Semantic AutoSuggest for Electronic Health Records[1]

Ulrich Beez, Bernhard G. Humm
Hochschule Darmstadt – University of Applied Sciences
Darmstadt, Germany
ulrich.beez@stud.h-da.de, bernhard.humm@h-da.de

Paul Walsh
NSilico Lifescience Ltd.
Bishopstown, Co. Cork, Ireland
paul.walsh@nsilico.com

***Abstract*— In traditional Electronic Health Record (EHR) applications, medical information is provided in free text fields resulting in ambiguities in the use of medical vocabulary. This paper presents a semantic AutoSuggest service that supports consultants entering free-text medical information while at the same time normalizing the medical vocabulary used. For this, several medical ontologies are semantically integrated. Particular focus is on usability, optimizing the relevance of suggested terms as well as supporting diversity.**

Keywords*— *Electronic Health Record (EHR), Semantic AutoSuggest; Medical Ontologies; Semantic Ontology Integration

I. INTRODUCTION

The use of EHRs is necessary to improve clinical research efficiency [1-3]. Medical information technology has recently advanced in many countries, and enormous amounts of clinical data are already stored as electronic medical records [2].

A large amount of EHR data is stored in free-text, providing the maximum flexibility for consultants to express case-specific issues. However, using free-text has a downside for mining EHR data, as medical terminology is used in diverse ways by various medical professionals and across different regions. For example synonyms are in widespread use in the medical community, along with abbreviations, and even misspellings. While this usually poses no problem for the human expert, it is difficult for software modules to deal with those kinds of ambiguities in a reliable way.

To cope with these issues, text mining approaches have been proposed to disambiguate texts in EHRs [4]. While such an analytic approach is unavoidable when dealing with existing legacy EHR data, we propose a constructive approach for new EHR applications: a semantic AutoSuggest service (see Fig. 1).

While typing input into a free text field, suggestions of medical terms of various categories (anatomy, symptom, disease, etc.) are being presented. Example: "ipilimumab (medication)" while typing "ip". While moving the mouse over an entry, an explanatory text is shown.

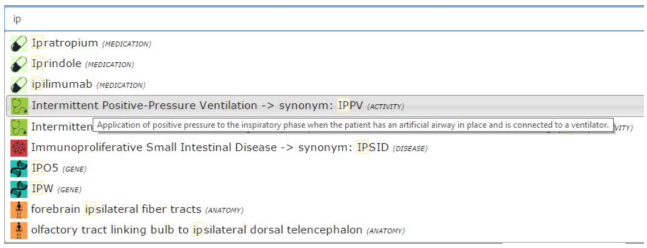


Fig. 1:Semantic AutoSuggest service

Semantic AutoSuggest not only improves usability by reducing typing effort for the consultant. As importantly, it normalizes the usage of medical terminology: instead of using synonyms, abbreviations or even misspelling terms, always the same preferred term is used for a concrete medical concept.

The remainder of this paper is structured as follows. Requirements are specified in Section II. Sections III and IV are the core of the paper and describe the concept and an implementation of the semantic AutoSuggest service. Section V evaluates our approach. Related work is reviewed in Section VI. Section VII concludes the paper and describes future work.

II. REQUIREMENTS

Having consulted extensively with clinicians involved in the treatment of melanoma, we have identified the following requirements:

[1] This work was funded by the European Commission, Horizon 2020 Marie Skłodowska-Curie Research and Innovation Staff Exchange, under grant no 644186.

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.85

1. A *semantic AutoSuggest service* shall support consultants editing EHRs by suggesting medical terms while typing.
2. The *semantically categorized medical terminology* shall contain all relevant terms for the respective clinical use cases.
3. The medical terminology used shall be *extensible*.
4. The Service shall sort terms according to *relevance* but at the same time present *diverse* terms, i.e., from various medical categories.
5. The meaning of terms should be outlined by an *explanation*.
6. The *response time* of the service shall be at typing speed.
7. Using the semantic AutoSuggest service shall *normalize the vocabulary* used in the EHR thus enabling semantic search and data mining of EHRs.

III. Semantic AutoSuggest for Electronic Health Records

A. Semantic Categories

One of the challenges in providing an AutoSuggest feature is to identify the semantic categories that are appropriate and necessary. These categories can best be identified for a concrete EHR application by methodically analyzing all free text fields. See Fig. 2 for prominent free-text fields of a melanoma care EHR application.

