An ontology-based method for secondary use of electronic dental record data

Titus KL Schleyer, DMD, PhD[1], Alan Ruttenberg, MS[2], William Duncan[2], Melissa Haendel, PhD[3], Carlo Torniai, PhD[3], Amit Acharya, BDS, MS, PhD[4], Mei Song, PhD[1], Thankam P. Thyvalikakath, MS, DMD, PhD[1], Kaihong Liu, PhD[1], Pedro Hernandez, DMD, MS[5]

[1]Center for Dental Informatics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA
[2]School of Dental Medicine, State University of New York at Buffalo, Buffalo, NY
[3]OHSU Library and Department of Medical Informatics, Oregon Health & Science University, Portland, OR
[4]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI
[5]Reparto Universitario, San Juan, PR

Abstract

A key question for healthcare is how to operationalize the vision of the Learning Healthcare System, in which electronic health record data become a continuous information source for quality assurance and research. . This project presents an initial, ontology-based, method for secondary use of electronic dental record (EDR) data. We define a set of dental clinical research questions; construct the Oral Health and Disease Ontology (OHD); anaylyze data from a commercial EDR database; and create a knowledge base, with the OHD used to represent clinical data about 4500 patients from a single dental practice. Currently, the OHD includes 213 classes and reuses 1,658 classes from other ontologies. We have developed an initial set of SPARQL queries to allow extraction of data about patients, teeth, surfaces, restorations, and findings. Further work will establish a complete, open and reproducible workflow for extracting and aggregating data from a variety of EDRs for research and quality assurance.

Introduction

Every day, patients ask their dentist questions. “How long will this filling last?,” “Why is my gum disease not getting better?” “Should I have an implant or a bridge?”. Gathering information prerequisite to answering these kinds of questions so it is available when needed, can be applied effectively and efficiently, and keeping it up-to-date as a part of the care process is a national priority[1].

How to operationalize the vision of the Learning Healthcare System is an important issue not only for medicine, but also for dentistry. Key questions include: What research questions can be answered by analyzing data in electronic dental records (EDR)? How can data from the multitude of current EDR systems be extracted in a standardized, reproducible, manner? How can dentistry create a cycle of continuous care improvement based on EDR data? And, finally, how can information gathered in dental visits be integrated into a Learning Healthcare System strategy to effectively contribute to total patient health[2]?

Typically, studies reusing electronic patient records extract data using a custom-developed, “one-off” mechanism, which is inefficient for data reuse on a broad scale[3]. Like electronic health records, most EDRs store information in a proprietary format. Extraction of data is impeded by several factors. First, incompatible database systems require idiosyncratic application programming interfaces to access data. Second, no two EDR databases are structured the same way. Third, even when the same kind of information is stored it may not be encoded in the same format, requiring conversions (for instance, when blood pressure measurements are stored as two integers vs. a single text string). Last, encodings may not map unambiguously to each other, such as when different EDRs record presence of “caries,” “root caries” and “incipient caries.”

We need a standardized approach that enables efficient access to information in EDRs and integration across different dental care providers and EDR systems. Our approach is to structure data from dental patient records using a realist approach. We interleave the construction of our Oral Health and Disease Ontology (OHD) with the re-encoding of the EDR data using the OHD. The OHD more directly represents what happens during dental visits. The OHD includes terms relevant to the diagnosis and treatment of dental maladies, and is publicly available[4,5]. Notably, we do not start from scratch. The OHD incorporates terms from a growing network of interoperable ontologies built using principles of the OBO Foundry[6]. In this paper, we report on initial efforts to represent dental patient data contained in an EDR and to build the supporting OHD. Our preliminary results describe a snapshot of the in-development ontology, selections from patient records represented using the OHD, and sample queries that retrieve relevant data. We conclude with a discussion of the benefits and challenges of our approach as concerns meaningful use of EDRs aggregated across practices, practice software, and with other sources of health information such as the EHR.

Methods

The data source for this project was a relational database of de-identified dental records for 7,337 patients from dental practice spanning the years 1999 – 2011, however only some 4500 had treatment records. The practice used Eaglesoft (Patterson Dental, Effingham, IL), which is one of the leading EDR systems in the US (18% market share). The database contained 232,270 records that pertained to patients' dental health history of which 54,000 dealt with restorative, endodontic and surgical procedures.

