world data is the second main contribution of our paper.

This paper is structured as follows. After an overview of the system that is given in section 2, we describe the requirements of our specific ontology model (section 3) including a short overview of existing medical ontologies. Section 4 describes evaluation results for the pilot phases at the Clinic of Neurology (Charité) and the Clinic of Neurology (Vivantes). The conclusions can be found in section 5.

## 2 Overview of the System

The current deployment of the CRDW system at the Charité Berlin is based on data from the IS-H/i.s.h.med CIS modules (SAP/Siemens), which are integrated with patient information extracted from documents of the GE radiological system. The data is loaded into the data warehouse by an ETL (Extract, Transform, Load) process. In the case of Vivantes, a custom-built HL7 adapter provides selected data of neurological case records for an instance of the system.

All data is stored in a pseudonymous manner in the CDS (clinical data storage), integrating data from both structured and unstructured sources. This means

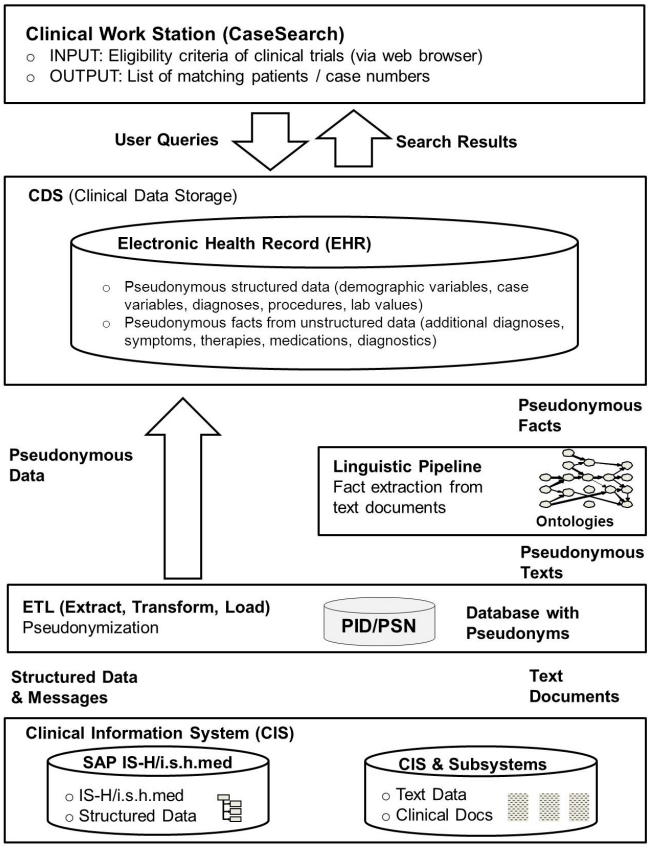


**Fig. 1.** Architecture of the CRDW System

that the patients' names, addresses, birth dates, etc. are replaced by aliases generated by the system in order to meet data protection regulations. In addition to the pseudonymization of structured data, our system is also able to remove identifying data in text, roughly following the American HIPAA privacy rule, cf. [25]. For removing personal information such as names and addresses, we combine using information from the CIS (e.g., name, address, date of birth) with computer linguistic methods, which, for instance, look for textual clues like "Herr" (Mr.), "Dr.", etc. in order to find occurrences of names in a document.

Within the system, the so-called *linguistic pipeline* is responsible the extraction of information from texts. The extraction is based on a linguistic analysis [12,8], which identifies terms, phrases, and sentences together with grammatical constructs such as negation ("not") and coordination ("and", "or"). This first, morphosyntactic part of the pipeline is based on the GATE framework [5]. For parsing, we use JAPE, the finite state transducer that comes with GATE.

The extraction of negated phrases is based on NegEX [3]. The POS-tagger (Apache OpenNLP) was trained on a tagged corpus of documents from both hospitals. In addition to the knowledge base, the pipeline also comprises resources that support spell-checking and the morphological analysis of words (lemmatization and compound resolution).

In the second part of the linguistic pipeline (so-called concept mapping), the identified terms and phrases are mapped to medical concepts in a semantic knowledge base, which contains etiological, morphological, topological and procedural information. This knowledge allows identifying the main pieces of information, assigning them to classes like "diagnosis", "therapy", "anatomical structure", and so on. For instance, in a phrase like "infarction of the MCA", "infarction" will be identified as a diagnosis, whereas the anatomical structure, "MCA", provides information about the location of the infarction.

Using the linguistic pipeline together with a knowledge base enables semantic search. For instance, the user can execute queries for synonymous terms like "stroke" and "cerebral infarction". The availability of taxonomical information allows searching with generalized queries like "infarction of a cerebral artery". This query then matches "infarction of the MCA" and variants hereof like "middle cerebral artery infarct", "media infarction", and even "The MRI depicts an infarct in the territory of the MCA". Note that a plain text search is bound to fail in these cases. Also, negated phrases do not count as hits in our system.

The software component that is responsible for finding eligible patients is called *CaseSearch*. It allows defining clinical trials together with their inclusion and exclusion criteria. These criteria are matched against both structured patient data and medical facts, which were extracted from documents by the linguistic pipeline. The *CaseSearch* provides web-based user interfaces that show current and past patients together with their eligibility for the defined trials. Translating the criteria of the study protocol into the language of the system is an important step of the overall process, since it has to take into account the availability and quality of data. It also requires knowledge of both the respective disease (and its treatment), and of processes at the respective clinic (patient paths, documentation, recruitment).