Prominent EHR field	Category and Icon
medication used by the patient	Medication
treatments the patient received	Activity
diagnoses and findings by the physicians	Symptom
patient's diseases	Disease
findings of the gene analysis	Gene
body parts where a melanoma occurs	Anatomy
other relevant health issues	all above

Fig. 2:Categorization of EHR free text fields

As shown in the table above, six distinct categories where identified for a melanoma management application: medication, activity, symptom, disease, gene, and anatomy. Some free-text fields require words from one category only, e.g., "medication used by the patient". Others can be filled with terms form multiple categories, e.g., "other relevant health issues".

B. Medical Ontologies

In the medical domain, numerous controlled vocabularies, thesauri and ontologies exist. They contain medical terms and, potentially, additional information such as explanations, synonyms, hyperonyms (broader terms), and domain-specific terms relationships. Following Liu and Özsu ([5] p. 360), we use the term "ontology" within this paper to refer to all kinds of classified terminology in the medical domain.

Whereas some medical ontologies are commercial (e.g., Unified Medical Language System® Metathesaurus®, SNOMED- CT, etc.), there are many open source ontologies available (for an overview see, e.g., www.ontobee.org).

Another challenge that needs to be addressed is how to select an ontology or a set of ontologies as the base vocabulary for the semantic AutoSuggest service. Again, the EHR application under consideration needs to be analyzed and relevant terms need to be identified. Then, the completeness of certain ontologies with respect to the identified relevant terms can be assessed.

For a melanoma EHR application, 50 relevant terms have been compared against prominent open source medical ontologies. See Fig. 3 for a subset of medication terms.

Category	Term	The Drug Ontology	National Drug File - Reference Terminology	Human Disease Ontology	Anatomical Entity Ontology	Foundational Model of Anatomy	Uber Anatomy Ontology	Gene Ontology	Ontology of Genes and Genomes	Symptom Ontology	Medical Subject Headings	National Cancer Institute Thesaurus
Medication		x	x								x	x
	Warfarin	x	x								x	x
	NOAC	-	-								x	(x)
	Antiplatelet	-	-								x	x
	Dacarbazine	x	x								x	x
	Ipilimumab	x	-								x	x

Fig. 3: Checking ontology completeness by sample terms

In analyzing the melanoma case study it was observed that no single ontology contains all relevant terms. Also, no set of ontologies covers all terms of different categories, e.g., The Drug Ontology all

relevant medications, the Human Disease Ontology all relevant diseases, etc.

It can be concluded that only by integrating several ontologies, a sufficiently comprehensive terminology can be established. For an overview of the ontologies selected, see Fig. 4.

However, the decision to semantically integrate various ontologies comes hand in hand with various integration issues, e.g., the handling of duplicates. We will go into more details in the next section.

Name	Anatomy	Symptom	Gene	Disease	Activity	Medication	License
The Drug Ontology (DRON)						x	open
National Drug File Reference Terminology (NDF-RT)						x	open
Human disease ontology (DOID)				x			open
Anatomical Entity Ontology (AEO)	x						open
Foundational Model of Anatomy (FMA)	x						open
Uber anatomy ontology (UBERON)	x						open
Gene Ontology (GO)			x				open
Ontology of Genes and Genomes (OGG)			x				open
VIVO-ISF					x		open
Symptom Ontology (SYMP)		x					open
Medical Subject Headings (MeSH)	x	x	x	x	x	x	Registration necessary
NCI Thesaurus (National Cancer Institute)	x	x	x	x	x	x	open

Fig. 4: Overview of medical ontologies and semantic categories covered

C. Integration of Medical Ontologies

When semantically integrating various ontologies, the following issues need to be addressed.

1. *Definition of a target data format*: Because ontologies use different data formats, a common target format is needed.
2. *Transformation of technical data formats*: Ontologies have different technical formats, e.g., XML, XLS, CSV, RDF. A transformation from the specific to the common format is required.
3. *Semantic field mapping*: Even if the technical formats are identical, e.g., XML, the individual field names and structure of the ontologies may differ. E.g., broader terms in MeSH are encoded as tree id whereas in other ontologies, the ids of the broader terms are listed.
4. *Semantic cleansing*: Some terms are "polluted" (have unwanted parts), e.g. "monooctanoin *[Chemical/Ingredient]*" in Ontology NDF-RT, where the text in italics needs to be removed.
5. *Semantic filtering*: Some ontologies contain terms that are not meaningful for the semantic AutoSuggest service, e.g. the term "Non-physical anatomical entity" in the Foundational Model of Anatomy. Those terms need to be filtered out.
6. *Duplicate handling*: Duplicate terms occur because of some terms are covered in various ontologies (e.g., "Warfarin" in The Drug Ontology and in MeSH), and even in various versions within the same ontology. For example, the Uber Anatomy Ontology contains the term "gene" twice; one time, marked as "deprecated", without any broader term, synonyms or definition. The goal of duplicate handling is to achieve a unique set of terms for the semantic AutoSuggest service.