Our team is composed of an interdisciplinary mix of dentists, informaticians, ontologists and clinical dental researchers. We first developed a research question that we felt could be answered with the data. We then met a number of times to bring each of us to a reasonable level of understanding of the domain and common informatics issues. Subsequently we acquired basic familiarity with the database structure by reviewing vendor-supplied documentation in the form of sample queries and explanations of what they did. Our clinical dental researcher wrote a target spreadsheet format to make as concrete as possible the deliverable for our work. Once this was in place we worked iteratively to develop the framework presented in Figure 1, each iteration developing part of (3), (4), and (5) focused on one or a small number of entities involved in restoration and subsequent dental work (1). We have implemented all steps except statistical analysis (6).

Development of guiding dental clinical research questions

The purpose of developing these questions was to (1) enable studies of interest to general practitioners; (2) reuse a small selection of data commonly stored in EDRs; and (3) focus the development of the OHD on a clearly defined, tractable subset of clinical data. Questions included: What is the time from one restoration to its replacement on the same tooth? Does the time between successive restorations depend on the restorative material, such as amalgam and composite? What findings, e.g. caries and fracture, are present on a tooth over time and how do these relate to restorations (e.g. cause for placing the restoration)?

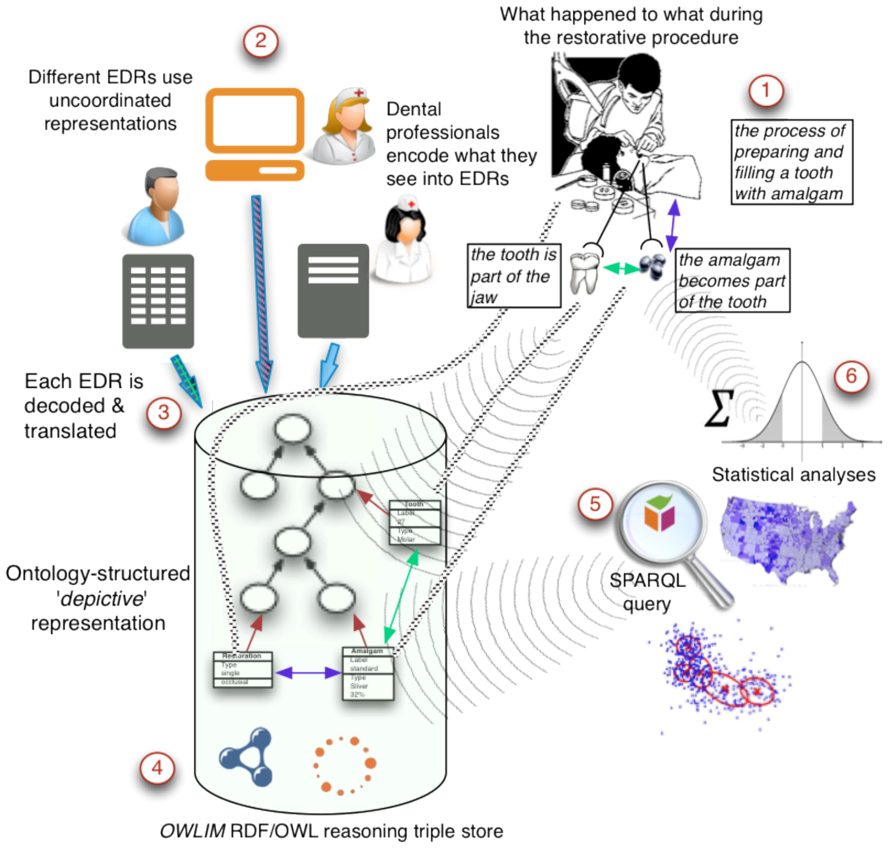
What happened to what during the restorative procedure
2
Different EDRs use uncoordinated representations
Dental professionals encode what they see into EDRs
1
the process of preparing and filling a tooth with amalgam
the tooth is part of the jaw
the amalgam becomes part of the tooth
Each EDR is decoded & translated
3
6
Σ
Statistical analyses
Tooth
Label
32
Type
Molar
5
SPARQL query
Ontology-structured 'depictive' representation
Restoration
Type
single
occlusal
Amalgam
Label
standard
Type
Silver
32%
4
OWLIM RDF/OWL reasoning triple store

Figure 1.Problem situation and solution architecture. What is seemingly a simple record of a cavity being filled with amalgam (1), is information that hard to aggregate from different EDRs (2). In our approach, we iteratively extract data out of EDRs (3) by writing scripts to translate data to an ontology-structured knowledge base (4). In each step we also develop queries (5) that incrementally provide the necessary data to do the statistical analysis (6).

