future internet

MDPI

Article

Merging Ontologies and Data from Electronic Health Records

Salvatore Calcagno [1], Andrea Calvagna [2], Emiliano Tramontana [2,*] and Gabriella Verga [2]

[1] Cyberetna, 95024 Catania, Italy
[2] Dipartimento di Matematica e Informatica, University of Catania, 95125 Catania, Italy; andreamario.calvagna@unict.it (A.C.); gabriella.verga@unict.it (G.V.)
* Correspondence: tramontana@dmi.unict.it; Tel.: +39-095-7383008

Abstract: The Electronic Health Record (EHR) is a system for collecting and storing patient medical records as data that can be mechanically accessed, hence facilitating and assisting the medical decision-making process. EHRs exist in several formats, and each format lists thousands of keywords to classify patients data. The keywords are specific and are medical jargon; hence, data classification is very accurate. As the keywords constituting the formats of medical records express concepts by means of specific jargon without definitions or references, their proper use is left to clinicians and could be affected by their background, hence the interpretation of data could become slow or less accurate than that desired. This article presents an approach that accurately relates data in EHRs to ontologies in the medical realm. Thanks to ontologies, clinicians can be assisted when writing or analysing health records, e.g., our solution promptly suggests rigorous definitions for scientific terms, and automatically connects data spread over several parts of EHRs. The first step of our approach consists of converting selected data and keywords from several EHR formats into a format easier to parse, then the second step is merging the extracted data with specialised medical ontologies. Finally, enriched versions of the medical data are made available to professionals. The proposed approach was validated by taking samples of medical records and ontologies in the real world. The results have shown both versatility on handling data, precision of query results, and appropriate suggestions for relations among medical records.

Keywords: Electronic Health Records; HL7; ontology; data integration; data analysis

check for updates

Citation: Calcagno, S.; Calvagna, A.; Tramontana, E.; Verga, G. Merging Ontologies and Data from Electronic Health Records. *Future Internet* **2024**, *16*, 62. https://doi.org/10.3390/fi16020062

Academic Editor: Petros Patias

Received: 13 January 2024
Revised: 5 February 2024
Accepted: 6 February 2024
Published: 17 February 2024

1. Introduction

The Electronic Health Record (EHR) can be seen as the natural evolution of the physical medical record, which consisted of hard-copy medical documents for a patient. Nowadays, electronic versions of medical records allow citizens to easily track the entire history of their healthcare life and share it with healthcare professionals. Moreover, doctors are better equipped to keep up with the large amount of data, more easily access them, and deliver a significant improvement in health services for patients. To facilitate the syntactic analysis of such data and some level of interoperability between different software systems documents follow a standard format: Health Level 7 (HL7) Clinical Document Architecture (CDA) [1,2]. Such a standard is a valuable tool for data sharing; however, its adoption has been slow due to the length and complexity of the defined formats. In fact, given the wide spectrum and complexity of the medical field, several templates have been defined within HL7, one for each type of clinical document (http://cdasearch.hl7.org, last accessed on 10 November 2023).

Documents produced according to HL7 present two critical issues that derive from its formats: (i) complexity in the organisation of data, which makes it difficult to read data, i.e., navigating the formats to extract data is a complex task, and this can be a significant impediment when developing software systems that extract data from an EHR [2]; (ii) though the terms are very specific, they present only the essential pieces of information, without any hints to other related terms or factors, such as, e.g., related diseases, symptoms, possible causes, etc. Moreover, pieces of data are scattered in the document, without evident relationships among them. There is a need for a way to organise data in an EHR to make it possible to discover and suggest potential diseases, causes, etc., related to the medical values and conditions.

To organise knowledge in a domain, some previous approaches have attempted to build ontologies automatically [3]; however, in the medical field it is paramount to have a highly reliable ontology curated by experts. A significant number of well-established ontologies exist in the biomedical realm (http://obofoundry.org, last accessed on 10 November 2023). Although ontologies have been used to guide the diagnosis process, they have not been related to EHR data [4].

This article proposes a solution to greatly mitigate the above two critical issues. Firstly, we propose to organise medical data originating from an EHR by using a more manageable format based on JavaScript Object Notation (JSON), through the embedding of original HL7 terms, which are the keywords characterising data held within a medical record. For this, a tool that automatically extracts data from the original HL7 format has been proposed and developed. This is not an easy task, due to the complexity of the several EHR formats. Secondly, to relate data in an EHR to their context, we propose to integrate extracted data and the most significant ontologies in the medical realm. The outcome is an enriched ontology, as well as a set of data, which provides more easily searchable data and assists doctors in their work while studying a patient's condition, writing new medical records, or when producing a diagnosis.

The EHR documents and ontologies employed in the development of the approach and in the tests performed to validate it actually derived from realistic medical data, that is, documents extracted from the official public repository of the HL7 organisation, whereas the main ontology employed was the Human Disease Ontology (HDO) [5], written by experts in the field and widely used.

This paper is organized as follows. Section 2 illustrates the related works. Section 3 discusses the EHR, existing standards with emphasis on HL7, and a description of the ontologies used to connect the medical data of patients. Section 4 describes the proposed devised approach. Section 5 contains the results obtained. Finally, conclusions are drawn in Section 6.

2. Related Work

In recent years, the high complexity of clinical documents has been largely recognised [6], and the importance of a correct interpretation of data for an appropriate diagnosis has been also highlighted [7–9]. Moreover, there exist some gaps in the representation of symptoms and clinical data in the structure of EHR clinical documents, because the most significant descriptions of conditions remain in unprocessed free text [10–12]. Indeed, the importance of sharing and using EHRs has also been shown [13].

JSON is a standard text-based format for representing structured data based on JavaScript object syntax. It is commonly used for transmitting data in web applications (e.g., sending some data from the server to the client, so it can be displayed on a web page, or vice versa). Many researchers have highlighted that JSON is a notation that is gradually replacing XML due to its relative simplicity, intuitiveness, compactness, and the ability to directly map the native data types of popular programming languages [14,15].

A first prototype conversion tool of clinical documents to a JSON format was introduced by Rinner et al. [16]. The goal of their work was to allow easier access to health data stored in the Austrian information system for EHR (ELGA) and convert CDA documents into Fast Healthcare Interoperability Resources (FHIR) resources, via XML. The limitation of such a proposal is that it implements a basic prototype with limited capabilities for converting from the XML document to an equivalent JSON file. Moreover, they do not interpret or enrich the processed data.