Target Data Format

The target data format for terms to be used in the semantic AutoSuggest service is kept minimal to facilitate fast searching, but this scheme can easily be extended. For the melanoma case study it was found that the following attributes suffice: label (the name of the term), category (anatomy, symptom, disease, etc.), definition (explanatory text), broader (hyponyms), synonyms. See Fig. 5.

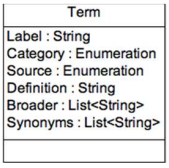


Fig. 5: Term attributes

Duplicate handling

The notion of duplicates can be interpreted in a number of ways, so we use the following definition: Two terms are considered duplicates if they have the same label or the label of one term is a synonym of the other. Similarity checks ignore case.

When duplicates are identified, then a heuristics-based ranking determines which of both terms to keep. The heuristics we are using has two components:

1. *Ontology ranking*: a statically assigned rank *r* for each ontology defines a partial ordering of all ontologies, e.g., $r_{MeSH} > r_{Drug\ Ontology}$.
2. Term ranking: the term completeness of a term is taken into account, e.g., terms with a definition, synonyms, and broader terms are ranked higher than terms without those additional information.

D. AutoSuggest

Goal of the AutoSuggest service is to offer relevant yet diverse terms to medical professionals as they enter data. The relevance of a search term is assessed using heuristic techniques taking into account the following ranking aspects:

1. A match with a label or synonyms (ignoring case)
2. A medical category match
3. A user input weighted position match: i.e., a match at the beginning of the first term is preferred to a match at the beginning of the following term, which is favored to a match within a term. For example a user's input is "war";
    - Highest match position boost: "Warfarin";
    - Second highest match position boost: "Venereal Warts";
    - Least match position boost: "Romanod-Ward Syndrom".
4. Term length: Terms with fewer words are preferred to terms with more words. For example for a user's input "war":
    - Highest length boost: "Warfarin";
    - Second highest length boost: "Venereal Warts";
    - Least length boost: "Warfarin Sodium 10 MG Oral Tablet".

Diversity, on the other hand, maximizes the number of categories of terms displayed. Relevance and diversity may be in conflict with each other. When displaying the *n* most relevant terms for a user input, then terms from only one or two categories may be displayed when *n* is small. On the other hand, if diversity is optimized, then less relevant terms may be favored over more relevant ones.

Various heuristic strategies can be applied to cope with those conflicting criteria. See Fig. 6.

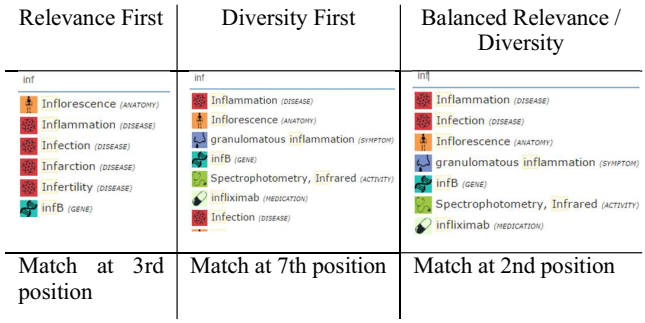


Fig. 6:AutoSuggest strategies

The AutoSuggest test is setup as a search for the term "Infection" while the user is typing "inf". All strategies display *n=7* results. The strategy "Relevance first" solely sorts according to the relevance rank as defined above. The strategy "Diversity first" alternates semantic categories in order to maximize category diversity. The strategy "Balanced Relevance / Diversity" uses a heuristic where the category with the highest ranks is displayed first with its top results. The amount of term slots per category in the result set is computed by a proportional assignment of the accumulated category ranks and *n*.

IV. Implementation

All concepts presented in the last section have been implemented within an EHR application for melanoma care. C# is used as the programming language for the backend and Javascript / HTML for the front-end. Apache Solr is used as search server. To communicate with Apache Solr, the solr.net client library is employed. Fig. 7 shows the integration of all modules into the existing application.