First pass development of the Oral Health and Disease Ontology (OHD)

OHD terms were added as needed to represent entities involved in the procedures for which data was requested. Once a tentative definition was sketched we looked for superclasses in a subject of existing OBO ontologies that aspire to follow the OBO Foundry principles. Given our research question, the clinical processes of interest were restorative procedures, dental procedures that were indicative of failure of those restorations, and clinical examinations that produced relevant information. Participants were patients, their teeth, the surface layers of those teeth, and materials used for restorations. The primary information entities used were the CDT billing codes, as well as clinical findings generated during examinations.

Based on our analysis of information requirements, we identified the following ontologies for reuse and/or specialization: the Ontology for General Medical Science (OGMS)[7] (entities related to health care [patient role, visit, disorder, symptom]; the Foundational Model of Anatomy ontology (FMA)[8] and the Common Anatomy Reference Ontology (CARO)[9] (anatomical description of teeth, tooth surfaces, jaws, etc.); the Ontology for Biomedical Investigations (OBI)[10] (properties to relate processes to entities, for instance a restoration material to the restoration procedure); and the Information Artifact Ontology (IAO)[11] (CDT billing codes[12], clinical findings, the relationship between them and what happened, and provenance information regarding the development of the OHD). Therefore what we call the OHD is an aggregation of terms imported from other ontologies as well as terms our team defined as subclasses or specializations of those terms.

Extracting from and translating information in the EDR

Each step of this phase focused on translating information about a single kind of entity, e.g. patient, tooth, procedure, tooth surface layer, and relations that connected this entity to others, e.g. surface layers being part of teeth, teeth participating in (in the ontological sense) procedures

Although we were clear on what entities we needed information about, in many cases it was not straightforward to determine how that information was encoded. We supplemented our initial understanding of the database by reviewing tool generate documentation about stored procedures, triggers, relationships between tables, and the types of data in each table's field. When necessary, we consulted the vendor in order to understand the table structure, obtain SQL queries that would return data for the use cases of our interest (for instance how to retrieve all the findings and procedures for each tooth). Once we had a clear understanding of what entities and relationships the data represented, we would develop a computer programs to extract the data, preparing for the next step.

The ontology-based knowledge base we constructed consists of the OHD, instances of these classes, and relations among those instances. In each round of development, the data we extracted provided partial information, for example in one round we extracted patient information, including birthdate and gender. For example, each patient was to be represented by an instance of a gender–specific subclass, and related to their birth date. At this point we sometimes needed to add new terms to the OHD. After adding any necessary terms to the OHD, we translated the information we had retrieved from Eaglesoft into OWL statements, which were added to our knowledge base.

We developed our SPARQL queries in tandem with this process, attempting, in each iteration, to get closer to generating the information specified by our clinical researcher. In a number of cases, using actual patient data made us realize that aspects of this specification were unclear, underspecified, or that the underlying EDR could not supply the information requested, and this, in turn, would lead to adjustments of the specification. In this way, we constructed both the OHD and the OHD-structured representations in manageable increments, a process we continue as we further develop our work.

Results

Here we show the results of our first integration of data from an EDR system with the OHD ontology and some preliminary analysis of the data. At this time the OHD comprises roughly 150 classes whose URI are in the OHD namespace, about 200 CDT code classes (subset of the complete set), 12 classes from OBI (selected terms), all 82 OGMS classes, 14 selected terms from IAO, 1 term from NCBI taxonomy (Homo Sapiens), about 1500 terms from the FMA (all parts of the jaw and maxilla, dentition and tooth sockets), 3 terms from CARO, and all 32 terms from an early draft of BFO2, and about a dozen relations. Note that most of the classes created *de novo* were related to procedures, visits, exams and CDT codes that weren't available in existing biomedical ontologies.