Rousseau et al. [15] aimed at providing a more complete clinical document to facilitate doctors' work by integrating Social Determinants of Health (SDoH) into EHRs, leveraging for this a specific ontology. Specifically, they combine the HL7 standard with the SNOMEDCT ontology (https://bioportal.bioontology.org/ontologies/SNOMEDCT, last accessed 10 January 2024), using JSON as a means to better connect data in medical records to external data sources. The limitation of this approach is that it was not designed to allow connecting multiple ontologies like our approach. Indeed, the authors concluded that the research and development teams of health IT and healthcare services strive to make information in existing semantic frameworks more accessible and connected to EHRs. However, they left the integration of data from other ontologies as future work.

Ontologies are among the most effective and most widely adopted means to represent knowledge in the medical (or any other) domain [17–20]. Several authors (i.e., [21]) have been using them to facilitate medical decision-making, e.g., to suggest diagnoses and treatments, also taking advantage of probabilistic algorithms applied to the ontology data. However, successfully processing clinical data to navigate complex ontologies down to a possibly correct diagnosis is not at all an easy task to accomplish, because probabilistic algorithms applied to the rigid complex structure of an ontology can fail when the clinical data are poorly specified or relate to little-known cases. Other authors [22–25] have been more concerned with expanding the domain of medical knowledge through an accurate review of the related concepts, e.g., the analysis of disease ontology terms or IDs (DOIDs), highlighting the importance of an enriched and complete ontology for the analysis of diseases. However, in our approach, the aim is to automatically create a connection between HL7 and ontologies as, e.g., DOIDs, and not to change ontologies, as they were carefully developed by experts.

A further obstacle is the complexity of the HL7 CDA standard itself, whose data format is very widespread but also difficult to manage and to apply in the actual daily routine [16]. To our knowledge, this paper is the first effort to improve the practical usability of clinical data in HL7 documents by truly enriching their semantics while also improving their syntactic clarity. Specifically, this is the first approach processing the HL7 XML formatted clinical documents in order to produce a single, both enriched and easier to understand (and process), JSON object. This allows EHRs to go beyond the mere need for storing and exchanging clinical records towards a new conception as data objects that can be really effective and practical tools supporting clinicians' everyday work.

In fact, the resulting medical document offers a clearer and more complete overall clinical picture of the patient since it is able to integrate the data in its medical records with data from multiple related medical ontologies. As a consequence, it facilitates the analysis of the patient by being able to provide the definitions of the terms and the relationships among all the basic pieces of clinical data, assisting clinicians in an in-depth analysis of the patient and the construction of a correct diagnosis. Moreover, having to manage a single JSON-formatted document object is of great added value not only because it facilitates its integration into multiple frameworks or software but also because it is more user friendly for automated processing and exchange.

3. Background

3.1. Electronic Health Record

Starting from an academic idea in the late 1980s, EHRs have evolved to become the centre of national health information strategies in most European countries [26], including Italy, and also outside Europe. The term EHR was officially introduced in Italy in the national legislation of Article 12 of Law 221/2012 [27].

In the real world, the potential of digital medical records even today is often not yet fully understood nor are the benefits for citizens (who can have a complete overview of their medical history at hand), clinicians (who manage to organise themselves more efficiently), and medical centres/clinics (which obtain both economic and time savings in the management of clinical work). For better sharing data between all the involved partners in the medical sector, i.e., patients, doctors, nurses and other involved professionals at any level, standards need to be met in order to minimise the effort.

The major standards for representing EHRs are as follows [28].

1. OpenEHR: a framework of standards for health care and medical research. Its mission is to facilitate the creation and sharing of health records by patients to clinicians via open-source, standards-based implementations [29].
2. HL7: Founded in 1987, HL7 is an ANSI accredited standards development non-profit organisation dedicated to providing a comprehensive framework and related standards for exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery, and evaluation of health services. HL7 is supported by over 1600 members from over 50 countries, including over 500 corporate members representing healthcare professionals, government stakeholders, taxpayers, pharmaceutical companies, vendors/suppliers, and consulting firms.
3. CEN TC251 EN 13606: standardisation in the field of health information and communication technologies (ICT) to achieve compatibility and interoperability between independent systems and allow modularity [30]. This includes health information structure requirements to support clinical and administrative procedures and technical methods to support interoperable systems; as well as safety, security and quality requirements.

3.2. Health Level Seven (HL7)

HL7 association deals with the management of health standards and makes documents available in the HL7 format. Such a format allows for the analysis and study of the various clinical documents. Hence, the term HL7 is used both as a name for the organisation and as a set of messaging standards [31]. It is probably the most widely used standard in the world for the electronic exchange of clinical information and it was created with the aim of providing a standard for the exchange, integration, sharing, and retrieval of electronic health information. For several years, the HL7 standard has been also criticized, highlighting its complexity, inconsistency, the documentation, and the high cost of integration with other existing software [32,33].

In this study, we focused on the analysis of medical documents in the CDA standard, which is an XML-based markup standard specifying the encoding, structure, and semantics of clinical documents for exchange, issued by HL7 in year 2000 [34]. A CDA document is an XML document that consists of a header and body. It is presented in this format:

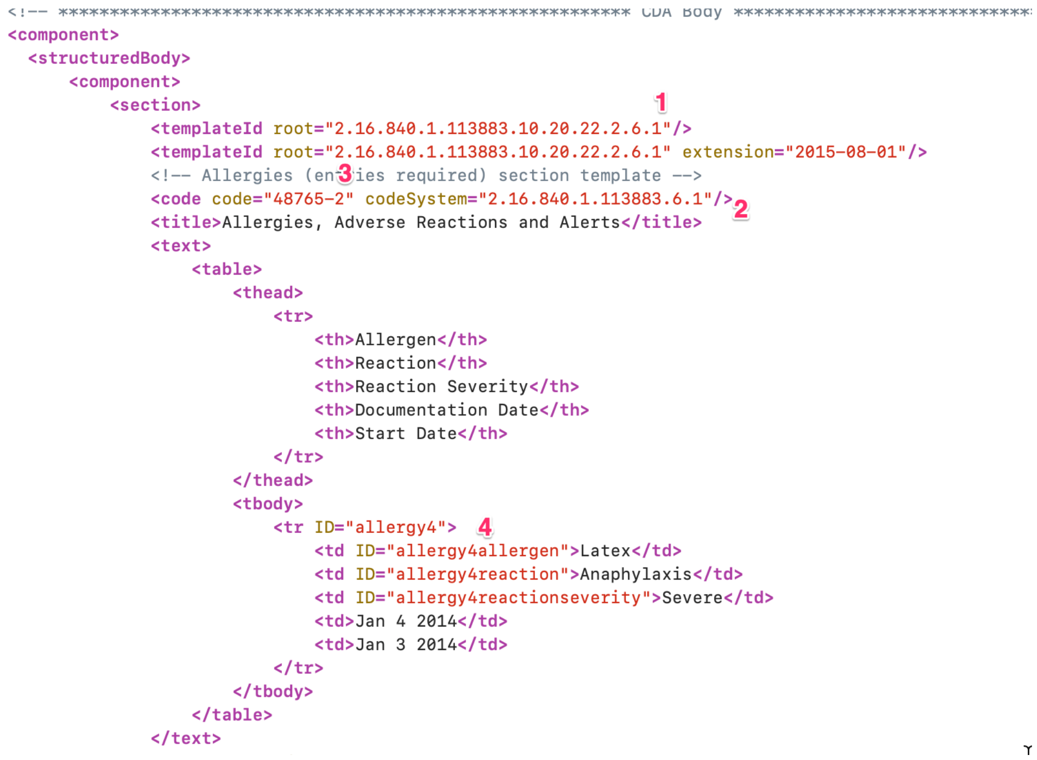
- Header: includes patient information, author, creation date, document type, provider, etc.
- Body: includes clinical details, diagnosis, medications, follow-up, etc. Presented as free and/or structured text in one or multiple sections and may optionally include coded entries, too. Due to the different sections containing different information, each document has a different body and tag which makes its consultation complex.

Among the main existing sections (https://github.com/HL7/C-CDA-Examples/tree/master, last accessed on 10 November 2023), there are the following:

1. Allergies Document: type that contains all the information on the allergies detected in the patient and any adverse reactions;
2. Encounters Document: type that contains information relating to the meetings sustained by the patient with the various health professionals;
3. Family History Document: type that contains information relating to the pathologies diagnosed in the patient's family members;
4. Immunizations Document: type that contains information relating to the vaccines carried out by the patient;
5. Interventions Document: type that contains the details relating to all the interventions carried out by the patient;
6. Problems Document: type that contains all the information relating to the health problems diagnosed to the patient.

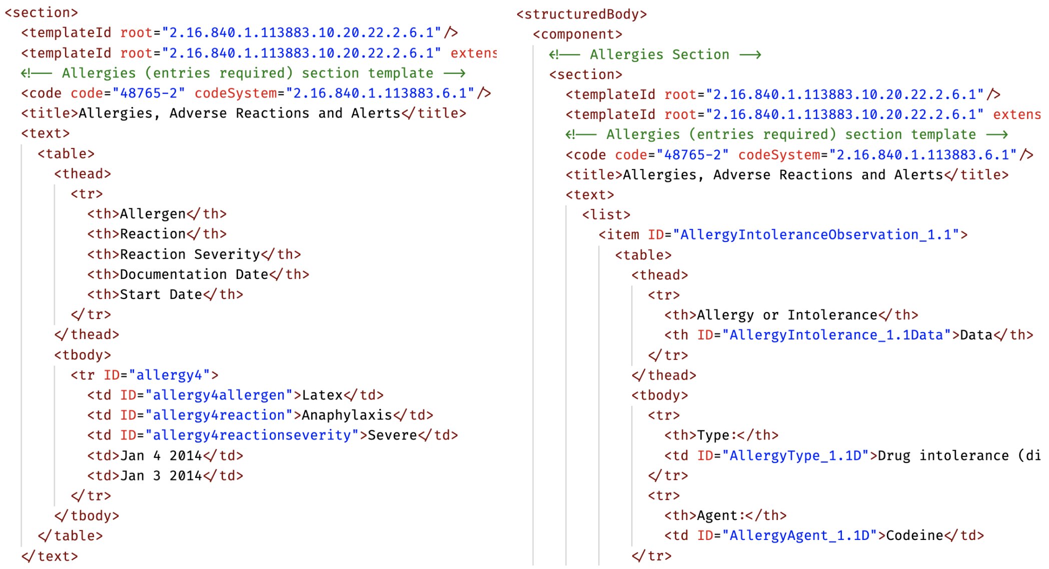
A CDA XML document is made up of several sections, each differing from another according to the medical field treated. For each XML tag present, it is possible to find attributes relating to codes of external dictionaries and/or tags encoding how to read and understand each section that has been populated; it follows that such documents are particularly complex to generate and read.

Figure 1 shows a fragment of XML code in a CDA document in which its inner structural complexity can be easily observed. In point (1): the "root" tag indicates the unique identifier of a populated section, following immediately below. Each section differs in the type of its contents and the standard used to populate it. In fact, in order to become acquainted with the actual content of the various sections, you need to access and download all the used types of data structures' definitions, available from the remote url: https://hl7.org/cda/stds/ccda/draft1/downloads.html, last accessed on 10 January 2024. Point (2), and specifically the "codeSystem" attribute within the "code" tag, shows which dictionary of codes has been used for the labels present in the section, which in this case is the Logical Observation Identifiers Names and Codes (LOINC) dictionary. LOINC dictionary terms definitions have to be accessed separately, which can be performed in one of two different ways: offline, by downloading the entire dictionary and carrying out local searches, or online using the appropriate remote search page, etc. (https://loinc.org/, last accessed on 10 January 2024). At point (3): the "code" tag value encodes the subsection used within the LOINC dictionary. In point (4), we find a tabular list of actual key–value data pairs and their "ID" attributes, which represent custom defined identifiers that can be used as references to this data from other parts of the CDA file. Additionally, another source of undesired intricacy in CDA documents is actually the way the XML data themselves have been organised at user compilation time. This is clearly shown in Figure 2 which confronts data belonging to the same type of section (allergies) however filled in very different ways, which leads to very apparently different documents, as they have different structures despite the fact that their tags carry the same information once correctly interpreted.



```
<!-- ********************************************************* CDA Body ******************************
<component>
  <structuredBody>
      <component>
          <section>
              <templateId root="2.16.840.1.113883.10.20.22.2.6.1"/>
              <templateId root="2.16.840.1.113883.10.20.22.2.6.1" extension="2015-08-01"/>
              <!-- Allergies (entries required) section template -->
              <code code="48765-2" codeSystem="2.16.840.1.113883.6.1"/>
              <title>Allergies, Adverse Reactions and Alerts</title>
              <text>
                  <table>
                      <thead>
                          <tr>
                              <th>Allergen</th>
                              <th>Reaction</th>
                              <th>Reaction Severity</th>
                              <th>Documentation Date</th>
                              <th>Start Date</th>
                          </tr>
                      </thead>
                      <tbody>
                          <tr ID="allergy4">
                              <td ID="allergy4allergen">Latex</td>
                              <td ID="allergy4reaction">Anaphylaxis</td>
                              <td ID="allergy4reactionseverity">Severe</td>
                              <td>Jan 4 2014</td>
                              <td>Jan 3 2014</td>
                          </tr>
                      </tbody>
                  </table>
              </text>
```