## 3 Constructing the Ontologies

In this section, we describe the requirements of our application, the chosen ontology model, and the knowledge engineering approach taken.

### 3.1 Analyzing Clinical Trials and Data Sources

We approached the problem of finding the right ontology formalism, structure and content from several directions. Since the main purpose of the CaseSearch is to find patients for clinical trials, we analyzed clinical trials in the field of *ischemic strokes*, in particular such that where conducted by the Clinic of Neurology of the Charité.

As an example, the clinical trial TRELAS [21,20] investigates Troponine T elevation in patients with ischemic stroke. In order to find patients meeting the criteria of this trial, we have to determine patients that suffer from a stroke or a non-ST elevation myocardial infarction (NSTEMI), and have a Troponine T level above $0.05\mu g/l$. Patients with a Creatinine value above $1.2mg/dl$ are excluced. Patients with a stroke can be found by looking at ICD-10 coded diagnoses in the CIS system, and by analyzing admission reports and radiological findings, which additionally allows determining the location of a stroke.

We analyzed stroke studies conducted at Charité with respect to their semantic structure and typical content. We were able to determine the following groups of criteria:

- *Main diagnosis* (IS-H/i.s.h.med)
- *Age and Sex* (IS-H/i.s.h.med)
- *Localization of infarction and other findings* (radiological report)
- *Time of identifying event* (admission report)
- *Severity:* in particular the NIH stroke score (admission report)
- *Symptoms* (admission report, discharge letter)
- *Lab data* (IS-H/i.s.h.med)
- *Prior medication, therapies* (admission report, previous cases)
- *Other diseases* (admission report, IS-H/i.s.h.med)
- *Medicolegal aspects:* pregnancy, ability to consent, risk factors, etc.

In a similar manner, our analysis of epilepsy trials conducted at Vivantes identified the following criteria as the most relevant:

- *Main diagnosis* (structured value, HL7)
- *Age and sex* (structured value, HL7)
- *Epilepsy type/type of seizures* (first-aid report, HL7/electronic document exchange)
- *Frequency of seizures and/or time of last seizure* (first-aid report)
- *Symptoms:* physical effects of seizures such as a bitten tongue, incontinence etc. (first-aid report)
- *Lab data (structured values)*
- *Prior medication and therapies* (first-aid report, HL7/electronic document exchange; admission report(s) of previous case(s), CIS)
- *Other diseases:* e.g., severe renal, hepatic, cardiologic or neurologic diseases (dto.)
- *Medicolegal aspects*

The analysis of the trials showed that in addition to features of the patient like age and sex, obviously diagnoses, symptoms, findings, and therapies are most important. When building the ontologies, we therefore focused on two main concept classes: *observations* (diagnoses, symptoms, and findings) and *therapies* (including medications). A concept in these classes can have properties like localizations (for which we introduced an *anatomy* class), and modifiers like *acute*, *chronic*, *left*, *right*, *frontal*, *parietal*, etc. These attributes for observations and therapies were collected in an *attribute* class.

Aside from the "has-attribute" relation, the criteria of trials rarely require semantic relations *between*, for instance, diagnoses and therapies. As an example, the doctor's letter might mention that a medication was prescribed in order to treat a specific disease. However, the criteria of trials rarely ask for this relationship. They usually just specify the diagnosis and the medication without detailing the semantic relationship between them. This means that this relation does not have to be extracted from documents and it does not have to be part of the ontologies.

Aside from determining what we needed to model, the question was what we actually did not need to model. For instance, for patient identification, the system does not have to be able to derive diagnoses based on symptoms: This was already done by the treating physicians. Also, we we found that, other than "part-of", we did not have to model functional relationships within the body or processes related to the respective disease or its treatment. We found that this knowledge is important when formulating the right criteria for the system, but it does not have to be part of the knowledge base.

Regarding the logical structure of the trial protocols, we found that both inclusion and exclusion criteria are relatively simple logical expression that can be expressed using AND, OR, and NOT. However, additional complexity arises from the fact that some criteria are time-dependent. For instance, it might be required that a patient has not been treated with a specific type of medication within in the last three months. Moreover, documents frequently state negated and uncertain information. This must be taking into account when matching trial criteria and extracted facts, see section 3.4. Also, quantification and counting is required to some extend.

As a general result of the analysis, we were able to verify that we cannot rely on a single type of information alone but need the combination of structured data and facts extracted from documents. For instance, the location of a stroke can only be found in text documents whereas the NIHSS (NIH stroke score) is frequently documented in the admission report in a structured manner. Both sources, however, are not 100 % complete. This means that the original criteria frequently have to be translated into system queries that combine (CRDW) criteria that are semantically similar, but pertain to different data sources.

Regarding the question of data quality and availability, we found that frequently the required information is not documented in the CIS (e.g., medicolegal aspects). Or it is not yet available, when searching for eligible patients. This is, for instance, the case for certain stroke studies, which require that a specific treatment is begun within only a couple of hours of the stroke event. Coded information is frequently only available after a couple of days. The same holds true of discharge letters, which are only available after the patient has already left the hospital.