The knowledge base provides the capability to query across all of the data in new and meaningful ways. For instance, we can query for the material used in filling restorations in a given practice during a certain time frame. Figure 3 shows the result for the material used between 1999 and 2011, grouped by year. Interesting here is that the overall number of restorations increased during 2003 - 2008 years, and that the percentage of amalgam was higher in the earlier years. Figure 4 is the SPARQL query used to retrieve the data for this chart.

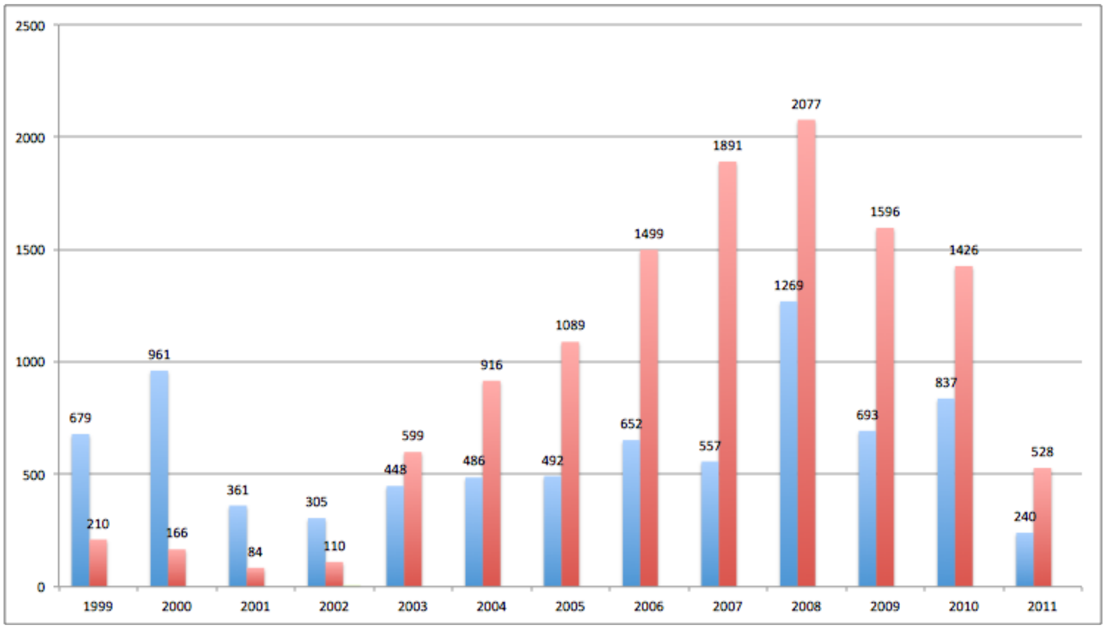

Figure 3. Material used in filling restoration per year. Red is Resin and blue is amalgam. Note that there was one case of gold in 2002.

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX obo:<http://purl.obolibrary.org/obo/>

PREFIX dental_restoration_material: <http://purl.obolibrary.org/obo/OHD_0000000>
PREFIX metal: <http://purl.obolibrary.org/obo/OHD_0000048>
PREFIX filling_restoration: <http://purl.obolibrary.org/obo/OHD_0000006>
PREFIX has_participant: <http://purl.obolibrary.org/obo/BFO_0000057>
PREFIX occurrs_on: <http://purl.obolibrary.org/obo/OHD_0000015>