Figure 1. Snippet of XML code displaying intrinsic syntactic complexity despite being extracted from a simple, basic example of CDA document. Points marked 1 to 4 highlight the following tags: root, codeSystem, code, and ID, respectively.



```
<section>
  <templateId root="2.16.840.1.113883.10.20.22.2.6.1"/>
  <templateId root="2.16.840.1.113883.10.20.22.2.6.1" extens
  <!-- Allergies (entries required) section template -->
  <code code="48765-2" codeSystem="2.16.840.1.113883.6.1"/>
  <title>Allergies, Adverse Reactions and Alerts</title>
  <text>
    <table>
      <thead>
        <tr>
          <th>Allergen</th>
          <th>Reaction</th>
          <th>Reaction Severity</th>
          <th>Documentation Date</th>
          <th>Start Date</th>
        </tr>
      </thead>
      <tbody>
        <tr ID="allergy4">
          <td ID="allergy4allergen">Latex</td>
          <td ID="allergy4reaction">Anaphylaxis</td>
          <td ID="allergy4reactionseverity">Severe</td>
          <td>Jan 4 2014</td>
          <td>Jan 3 2014</td>
        </tr>
      </tbody>
    </table>
  </text>
```

```
<structuredBody>
  <component>
    <!-- Allergies Section -->
    <section>
      <templateId root="2.16.840.1.113883.10.20.22.2.6.1"/>
      <templateId root="2.16.840.1.113883.10.20.22.2.6.1" extens
      <!-- Allergies (entries required) section template -->
      <code code="48765-2" codeSystem="2.16.840.1.113883.6.1"/>
      <title>Allergies, Adverse Reactions and Alerts</title>
      <text>
        <list>
          <item ID="AllergyIntoleranceObservation_1.1">
            <table>
              <thead>
                <tr>
                  <th>Allergy or Intolerance</th>
                  <th ID="AllergyIntolerance_1.1Data">Data</th>
                </tr>
              </thead>
              <tbody>
                <tr>
                  <th>Type:</th>
                  <td ID="AllergyType_1.1D">Drug intolerance (di
                </tr>
                <tr>
                  <th>Agent:</th>
                  <td ID="AllergyAgent_1.1D">Codeine</td>
              </tr>
```

Figure 2. Comparison of two XML files with different structures but the same type of content.

3.3. *Human Disease Ontology (HDO)*

HDO [5] was developed as a standardised ontology for human disease with the aim of providing the biomedical community with consistent, reusable, and sustainable vocabulary of human diseases, including terms, descriptions, phenotype characteristics, and related medical concepts. The ontology represents over 8000 diseases that can be found through names, synonyms, and definitions. Figure 3 shows the main classifications of diseases (on the left) and next to it an example of the branching of the disease caused by an infectious agent (on the right).

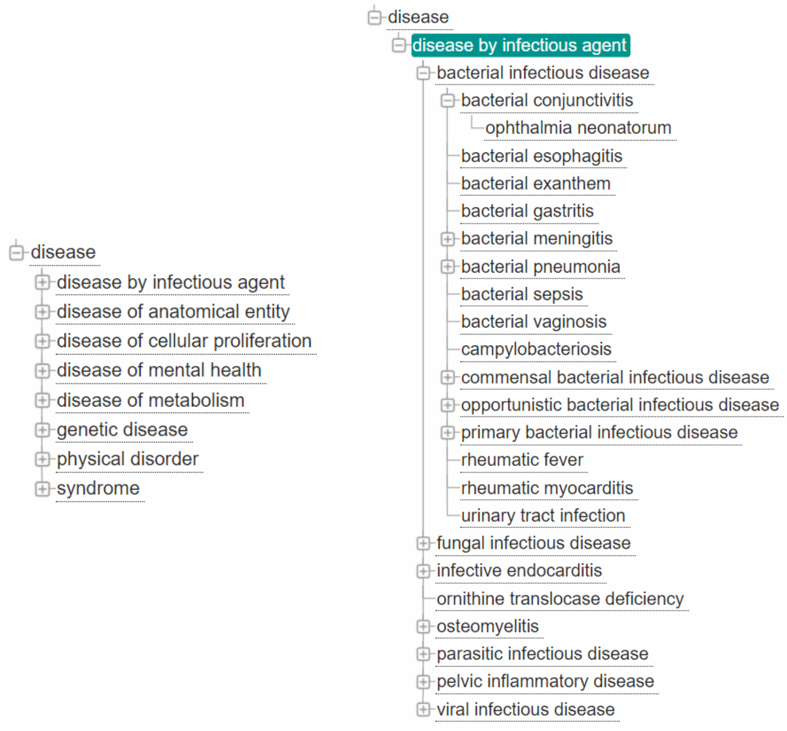


Figure 3. On the **left**, the main branches of the Human Disease Ontology and on the **right** some classes that derive from the first disease, or disease by infectious agent.