### 3.2 Technical Considerations

When starting to develop a prototype of the system, it was necessary to make a decision for a specific semantic technology. In order to get started, we decided to

set up a SESAME server (`http://www.openrdf.org/`) and to use RDFS (RDF Schema) for modeling some basic concepts.

In RDFS, one can, for instance, establish subclass relationships between concepts and subproperty relationships between properties. In our ontologies, we use the following primitives:

- `rdfs:subClassOf` for the subclass relationship
- `rdfs:subPropertyOf` for properties
- `rdfs:label` for specifying synonyms

However, it is not possible in RDFS to define a concept as the conjunction, disjunction or negation of other concepts. Neither is it possible to state rules that derive properties or class membership for instances.

The lack of rules is partly compensated by the query interface of the CRDW. In our system, we allow conjunction and disjunction of positive criteria plus a conjunction of negated criteria. This means that although we cannot state rules or complex concept definitions, the user can use logical combinations of search criteria. For instance, instead of inferring that a patient has diabetes from the fact that he or she is treated with Metformin, the user of our system can issue a search for `diabetes OR metformin` (potentially plus other indicators for diabetes).

An example, in which (simple) rules are helpful, are implicit attributes. For instance, we have the concept `epilepsy` in our ontology plus its potential attributes `generalized` or `focal`. There is a special type of epilepsy, called matutinal epilepsy, which is always of type `generalized`. Therefore, in a medical document, one does never find the term "generalized matutinal epilepsy", since "generalized" is redundant. This means that this information must be inferred by the system. In the current state of the system, the user has to define the query `epilepsy(generalized) or matutinal epilepsy`. Alternatively, we allow internal rewriting of linguistic expressions to include more information.

As part of our Scrum [22] software development process, we carefully evaluated several alternatives and extensions to RDFS: PROLOG [11], F-Logic/Object Logic [10], Datalog [7], Production Rules (Drools) [2], RIF [9], OWL 2 [27], SPARQL Rules (SPIN) [14]. Some of these languages are very powerful but lack built-ins for modeling ontologies (e.g., PROLOG, Drools, Datalog). The remaining approaches pertain to ontologies, however many of them could not be considered mature enough to be included into a commercial software product, which is the ultimate goal of our project. In general, we found SPIN the most attractive approach since it is relatively powerful but lightweight, and it features negation as failure. However, we found that the problems introduced by not having rules were not critical. We therefore postponed the introduction of SPIN to future versions of our system.

### 3.3 Extracting Patient Information from Documents

An important insight of the project was that not only the ontology contents but also its structure has to support the task of extracting information from texts.

As stated in the previous section, we wanted to extract diagnoses and therapies from text sources. In the first versions of the system, one of the problems was that attributes were frequently attached to the wrong diagnoses. In order to help the concept mapping algorithm with attaching properties to the right main terms, we decided to specify possible attributes for each concept in order to reduce ambiguities.

Consider, for instance, the phrase "acute MCA infarction". In the ontology, infarction is specified to have potential attributes "`:Acute`" and `:ArteriaCerebriMedia`. This is achieved by the following declaration:

```
:Infarction rdf:type rdfs:Class ;
       rdfs:subClassOf :Observation;
       :label "Infarction" ;
       :hasLocalization :Artery ,
              :Brain ,
              :Heart ;
       :hasAttribute :Position ,
              :Modifiers ,
              :InfarctionAttributes .
```

`:Acute` is defined to be a subclass of `:Modifiers`.

As can be seen from the example, we extensively made use of the meta-modelling feature of RDFS that allows to treat classes as instances. In terms of general AI terminology, our approach can be seen as frame-based [19]: We specify slots for each disease, which are filled by the linguistic pipepline based on information given in the text.

### 3.4 Mapping Patient Information to Criteria

Clinical trials usually pertain to information that is assumed to be certain. Mapping positive facts to criteria is therefore straightforward. However, radiological findings, admission reports, and doctor's letters frequently contain negated and uncertain information. The handling of such information can be a relatively difficult topic in the clinical context since there does not seem to be a strategy that is always the right one.

The criteria of a clinical trial are divided into inclusion and exclusion criteria, the latter of which can be considered negated criteria. Exclusion criteria are usually evaluated in a "closed world manner". This means that an exclusion criterion only matches whenever there is no corresponding fact that can be considered certain. In some cases, however, doctors preferred a more conservative "open world approach", with exclusion criteria matching only explicit negative statements. We are therefore planning on allowing the user to choose between open and closed world semantics.

A related issue is the handling of uncertain facts (e.g., suspected diagnoses). For instance, we found it crucial that the system avoids *false negatives*, i.e.