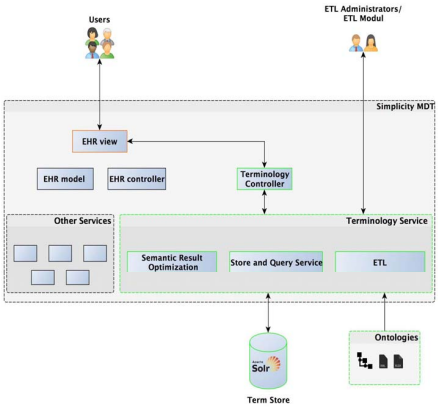


Fig. 7:System architecture

The implementation operates in two modes as indicated by the "Users" and "ETL Administrators/ETL Module" icons at the top of the diagram. Both modes are explained below.

Query Mode

- Input: search term, category weights (context) and total suggestion count
- Output: ordered list of terms.

See Fig. 8 for an UML sequence diagram.

1. The AutoSuggest service is invoked by the client.
2. The request is forwarded to the term store (search server).
3. The term store performs a lookup using partial matching on the term field as well as the synonym field.
4. Potential duplicates are removed.
5. The search result is semantically optimized, re-ordered according to the AutoSuggest strategy chosen.
6. Search terms are converted to the output format.

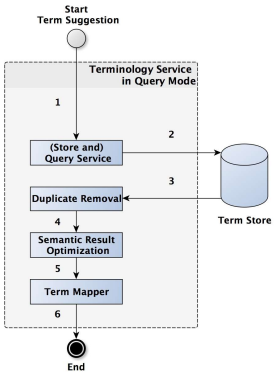


Fig. 8:Query mode

Extract, Transform, Load (ETL) Mode

- Input: none
- Output: duplicate-free set of terms

See Fig. 9 for an UML sequence diagram.

1. The ETL process is started.
2. All source configurations are loaded with individual parameters for the appropriate loader per ontology.
3. Each source configuration is invoked, exclude lists are applied and the mapping to the term class takes place.
4. All terms are semantically filtered.
5. Duplicate terms are removed.
6. – 8. The term set is serialized to a file.
9. – 11 The terms are stored in an Apache Solr index.

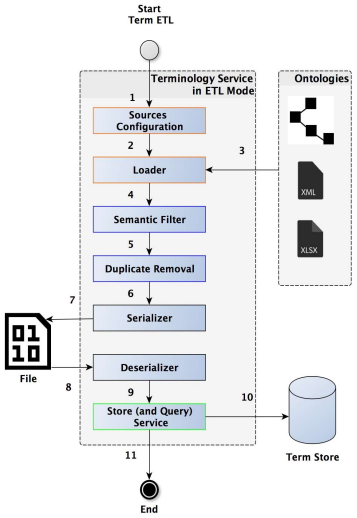


Fig. 9:Extraction, Transformation, Load

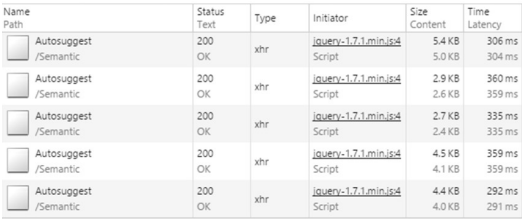
With a growing number of ontologies (and terms), the ETL processing time increases, making a full ETL unfavorable during application start. In the diagram shown in Figure 9, a file is displayed (step 7, 8). This file decouples the slower extract and transform part from the loading part, which has a requirement for quick response.

V. Evaluation

We compare the solution presented with the requirements specified in Section II.

1. – 3 Due to the ontology integration approach (ETL), the medical terminology is extensible. All relevant terms for the respective clinical use cases, in our case melanoma care, are covered. Using the semantic AutoSuggest service supports consultants editing EHRs by suggesting medical terms while typing.
4. Different term sorting strategies are provided, allowing to sort terms according to relevance but at the same time presents terms from diverse categories.

5. The meaning of terms is outlined by an explanation if such an explanation is provided in the source ontology.
6. The response time of the service is at typing speed. See Fig. 10 for latency measurements while typing the term "infection", which range from approximately 290 to 350 milliseconds.
7. First evaluations with leading cancer specialists in Ireland suggested the usefulness of the semantic AutoSuggest service. However, more thorough usage evaluations are needed.