4. Description of Our Approach

Given the complexity of a document in HL7 CDA format, a tool has been implemented that converts an HL7 CDA document from XML format into a JSON format. This innovative process allows for a more comfortable, easy-to-read clinical document, which can also be easily integrated in other medical software processing frameworks. In fact, thanks to its much more concise syntax and its well established semantics, JSON is nowadays often a preferred data format over XML. It is then easier to work with the new obtained JSON data representation in various ways. In this work, we propose to merge the obtained JSON data with additional knowledge extracted from related domain ontologies to offer a more complete and detailed document that is easier to read and understand due to the use of a more explicit and broader vocabulary. The end result is a clearer and more documented clinical picture of the patient, giving clinicians stronger support to make a diagnosis.

The implemented approach consists of three phases, discussed in detail in the following.

1. Data acquisition: medical records in XML format are processed for conversion into JSON format.
2. Ontological mapping: the obtained JSON keywords are analysed and mapped to the corresponding semantic definitions found in related ontologies, where the meaningful keywords allowing the mapping are derived from a data cleaning and analysis subprocess.
3. Document enrichment: the semantic information is integrated with the JSON medical record to obtain a more informative document.

To test the proposed approach, a real medical data repository was used, which was created and publicly shared by the HL7 organisation to allow the validation and benchmarking of medical applications with simulated but realistic clinical documents (https://github.com/HL7/C-CDA-Examples/tree/master, last accessed on 10 November 2023). The proposed document processing and integration approach has been implemented in a tool, entirely developed in the Java language. Below, a snippet of the main algorithm is shown which implements the high-level steps carried out during the processing.

```
ClinicalDocument clinDocument = DocumentParser.buildClinicalDocument(xml);
List<Keyword> jsonKeys = clinDocument.extractKeys();
List<Label> labelsList = OntologyManager.retrieveLabels();
List<Label> similarLabels = clinDocument.textSimilarity(jsonList, labelsList);
List<OntLabelDetails> infoLabels;
for (Label label : similarLabels){
    infoLabels.add(OntologyManager.retrieveData(label));
}
clinDocument.dataUnion(jsonList,similarLabel,infoLabels);
```

Classes and methods in the above code snippet are described in the following.

- ClinicalDocument: a class representing a JSON object that has been built up by transforming an input CDA XML file.
- DocumentParser: a class that deals finds and extract data from the XML in a CDA.
- OntLabelDetails: a class holding data belonging to an ontology class, such as definition and synonyms.
- buildClinicalDocument(): a method that reads the CDA XML and processes it. Given an id for the XML file as input, it calls all the methods for checking and reading the file according to its tree structure. The tree is read by starting from the root, looking for all the sections to export, which are found by means of the section tag. For each section tag found, the corresponding XML text is passed as input to appropriate methods and that will extract text data and encapsulate them in instances of appropriate Java classes. After reading all the tags, all the created instances are represented as an object of class ClinicalDocument, returned to the caller.
- extractKeys(): a method that extract keywords from JSON objects.
- retrieveLabels(): a method that extract labels from an ontology.

- textSimilarity(): a method that matches JSON keywords and ontology labels, For each pair of values consisting of a value taken from the JSON and one from the ontology, a similarity value is computed. Then, for the pairs having a similarity score greater than a threshold (which in our experiments was 0.7), the corresponding label is included in the list returned as a result from the method.
- retrieveData(): a method that finds and returns data corresponding to a given label.
- dataUnion(): a method that builds a comprehensive JSON document aggregating all original CDA data and explanations found in the ontology.

4.1. Data Acquisition

CDA XML file documents representing patients' records are converted into a more readable JSON format that is easier to analyse and later integrate into different frameworks than its originating one. A tool has been developed to generate JSON files starting from the XML file in HL7 CDA format given in input, as a micro-service in Spring-Boot using the MVC pattern (https://spring.io/guides/gs/spring-boot/, last accessed on 10 November 2023). The JSON file contains a JSON data object which holds all the information present in the input XML document, e.g., the patient's personal data, vaccines carried out, etc. The javax.xml and org.w3c libraries were used to parse the whole XML file, thus handling a tree structure. While reading the entire XML for each identified node, known section tags were searched. For example, when tags like 48765-2 or 10157-6 were found (see Figure 1), which are the roots of different and articulated sections of medical data structures, then the respectively dedicated parsers are started to process the corresponding data.

For each analysed clinical document (as discussed in Section 3.2), an instance of a specialised Java class is set up in order to represent the general structure of the clinical document, i.e., it contains (i) the basic information to define the document and (ii) a reference to the patient to whom it has been dedicated, with general information: name, surname, email, etc. This class will also be the main container of a set of specialised data objects corresponding to any specific clinical data sections of the original CDA document.

Listings 1 shows an example of XML data excerpted from a sample medical document and Listing 2 the correspondent snippet of JSON conversion output. Similarly, Listings 3 and 4 are an XML portion of a medical document and its JSON equivalent.

Listing 1. Example 1: XML file part of a clinical document related to patient data.

```xml
<code code="BRO"
 codeSystem="2.16.840.1.113883.5.111"
 codeSystemName="HL7 RoleCode"
 displayName="Brother">
        <originalText>
            <reference value="#FamHist1rel"/>
        </originalText>
</code>
<subject>
        <sdtc:id extension="98765432-1"
         root="1.3.6.1.4.1.16517.1"
         xmlns:sdtc="urn:hl7-org:sdtc" />
        <name>James</name>
        <administrativeGenderCode code="M"
         codeSystem="2.16.840.1.113883.5.1" />
        <birthTime value="1982" />
</subject>
```

Listing 2. JSON object extracted from the XML of Example 1.

```json
{
    "name": "James",
    "birthTime": "1982",
    "relationship": "Brother",
    "deceased": false,
    "deceasedTime": null,
}
```

Listing 3. Example 2: XML file part of a clinical document relating to a visit made.