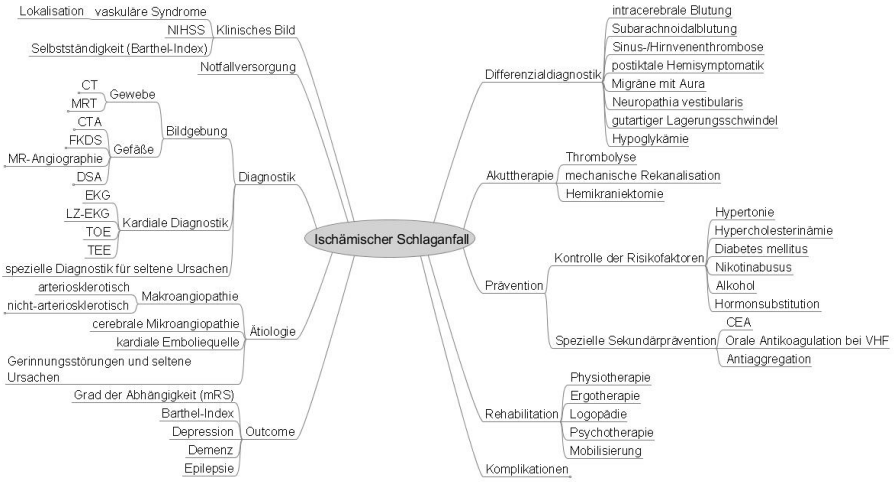


**Fig. 2.** Mindmap for ischemic stroke (simplified)

patients that are not suggested for a trial although they are potential candidates. This means that if there is an *uncertain* fact for an exclusion criterion, the respective criterion is not supposed to match. In contrast, inclusion criteria are (usually) required to match also uncertain facts. Since this is not always the case, though, we are planning on letting the user decide the matching behavior of each criterion.

### 3.5 Modeling Diseases

In order to better understand the structure of disease models, we asked the doctors in our group to draw mindmaps of the diseases. We were primarily interested in the following diseases: *ischemic stroke*, *idiopathic parkinson disease*, *multiple sclerosis*, and *epilepsy*. For instance, the mindmap of *ischemic stroke* consists of the branches: clinical picture, etiology, diagnostics, differential diagnoses, acute therapy, emergency medical care, complications, outcome, rehabilitation, prevention, see fig. 2. Similar branches can also be found in the mindmaps for the other diseases.

For the task of patient identification, however, we found that is only necessary to represent the concepts present in the mindmap, but not the semantic relations corresponding to the branches and their labels: For instance, we can can find stroke patients suffering from a paresis by simply issuing the query `stroke AND paresis`. In order to be able to answer this query, however, it is not necessary to represent the fact that a paresis is a potential symptom of a stroke in the ontology. This confirmed that we could use the relatively simple ontology model sketched in the previous section.

Other than understanding the required structure of our ontologies, in the beginning the mindmaps served as a basis for defining the ontologies. Later on, we added concepts mentioned in annotated documents and also incorporated concepts that are found in the ICD-10 classification of disease.

### 3.6 Enriching the Ontologies

In order to identify further concepts that needed to be part of the ontology, and also in order to be able to construct a set of test sentences, the clinical doctors were asked to annotate phrases in a set of example documents chosen by them. In the beginning, they just used a text marker on a print-out of the documents. Based on the annotated documents, the ontology models were extended by a knowledge engineer. In the future, however, we plan to utilize an annotation tool that allows the doctors to annotate the relevant phrases graphically followed by a semi-automatic step of ontology extension.

At the moment, the ontologies are being developed by several people who work at different locations and have different backgrounds. Yet, there is a strong overlap between ontologies for separate diseases because co-morbidity has also to be modeled to some extend, and anatomical and attributive information have to be shared between ontologies. For now, we decided to use independent ontology modules. In order to be able to use different modules in parallel, we plan to use techniques of ontology alignment [23,26].

### 3.7 Available Clinical Ontologies

In our project, we also investigated if we can use already existing knowledge bases. The results are summarized in the following.

**UMLS** (Unified Medical Language System) [1] is a so-called meta-thesaurus, which combines several thesauri by the means of a common semantic network. Since most of the resources are not available in German, and license conditions are frequently problematic for the use in a commercial software, we were not able to use this powerful resource in our project. UMLS is also used for the linguistic component in I2B2 [13], a system that has a similar purpose as the CRDW.

**MESH** (Medical Subject Headings) [16] is a controlled vocabulary for indexing the MEDLINE/PUBMED database. There exists also a German version (MESH GER), which we licensed for the project. However, in the end we did not use it in our software since the concepts did not meet the requirements of an ontology and were not consistent with our modeling strategy. Consequently we favored modeling the ontology from scratch with the help of domain experts.

As an example, there is a MESH GER concept called "Infarkt, A. cerebri media". The first problem is that we need two separate concepts in our ontology: a diagnosis and a localization. Moreover, the MESH GER concept comprises non-synonymous labels like "A.-cerebri-media-Syndrom", "A.-cerebri-media-Embolus", "A.-cerebri-media-Thrombose", "Left Middle Cerebral Media Infarction", "Right Middle Cerebral Media Infarction" and several variants of

"Infarkt, Arteria cerebri media". As a third problem, we found that many concepts relevant for us where missing.

**OpenGalen**: The GALEN Common Reference Model (CRM) is a clinical terminology, which was developed in a project funded by the European Union. The English version is available as an OWL download whereas we could not find any German version. In general, we found the structure of the GALEN common reference model much too complex for our project. We had the feeling that constructing a simpler ontology from scratch is preferable to adapting the structure of the GALEN model for our purposes.