Name / Path	Status / Text	Type	Initiator	Size / Content	Time / Latency
Autosuggest /Semantic	200 OK	xhr	jquery-1.7.1.min.js:4 Script	5.4 KB 5.0 KB	306 ms 304 ms
Autosuggest /Semantic	200 OK	xhr	jquery-1.7.1.min.js:4 Script	2.9 KB 2.6 KB	360 ms 359 ms
Autosuggest /Semantic	200 OK	xhr	jquery-1.7.1.min.js:4 Script	2.7 KB 2.4 KB	335 ms 335 ms
Autosuggest /Semantic	200 OK	xhr	jquery-1.7.1.min.js:4 Script	4.5 KB 4.1 KB	359 ms 359 ms
Autosuggest /Semantic	200 OK	xhr	jquery-1.7.1.min.js:4 Script	4.4 KB 4.0 KB	292 ms 291 ms

Fig. 10: Latency measurements of the semantic AutoSuggest service

VI. Related Work

There are numerous publications on AutoSuggest (a.k.a. autocomplete) services. For a good overview see [6]. However, traditional AutoSuggest services as in web search do not include semantics, e.g. category, synonyms, and description.

Hyvönen and Mäkelä describe in [7] various applications of semantic autocompletion in the domains of museums, media (video and audio), and yellow pages. As in our approach, hierarchy information from ontologies is used for categorizing terms to be suggested and synonyms are also used to suggest preferred terms. Our approach differs in a number of respects as we apply our approach using multiple ontologies in the domain of EHRs.

In [8], a semantic autocompletion service for medical terminology in a healthcare application is evaluated. As in our approach and in [7], ontologies are used, with the difference that commercial ontologies such as SNOMED CT are targeted. However, all those approaches are based on the hierarchy structure of a single ontology. From our experience, we agree with Bowker and Star when they state: "*Classifications that appear natural, eloquent, and homogeneous within a given human context appear forced and heterogeneous outside of that context*" [9]. This is why we introduce the semantic ontology integration (ETL) which not only allows to widen the terminology base but also to clean and prepare the terminology for the particular purpose of semantic AutoSuggest.

The concept of optimizing the conflicting criteria of relevance and diversity has been described in [10], but in a different domain (literature retrieval in a library) and a different context (facet recommendation for search refinement). To the best of our knowledge, this concept has not yet been applied to semantic AutoSuggest and not in the medical area.

VII. Conclusions and Future Work

We have presented a semantic AutoSuggest service that supports medical users who maintain EHR data by normalizing the use of medical vocabulary. Particular focus has been on the usability of the service optimizing relevance and diversity of suggestions.

The service has been implemented on top of a commercial EHR tool. After thorough user experience testing, it is intended to roll out the service to clinical end users. Possible future improvements include the personalized ranking of terms according to usage patterns in the EHR tool.

References

[1] Embi PJ, Payne PRO: *Clinical Research Informatics: Challenges, Opportunities and Definition for an Emerging Domain*. Journal of the American Medical Informatics Association : JAMIA. 2009;16(3):316-327.

[2] Yamamoto K, Sumi E, Yamazaki T, Asai K, Yamori M, Teramukai S, Bessho K, Yokode M, Fukushima M: *A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research*. BMJ Open 2012, 2:e001622.

[3] Prokosch HU, Ganslandt T.: *Perspectives for medical informatics*. Methods Inf Med 48.1 (2009): 38-44.

[4] Jensen PB, Jensen LJ, Brunak S: *Mining electronic health records: towards better research applications and clinical care*. Nature Reviews Genetics 13.6 (2012): 395-405.

[5] Liu L, Özsu MT, (Eds.). *Encyclopedia of Database Systems*. Springer US, 2009.

[6] Bast H, Weber I: *Type less, find more: fast autocompletion search with a succinct index*. ACM SIGIR. 2006:364–371.

[7] Hyvönen E, Mäkelä E: *Semantic autocompletion*. ASWC. 20062006:4–9.

[8] Sevenster M, Zharko A: *SNOMED CT saves keystrokes: quantifying semantic autocompletion*. AMIA Annual Symposium Proceedings. Vol. 2010. American Medical Informatics Association, 2010.

[9] Bowker GC, Star SL: *Sorting Things Out: Classification and its consequences*. Cambridge, MIT Press.1999

[10] Deuschel T, Greppmeier C, B, Humm BG, Stille W: *Semantically faceted navigation with topic pies*. Proceedings of the 10th International Conference on Semantic Systems (SEMANTiCS 2014), 2014.