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/xsl" href="CDA.xsl"?>
<ClinicalDocument xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="urn:hl7-org:v3" xmlns:voc="urn:hl7-org:v3/voc"
xmlns:sdtc="urn:hl7-org:sdtc">
  ... <!-- list of tags related to the xml encoding codes  -->
  <recordTarget>
    <patientRole>
        <id extension="444222222" root="2.16.840.1.113883.4.1" />
        <telecom value="tel:+1(555)555-2003" use="HP" />
        <patient>
            <name use="L">
                <given>Eve</given>
                <family qualifier="SP">Betterhalf</family>
            </name>
            <administrativeGenderCode code="F" displayName="Female"
            codeSystem="2.16.840.1.113883.5.1"
            codeSystemName="AdministrativeGender" />
            <birthTime value="19750501" />
            <maritalStatusCode code="M" displayName="Married"
            codeSystem="2.16.840.." codeSystemName="MaritalStatusCode" />
            <religiousAffiliationCode code="1013"
            displayName="Christian (non-Catholic, non-specific)"
            codeSystem="2.16.840.1.113883.5.1076"
            codeSystemName="HL7 Religious Affiliation" />
            ... <!-- patient personal information -->
        </patient>
    </patientRole>
  </recordTarget>
  <component>
  <structuredBody>
    <component>
      <section>
        <templateId root="2.16.840.1.113883.."/>
        <templateId root="2.16.840.1.113883.." extension="2015-08-01"/>
        <code code="48765-2" codeSystem="2.16.840.1.113883.6.1"/>
        <title>Allergies, Adverse Reactions and Alerts</title>
        <text>
            <table>
              <thead><tr>
                <th>Allergen</th><th>Reaction</th><th>Reaction Severity</th>
                <th>Documentation Date</th><th>Start Date</th>
              </tr></thead>
              <tbody><tr ID="allergy2">
                <td ID="allergy2allergen">Penicillin</td>
                <td ID="allergy2reaction">Anaphylaxis</td>
                <td ID="allergy2reactionseverity">Severe</td>
                <td>Jan 4 2014</td>
                <td>2006</td>
              </tr></tbody>
            </table>
         </text>
        <entry>
        ... <!-- more details of the section, for~example information about the doctor
     who performed it-->
        </entry>
        </section>
    </component>
   </structuredBody>
  </component>
</ClinicalDocument>
```

4.2. *Ontological Mapping of Clinical Data*

Each clinical document in the EHR of a given patient describes specific clinical information of that patient, e.g., their allergies or vaccines. Thus, once a new CDA document has been processed for data acquisition, a set of labels has been already extracted from it. Those are standardized medical keywords used to unambiguously refer to specific diseases, symptoms, drugs, allergic reactions, etc., that describe the medical story of the patient: the reason for the check, the name and type of the problem encountered, its severity, reactions, date of examinations, etc. Such data can be further enriched with the use of appropriate ontologies. Specifically, we match data extracted from the CDA document with appropriate semantic data extracted from related existing medical ontologies, facilitating doctors' work, e.g., by offering them readable descriptions of the keyword terms and advice and/or suggestions on further related pathologies, i.e., further allergies, diseases, vaccines, etc.

While there is no limitation on the variety of ontologies that could be leveraged, in this work we mainly describe the matches found within HDO, also referred as DOID, which is a well-known comprehensive ontology of human diseases including their descriptions, synonyms, antonyms, etc. Other ontologies used in our experiments were OBIB, IDO, and OHD; however, a much smaller number of ontology classes could be found to match clinical data compared to HDO.

In more detail, in order to match the relevant semantic data to the data extracted from a CDA, the following steps are performed.

- Data extrapolation: for each medical keyword, among the whole set of its associated key–value data pairs we filter only those corresponding to Type and Name keys (see e.g., lines 19 and 23 in Listing 4). This task is easily accomplished thanks to the previous CDA XML data conversion into a more easily processed JSON file, from where it is easier to extrapolate the most relevant keywords.
- Data cleaning: in order to make the extracted data actually suitable, data cleaning has to be carried out: that is, elimination of special characters and making all characters lowercase.
- Data filtering: a further reduced set of keywords is obtained by selecting only the keywords whose corresponding values has a string length with a maximum of three words. This is a heuristic that was introduced, as more words would likely indicate a very specific issue that would be unlikely to be found within an ontology.
- Ontology filtering: the selected ontology is processed to extract a list of useful terms that could be related to the patient's medical record. This is achieved by parsing the ontology with the supporting library Jena (https://jena.apache.org, last accessed on 10 November 2023), which converts all its semantic concepts into corresponding class objects whose names can then be easily combined with the data of the patient's clinical document.

Listing 4. JSON object extracted from the XML of Example 2.

```
1  {
2     "title":"Patient Chart Summary",
3     "date":"201308151030-0800",
4     "patient":{
5        "id":"444222222",
6        "name":"Eve",
7        "last_name":"Betterhalf",
8        "gender":"female",
9        "maritalStatus":"married",
10       "birthTime":"19750501",
11       "religious":"Christian (non-Catholic, non-specific)"
12    },
13    "allergiesSection":{
14       "title":"Allergies, Adverse Reactions and Alerts",
15       "date":"20140104123506-0500",
16       "allergies":[
17          {
18             "allergen":"penicillin",
19             "type":"drug allergy",
20             "documentation_date":"Jan 4 2014",
21             "reactions":[
               {
                  "name":"Anaphylaxis",
                  "reaction_severity":"Severe",
                  "start_date":"2006"
               }
            ]
         }
      ]
   },
   ...
}
```

The following code snippet shows the ontology filtering process in detail.

```
/** Extrapolating class names supported by the Apache Jena library **/
private List<String> extrapolationByOntology(){
    OntModel model = ModelFactory.createOntologyModel(OntModelSpec.OWL_DL_MEM);
    model.read(ontologyPath);
    listLabel = model.listClasses().toList().stream()
       .filter(s -> s.getURI() != null && s.getLabel(null) != null)
       .map(s -> s.getLabel(null))
       .toList()
    return listLabel;
}
```

The last step implementing the ontological matching is the search for correspondences between the clinical keywords data obtained from the clinical document and the available ontology classes. This, in fact, allows us to later build an enriched representation of the patient's EHR, where the set of additional related data is bound to the JSON converted clinical document. To correctly find the semantic relationships between the ontology concepts and the clinical data terms we used a well-known natural language processing (NLP) technique: we applied a filtering layer based on the Jaccard distance, which is widely adopted as an effective criteria to compute the similarity between two sets of words with the aim of revealing their association. This step is performed as shown in detail in the following code snippet:

```
/** Calculates the similarity between text using the JaccardDistance class belonging to
     the package org.apache.commons.text.similarity **/
private List<Label> textSimilarity(List<Keyword>jsons,List<String> labels){
    List<Label> toReturn = new ArrayList<Label>();
    for (Keyword obj : jsons) {
        String keys = obj.toString();
        for (String str : labels) {
        double jaccardResult = jaccardDistance.apply(keys, str);
        if (jaccardResult >= 0.7) {
                toReturn.add(new Label(keys, str, jaccardResult));
    }}}
    return toReturn;
}
```

4.3. Document Enrichment

The final stage consists of the actual building of a single manageable document that can be viewed by doctors, aggregating the enhanced clinical data and the additional semantic information obtained into a single JSON file. This will thus contain added notes to the original medical data, i.e., detailed explanations of the diseases or synonyms or opposites of them.

The following section discusses the result of the enrichment for real sample data, and gives a graphical representation of data resulting from the merging between clinical data and an ontology.

5. Validation

In order to validate the proposed approach, we processed a set of CDA documents extracted from the HL7 official public test data repository as an overall functional test case. Specifically, we processed a set of HL7 files simulating the medical story of the imaginary patient "Eve Betterhalf". Each file represents a report of one of her visits performed by a clinical specialist. The considered patient's fascicle totalled 35 CDA files, the union of which constituted her EHR. Each processed file contained several clinical keywords that our tool matched to concepts in the HDO ontology, allowing to build a new enhanced health record, embedding clear and precise semantic explanations of all the clinical terms and notions used in the original document.

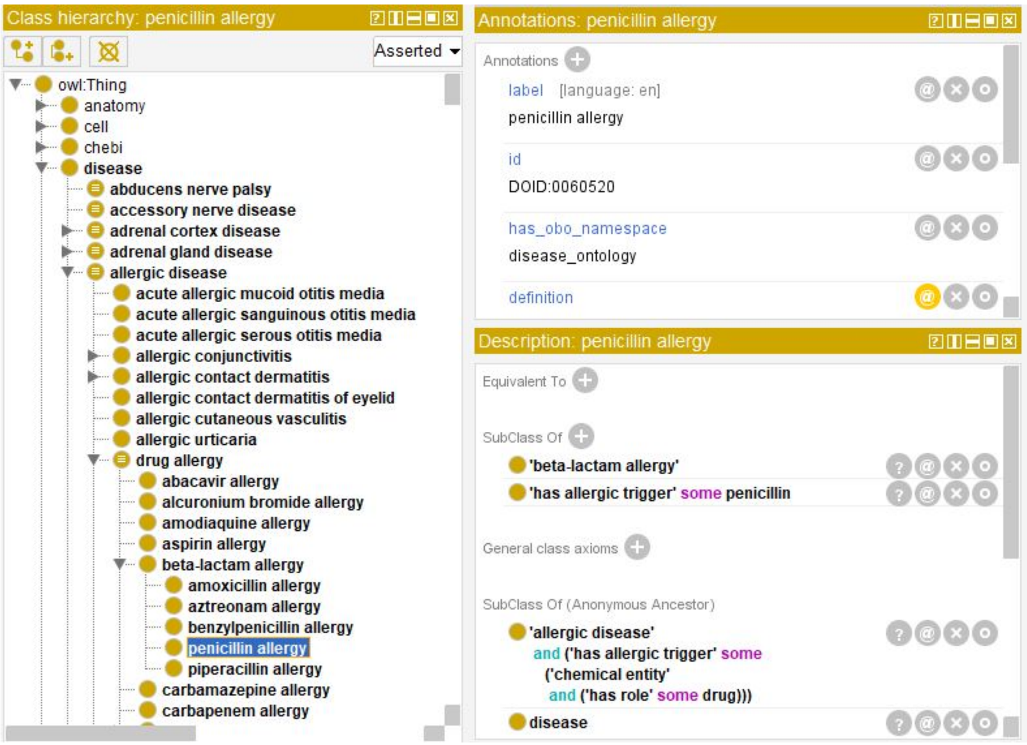
Table 1 reports a short list of terms, showing some example keywords and terms which have been actually subject to this embedding process during the tests along with their original reference IDs. Typology is determined according to the derived JSON. Name defines a subclass concept related to the corresponding type and IRI is the univocal id present on the ontology which belongs to the area of the domain which matches the name and the typology (an IRI would need to have the prefix http://purl.obolibrary.org/obo/ (accessed on 10 January 2024)). After processing, the whole original health record has been reformatted as a reduced set of self-contained, easier-to-read JSON-encoded documents, carefully preserving (and improving on) the original medical data content.

Table 1. List of some terms present in the processed HL7 format files for which a match has been found in the HDO.

Typology	Name	IRI
Allergy	Egg	DOID_4377
Allergy	Latex	DOID_0060532
Allergy	Penicillin	DOID_0060520
Diagnosis	Congestive heart failure	DOID_6000
Diagnosis	Pneumonia	DOID_552
Reactions	Anaphylaxis	SYMP_0000895
Reactions	Rash	SYMP_0000487
Reactions	Ventricular tachycardia	SYMP_0000827
Reactions	Nausea	SYMP_0000458
Reactions	Urticaria	DOID_1555