**SNOMED/SNOMED CT:** SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms, [18]) is a well-known health care terminology. SNOMED CT consists of a "IS_A" hierarchy along with the possibility to define concepts based on attributes. SNOMED, the predecessor of SNOMED CT, was defined using 11 groups of concepts. Since there is no (available) German version of either SNOMED or SNOMED CT, we were not able to use it in our project. The same reason prevented us from using the Foundational Model of Anatomy (**FMA**, [17]) and **RADLEX**.

**ICD-10:** The "International Statistical Classification of Diseases and Related Health Problems, 10th Revision" (ICD-10) is the most important classification of diagnoses. It is widely used for billing purposes in German hospitals. Although the ICD-10 is quite broad on the one hand, it is not fine-grained enough for our purposes: for instance, with respect to the diagnosis "stroke", one is usually interested in the specific location of the stroke. However, the ICD-10 only distinguishes between cerebral and pre-cerebral arteries. **OPS** plays a similar model as ICD-10. We use it in our software for searching structured data, but do not use it for extracting therapeutic information from texts.

## 4 Evaluation

We evaluated the performance of the CRDW with respect to the assessments of the trial team of the Center for Stroke Research Berlin (CSB), Charité, see section 4.1. The results of this evaluation are based on both structured and unstructured data. In order to evaluate how useful unstructured data for patient recruitment are, we also compared different versions of the data set. Details of this experiment are given in section 4.2. In cooperation with the trial team of the Center for Epilepsy at the Vivantes Humboldt-Klinikum, we furthermore evaluate the usability of the CRDW in productive use. First results of the qualitative evaluation can be found in section 4.3.

### 4.1 Charité – Universitätsmedizin Berlin

In the evaluation, we considered 16 stroke trials, which are currently being conducted at the Clinic of Neurology, and compared the performance of the CRDW to that of the trial team. whose assessment of patients was considered the gold standard. The data were collected from January to March 2013. For each trial, we determined the following quantities:

**Table 1.** System performance for several clinical trials: sample size, number of true positives, false positives, true negatives, false negatives, recall (sensitivity), specificity, precision (positive predictive value), negative predictive value, F-measure

| Trial | N | TP | FP | TN | FN | Recall | Spec. | Prec. | NPV | F |
|---|---|---|---|---|---|---|---|---|---|---|
| Trial 1 | 268 | 1 | 20 | 246 | 1 | 0.50 | 0.92 | 0.05 | 1.00 | 0.09 |
| Trial 2 | 257 | 39 | 126 | 82 | 10 | 0.80 | 0.39 | 0.24 | 0.89 | 0.36 |
| Trial 3 | 177 | 0 | 2 | 175 | 0 | 1.00 | 0.99 | 0.00 | 1.00 | 0.00 |
| Trial 4 | 268 | 1 | 23 | 242 | 2 | 0.33 | 0.91 | 0.04 | 0.99 | 0.07 |
| Trial 5 | 136 | 1 | 11 | 124 | 0 | 1.00 | 0.92 | 0.08 | 1.00 | 0.15 |
| Trial 6 | 253 | 51 | 90 | 101 | 11 | 0.82 | 0.53 | 0.36 | 0.90 | 0.50 |
| Trial 7 | 255 | 44 | 120 | 80 | 11 | 0.80 | 0.40 | 0.27 | 0.88 | 0.40 |
| Trial 8 | 267 | 17 | 22 | 228 | 0 | 1.00 | 0.91 | 0.44 | 1.00 | 0.61 |
| Trial 9 | 259 | 13 | 27 | 219 | 0 | 1.00 | 0.89 | 0.33 | 1.00 | 0.49 |
| Trial 10 | 251 | 0 | 0 | 250 | 1 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| Trial 11 | 268 | 0 | 7 | 261 | 0 | 1.00 | 0.97 | 0.00 | 1.00 | 0.00 |
| Trial 12 | 236 | 8 | 61 | 166 | 1 | 0.89 | 0.73 | 0.12 | 0.99 | 0.21 |
| Trial 13 | 269 | 6 | 35 | 226 | 2 | 0.75 | 0.87 | 0.15 | 0.99 | 0.24 |
| Trial 14 | 254 | 52 | 118 | 79 | 5 | 0.91 | 0.40 | 0.31 | 0.94 | 0.46 |
| Trial 15 | 245 | 2 | 31 | 209 | 3 | 0.40 | 0.87 | 0.06 | 0.99 | 0.11 |
| Trial 16 | 245 | 3 | 2 | 236 | 4 | 0.43 | 0.99 | 0.60 | 0.98 | 0.50 |

- N: total number of patients considered
- TP (true positives): number of eligible patients (according to the trial team) that were found eligible by the CaseSearch
- FP (false positives): number of non-eligible patients that were found eligible by the CaseSearch. False positives increase the workload when using our system.
- TN (true negatives): number of non-eligible patients that were found non-eligible by the CaseSearch
- FN (false negatives): number of eligible patients that were found non-eligible by the CaseSearch. This is the most critical quantity since it means that the user of the system might miss potential candidates.