SELECT DISTINCT ?material ?year (COUNT(?material) as ?m)
WHERE {
  ?procedure_instance rdf:type filling_restoration: .
  ?procedure_instance has_participant: ?material_instance .
  ?material_instance rdf:type ?material_type .
  ?material_type rdfs:subClassOf dental_restoration_material: .
  ?procedure_instance occurs_on: ?procedure_date .
  BIND (YEAR(?procedure_date) as ?year)
  ?material_type rdfs:label ?material .
FILTER (?materialtype != dental_restoration_material:)
FILTER (?materialtype != metal:)
}
GROUP BY ?material ?year
ORDER BY ?year ?material
```

Figure 4: SPARQL query to retrieve data for use in Figure 3. The query asks for filling restorations procedures and the date that they occurred, determining the material type by asking which participant in the procedure was a dental restoration material. This query takes advantage of the recently updated SPARQL 1.1. Prefixes are printed in gray to better see the query

To the best of our knowledge, the approach presented in our paper is the first that leverages Semantic Web Technologies for structuring and mining EDR data.

These preliminary results show the feasibility of extracting semantic structured data from an EDR system. The main advantages are the flexibility and the complexity of queries that can be performed and the advanced analysis of patient dental records that this will enable. In addition, having dental data semantically structured facilitates the integration of data coming from different EDRs and potentially EHRs and enables the usage of other semantic biomedical data (available for instance as Linked data [cite / link]) for data analysis.

An important contribution of our approach is related to the data quality and reliability. During the mapping process we have identified redundant, meaningless and even incorrect data (related, for instance, to having the proper value of some fields changed in order to facilitate display of the data such as adding a prefix). While our method won't fix all the possible issues with the source data, it can enhance the data quality by fixing duplication/errors, by identifying incorrect practices in data entry, and by removing redundant legacy data.

Our work also led to improvements of ontologies we have reused. Our observation that dentists were concerned with surface layers of teeth prompted the FMA to include such entities in a subsequent version. We also identified some errors due to the fact that maxillary dentition was asserted to be *part_of* secondary dentition (thereby also including primary maxillary dentition in secondary dentition). Identifying such errors and having them fixed in the source ontologies benefits all others using these ontologies.

Together with these benefits, the proposed approach poses also some challenges. First of all there is the difficulty of interpreting the structure and the content of the data sources. The involvement of the Eaglesoft vendor personnel and of a dental practice was fundamental in order to ensure that the source data were translated correctly into our knowledge base. The same level of involvement should be considered while replicating our approach for a different vendor / practice combination. Still, by having our work be open source we hope to make it possible that such effort is done once, instead of each time a different group wants access to such data.

Another challenge identifying and properly translating into our knowledge base events or findings that rely on what might be missing data, or on complex patterns of findings or procedures. For instance, most of the time it is not possible to characterize if a "missing tooth" finding on a patient's first visit being the result of a concurrent extraction or having not initially formed or having been lost due to advanced periodontal disease.

Our immediate plans are to do more extensive analysis on the data, using a SPARQL extension to R to enable our dental clinical researchers to answer more complex clinical questions without the assistance of an intermediary. We also plan to make available de-identified data for the analysis result as Linked Data, and to continue developing OHD and translation methods until the full content of the EDR is available.

Acknowledgements

We would like to acknowledge Dan Caplan, who worked with us to define analysis requirements, and Jim Garrett of Eaglesoft who provided his time and expertise to help us understand the database structure. Michael Breene and Gunther Kohn provided advise and background on EDR implementation. Michael Dziabiak, assisted with editing.

References

1. Olsen L, Aisner D, McGinnis JM, editors, Medicine RoE-b. The learning healthcare system: workshop summary. Washington, D.C.: The National Academis Press; 2007.
2. Powell V, Din FM, Acharya A, Torres-Urquidy MH, eds. Integration of medical and dental care and patient data. New York: Springer; 2011=In print.
3. Pearson JF, Brownstein CA, Brownstein JS. Potential for electronic health records and online social networking to redefine medical research. Clin Chem. 2011/2;57(2):196-204.
4. Oral Health and Disease Ontology. [cited 2012 9/28]; Available from: http://purl.obolibrary.org/obo/ohd/dev/ohd.owl.
5. Browsable version of Oral Health and Disease Ontology. [cited 2012 9/28]; Available from: http://purl.obolibrary.org/obo/ohd/browse.
6. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007/11;25(11):1251-5.
7. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. Summiton TranslatBioinforma. 2009;2009:116-20.
8. Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003/12;36(6):478-500.
9. Haendel MA, Neuhaus F, Osumi-Sutherland D, Mabee PM, Mejino JLVJ, Mungall CJ, et al. CARO - The common anatomy reference ontology. In: Burger A, Davidson D, Baldock R, editors. London: Springer; 2008. p. 327-50.
10. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, et al. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010;1 Suppl 1:S7-.
11. The information artifact ontology. [cited]; Available from: http://purl.obolibrary.org/obo/iao.owl.
12. Association AD. CDT 2011-2012: current dental terminology: the ADA practical guide to dental procedure codes. Chicago: American Dental Association; 2010.