Table 1 shows the evaluation results when using both structured and unstructured data. For the 16 selected neurological trials, the table shows the recall (sensitivity) $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$, specificity $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$, precision $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$ (positive predictive value), the negative predictive value $\frac{\mathrm{TN}}{\mathrm{TN+FN}}$, plus the so-called F-measure. The F-measure is the harmonic mean of precision and recall. It is defined as

$$F = \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}\,.$$

There is a number of trials, for which there is only a small number of candidates (= TP + FN). This means that the computed quantities are not very reliable and might attain very high or low values. For instance, if there are no

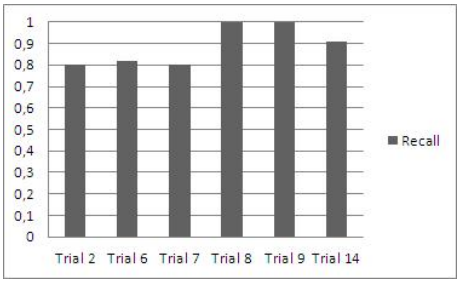


**Fig. 3.** Sensitivity/Recall (only trials with 10 or more candidates)

candidates for a clinical trial, the recall is defined as 1.0. In the following, we therefore considered only trials with more than 10 candidates.

Fig. 3 shows the recall (sensitivity) of the system for the six remaining trials. The diagram shows that, in all six cases, we were able to achieve a good to high sensitivity. This means that the system suggests most of the eligible patients. Achieving good recall was of uttermost importance for the task of patient recruitment.

The specificity (i.e., recall of non-candidates) corresponds to the probability that a patient is not suggested by the system if he or she is not eligible for a trial (see fig. 4). The specificity values are medium to high.

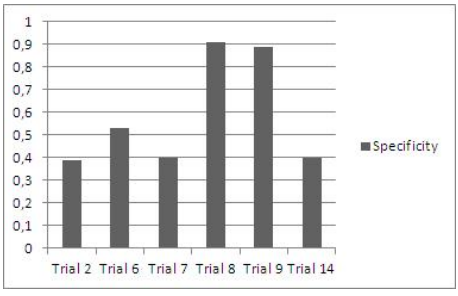


**Fig. 4.** Specificity

Precision is an estimate of the probability that whenever the system suggests a patient for a clinical trial, he or she is actually eligible. Compared to recall and specificity, the precision attains lower values (see Fig. 5). This means that the system tends to incorrectly suggest patients as candidates. This reduces the potential amount of time that can be saved when working with the system.

One problem regarding precision is that not all necessary information is documented in the clinical information system. Some information is just missing or incomplete (e.g., NIH stroke score, medications) due to the work load in the ER. Other information can only be obtained by talking to the patient or by further

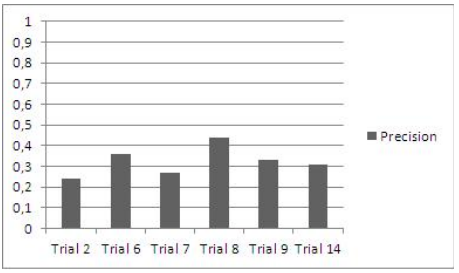


**Fig. 5.** Precision

examinations. This means that it is usually not possible to attain a precision of 1.0. In order to still improve the performance of the CRDW, we are currently in the process of increasing the logical expressiveness of our query interface, which does not correspond to full SPARQL yet. Other improvements concern the handling of negation and uncertainty as described earlier.

The negative predictive value (i.e., precision for non-candidates) corresponds to the probability that, if the system predicts non-eligibility, the corresponding patient is not a candidate. The negative predictive values are relatively high meaning that negative predictions have a relative high probability of being correct. For reasons of space, we leave out the respective diagram.

### 4.2 Structured vs. Unstructured Data

In order to determine the usefulness of our ontology-based approach, we considered two variants of the data and the study criteria:

- S+U: This corresponds to the complete data set, comprising unstructured data (U) as well as structured data (S). The trial criteria for "S+U" were defined by medical doctors, specialized in the field of neurology. When setting up the criteria, their focus was on recall while at the same time achieving acceptable levels of specificity and precision.
- U: Unstructured data only. Possible criteria correspond to coded diagnoses (ICD-10) and procedures (OPS), lab values, age, sex, and some structured information from the admission report like the stroke scores. For each trial, the criteria for "version S" were obtained by removing those conditions for "version S+U" that pertain to information contained in documents.

The table 2 shows recall, specificity, precision, negative predictive value, and F-measure averaged over all 6 trials. Using both structured and unstructured data results in a high recall and a medium specificity. Dropping conditions from the criteria that pertain to texts results in lower averaged recall, precision, specificity, negative predictive value, and F-measure.

Since the number of trials, 6, is relatively small, we could not show that the differences are statistically significant. We therefore performed a statistical

**Table 2.** Recall, specificity, precision, negative predictive value, $F$ (averages over all trials)

| | Rec. | Spec. | Prec. | NPV | F |
|---|---|---|---|---|---|
| S+U | 0.89 | 0.59 | 0.32 | 0.94 | 0.47 |
| S | 0.81 | 0.54 | 0.31 | 0.92 | 0.43 |

analysis of the single trials. For recall and specificity it is possible to determine significant differences using McNemar's test. The list of patients is considered a related sample for the two versions of the trial. In the case of recall, a patient is labeled with 1.0 if he or she is found eligible by both the trial team and our software. If our software fails to identify an eligible patient, he or she is labeled with 0.0. All other case are not included in the sample. Specificity is handled analogously.

Table 3 summarizes recall and precision for both versions of all trials. With respect to recall, version "S+U" is better or equal than "U" in 14 cases out of 16. We ran the SPSS McNemar's test in order to determine the trials for which there are differences which are statistically significant. The table gives the p-values for the one-sided tests. In the table, a "+" occurs whenever "S+U" performs significantly better than "U", i.e., if $p \leq 0.05$ holds. If "U" is significantly better than "S+U", there is a "-". For two trials, we found that "S+U" performed significantly better than "U". "=" means that the null hypothesis cannot be rejected, i.e., equal probabilites are assumed. There are no trials, for which "U" performed significantly better with respect to recall. With respect to specificity, "S+U" is better or equal than "U" in 11 cases out of 16. In eight cases, "S+U" is significantly better than "U". The reverse is true in only 4 cases.

### 4.3 Vivantes - Netzwerk für Gesundheit GmbH

As a state-owned hospital group, the prime objective of Vivantes is to provide community health care services. Nevertheless, clinical research is in strong focus, and a large number of clinics engage in the execution of (mostly sponsor-initiated) clinical trials. The crucial requirement for the CRDW thus is that it has to fit perfectly in to the interwoven processes of health care and clinical research. Starting from this situation, we evaluate the usability of the CRDW system in production use.

In our experiment, a custom-built HL7 adapter provides selected data of neurological case records for an instance of the CRDW. The data comprise both structured data and clinical documents out of the HL7 data stream. Since March 2013, members of the trial team of the Neurological Clinic at the Vivantes Humboldt-Klinikum are using the CaseSearch application for case identification. At present, the trial team is exclusively conducting epilepsy trials.

**Table 3.** Recall and specificity: results of McNemar's test (SPSS). Explanation see text.

| Trial | Rec. (S+U) | Rec. (S) | p-value | Decision | Spec. (S+U) | Spec. (S) | p-value | Decision |
|---|---|---|---|---|---|---|---|---|
| Trial 1 | 0.50 | 0.50 | 0.500 | = | 0.92 | 0.77 | 0.000 | + |
| Trial 2 | 0.80 | 0.63 | 0.019 | + | 0.39 | 0.59 | 0.000 | - |
| Trial 3 | 1.00 | 1.00 | | n.a. | 0.99 | 0.77 | 0.000 | + |
| Trial 4 | 0.33 | 0.33 | 0.500 | = | 0.97 | 0.97 | 0.500 | = |
| Trial 5 | 1.00 | 1.00 | 0.500 | = | 0.92 | 0.77 | 0.000 | + |
| Trial 6 | 0.82 | 0.74 | 0.090 | = | 0.53 | 0.59 | 0.022 | - |
| Trial 7 | 0.80 | 0.65 | 0.028 | + | 0.40 | 0.57 | 0.000 | - |
| Trial 8 | 1.00 | 0.94 | 0.500 | = | 0.91 | 0.95 | 0.001 | - |
| Trial 9 | 1.00 | 1.00 | 0.500 | = | 0.89 | 0.29 | 0.000 | + |
| Trial 10 | 0.00 | 0.00 | 0.500 | = | 1.00 | 1.00 | 0.500 | = |
| Trial 11 | 1.00 | 1.00 | | n.a. | 0.97 | 0.22 | 0.000 | + |
| Trial 12 | 0.89 | 1.00 | 0.500 | = | 0.73 | 0.34 | 0.000 | + |
| Trial 13 | 0.75 | 0.75 | 0.500 | = | 0.87 | 0.78 | 0.000 | + |
| Trial 14 | 0.91 | 0.91 | 0.500 | = | 0.40 | 0.44 | 0.170 | = |
| Trial 15 | 0.40 | 0.60 | 0.500 | = | 0.87 | 0.69 | 0.000 | + |
| Trial 16 | 0.43 | 0.43 | 0.500 | = | 0.99 | 0.99 | 1.000 | = |

We evaluate the experiment by conducting semi-structured qualitative interviews (i.e., interviews with open questions that allow for dialogue and discussion) with the users. First results of the ongoing evaluation phase are:

- Recall: Compared to the previous procedure of the trial team (reviewing new cases in the clinic information system), the CRDW shows a high recall rate. Users did not detect any patients matching a clinical study that the system had missed.
- Precision: The list of matching cases generated by the CRDW is rather unspecific. This is due to the fact that some exclusion criteria cannot be checked automatically, since the corresponding data is ambiguous or is not available in a digital format at all. Nevertheless, the system reliably sorts out cases with neurological diagnoses other than epilepsy. It also sorts out cases with types of epilepsy and/or types of seizures that are excluded by the trial protocols.
- Term extraction: The system recognizes essential synonyms and hyponyms. In almost all cases, the various types of epilepsy and epileptic seizures as well as the names of drugs and drug agents were correctly recognized.

Note that all results are still preliminary and will have to be substantiated by future experiments.

## 5 Conclusion

In this paper, we described a case study in using ontologies for information extraction from clinical documents. We demonstrated that we managed to build a system with a high sensitivity – a requirement for the task of patient recruitment. Improving precision, however, is still an issue. Future work will focus on the elimination of false positives by allowing logically more complex criteria.

The experimental data suggest that the process of patient identification benefits from extracting facts from structured data. We are planning to obtain more reliable results by considering more patients and trials. Moreover, the software will be tested by other departments, too.

A lesson learned in the area of ontologies is that it can be much easier to construct an ontology for a specific application instead of building or even using a general-purpose ontology. However, we also feel that the lack of German language resources hinders progress in the domain of semantic technologies suitable for German text and web resources.

**Acknowledgments.** The project is partially funded by TSB Technologiestiftung Berlin, Zukunftsfonds Berlin, and co-financed by the European Union – European Fund for Regional Development.

We would like to thank all computer scientists, linguists, ontologists, medical doctors, study nurses, and administrative staff who participated in the development of the software.

## References

1. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Research 32(Database-Issue), 267–270 (2004)
2. Browne, P.: Jboss Drools Business Rules. From technologies to solutions. Packt Publishing, Limited (2009)
3. Chapman, W.W., Hilert, D., Velupillai, S., Kvist, M., Skeppsted, M., Chapman, B.E., Conway, M., Tharp, M., Mowery, D.L., Deleger, L.: Extending the negex lexicon for multiple languages. In: Proceedings of the 14th World Congress on Medical and Health Informatics, MEDINFO 2013 (2013)
4. Cowie, J., Wilks, Y.: Information extraction. In: Handbook of Natural Language Processing, pp. 241–260 (2000)
5. Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K.: Getting more out of biomedical documents with gate's full lifecycle open source text analytics
6. Dugas, M., Lange, M., Berdel, W., Müller-Tidow, C.: Workflow to improve patient recruitment for clinical trials within hospital information systems - a case-study. Trials 9(1), 2 (2008)
7. Gallaire, H., Minker, J., Nicolas, J.-M.: Logic and databases: A deductive approach. ACM Comput. Surv. 16(2), 153–185 (1984)
8. Jurafsky, D., Martin, J.H.: Speech and Language Processing, 2nd edn. Prentice Hall Series in Artificial Intelligence. Prentice Hall (May 2008)
9. Kifer, M.: Rule interchange format: The framework. In: Calvanese, D., Lausen, G. (eds.) RR 2008. LNCS, vol. 5341, pp. 1–11. Springer, Heidelberg (2008)

10. Kifer, M., Lausen, G., Wu, J.: Logical foundations of object-oriented and frame-based languages. Journal of the ACM 42(4), 741–843 (1995)
11. Lloyd, J.W.: Foundations of Logic Programming, 2nd edn. Springer (1987)
12. Müller, F.: A finite-state approach to shallow parsing and grammatical functions annotation of German. PhD thesis, University of Tübingen (2005)
13. Murphy, S.N., Mendis, M.E., Berkowitz, D.A., Kohane, I., Chueh, H.: Integration of clinical and genetic data in the i2b2 architecture. In: AMIA Annu. Symp. Proc., p. 2009 (2006)
14. Polleres, A.: From SPARQL to rules (and back). In: Williamson, C.L., Zurko, M.E., Patel-Schneider, P.F., Shenoy, P.J. (eds.) WWW, pp. 787–796. ACM (2007)
15. Reeve, L.: Survey of semantic annotation platforms. In: Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 1634–1638. ACM Press (2005)
16. Rogers, F.B.: Medical subject headings. Bull. Med. Libr. Assoc. 51, 114–116 (1963)
17. Rosse, C., Mejino, J.V.L.: A reference ontology for biomedical informatics: the foundational model of anatomy. J. Biomed. Inform. 36, 478–500 (2003)
18. Ruch, P., Gobeill, J., Lovis, C., Geissbühler, A.: Automatic medical encoding with SNOMED categories. BMC Medical Informatics and Making 8, 6 (2008)
19. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn. Pearson Education (2003)
20. Scheitz, J.F., Mochmann, H.C., Witzenbichler, B., Fiebach, B., Audebert, H.J., Nolte, C.H.: J. Neurol. 25 (2012)
21. Scheitz, J.F., Mochmann, H.C., Nolte, C.H., Haeusler, K.G., Audebert, H.J., Heuschmann, P.U., Laufs, U., Witzenbichler, B., Schultheiss, H.P., Endres, M.: Troponin elevation in acute ischemic stroke (TRELAS) – protocol of a prospective observational trial. M. BMC Neurol. 11(98) (2011)
22. Schwaber, K., Beedle, M.: Agile Software Development with Scrum, 1st edn. Prentice Hall PTR, Upper Saddle River (2001)
23. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. IEEE TKDE 99 (2011)
24. Staab, S., Studer, R.: Handbook on Ontologies, 2nd edn. Springer (2009)
25. Szarvas, G., Farkas, R., Busa-Fekete, R.: Research paper: State-of-the-art anonymization of medical records using an iterative machine learning framework. JAMIA 14(5), 574–580 (2007)
26. Todorov, K., Geibel, P., Kühnberger, K.-U.: Mining concept similarities for heterogeneous ontologies. In: Perner, P. (ed.) ICDM 2010. LNCS, vol. 6171, pp. 86–100. Springer, Heidelberg (2010)
27. Yu, L.: A Developers Guide the Semantic Web. Springer (2011)