We illustrate this in detail with an actual enhanced health record extracted from the set of processed documents obtained from the original CDA document (partially) shown in Figure 1. The original XML CDA file describes the patient's allergy to penicillin (among others) and the severity of the allergic reaction. Processing the hard-to-read XML medical record, a simplified version is created in JSON, where it can be more easily deduced that she is allergic to the drug penicillin and that she had anaphylaxis as a severe allergic reaction. Specifically, two keywords have been extracted, penicillin and anaphylaxis, and they were semantically defined as part of the knowledge domain of the HDO ontology.

As shown in Figures 4 and 5, the ontological data records for those two keywords are detailed and rich and allow additional semantic knowledge to be extracted from those records and embedded into to the medical record. The resulting new health record is represented in Figure 6, as shown with a basic JSON file reader and merged in one clearly readable data structure (shown in the centre of the figure). All of the original information about not only that one allergy cited above, but about all three types of allergies that the patient was suffering were included. In fact, those allergies' data were originally split on separate CDA files that were cluttered with XML syntactical artefacts and alphanumeric codes and tags. It also enriched the overall allergies data record with textual descriptions of the terms for related diseases and symptoms, following the semantic connections in the associated ontology (lollipops on the sides of Figure 6). This will help the doctor remember additional drug- and allergy-related concepts, having a handy and valid help to accordingly make a more aware and correct diagnosis. In the centre of Figure 6, there is a table representing the JSONs coming from three HL7 CDA files, each relating to a visit made by the patient, whereas inside the coloured boxes there are data taken from the HDO ontology, offering additional support to the doctor.



label [language: en]
penicillin allergy

id
DOID:0060520

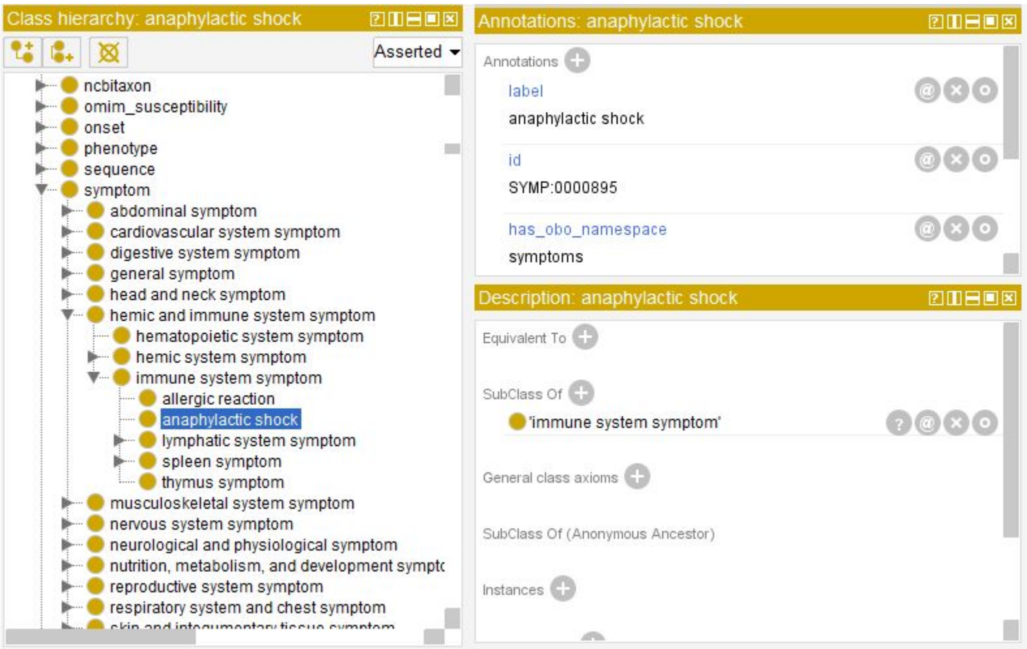
has_obo_namespace
disease_ontology

definition

SubClass Of
'beta-lactam allergy'
'has allergic trigger' some penicillin

SubClass Of (Anonymous Ancestor)
'allergic disease'
and ('has allergic trigger' some
('chemical entity'
and ('has role' some drug)))
disease

Figure 4. A representation of the class penicillin allergy found in HDO.


Class hierarchy: anaphylactic shock
Asserted
ncbitaxon
omim_susceptibility
onset
phenotype
sequence
symptom
abdominal symptom
cardiovascular system symptom
digestive system symptom
general symptom
head and neck symptom
hemic and immune system symptom
hematopoietic system symptom
hemic system symptom
immune system symptom
allergic reaction
anaphylactic shock
lymphatic system symptom
spleen symptom
thymus symptom
musculoskeletal system symptom
nervous system symptom
neurological and physiological symptom
nutrition, metabolism, and development sympto
reproductive system symptom
respiratory system and chest symptom
skin and integumentary tissue symptom
Annotations: anaphylactic shock
Annotations
Description: anaphylactic shock
Equivalent To
General class axioms
SubClass Of (Anonymous Ancestor)
Instances

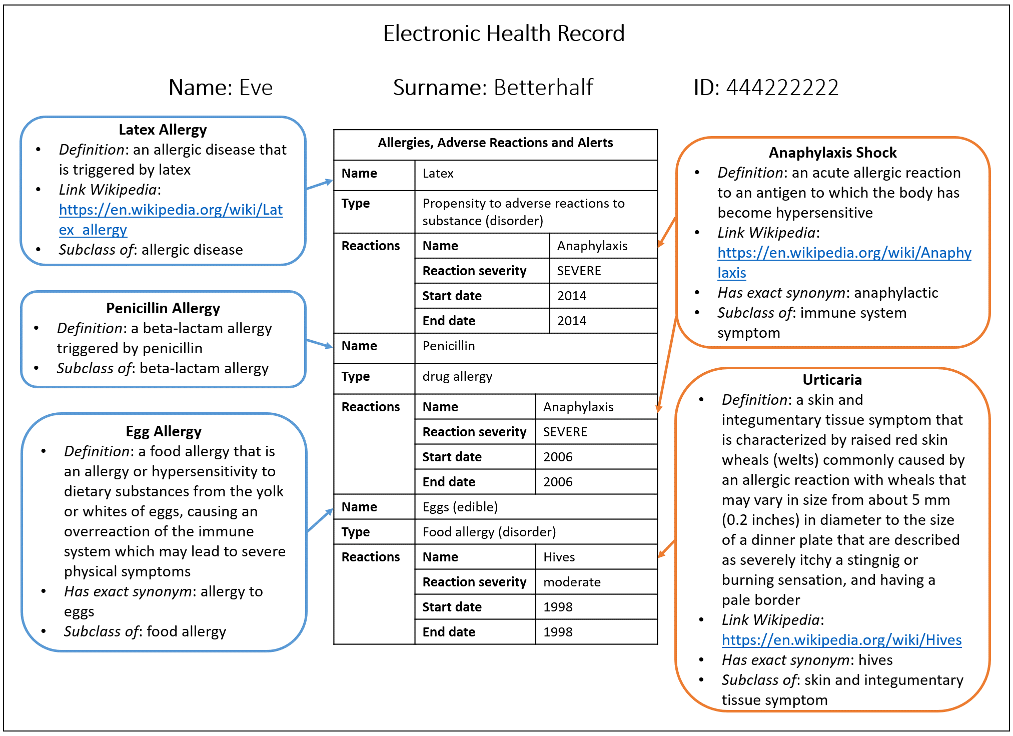
label
anaphylactic shock

id
SYMP:0000895

has_obo_namespace
symptoms

SubClass Of
'immune system symptom'

Figure 5. A representation of the class anaphylactic shock found in HDO.



Electronic Health Record

Name: Eve Surname: Betterhalf ID: 444222222

Latex Allergy
- *Definition*: an allergic disease that is triggered by latex
- *Link Wikipedia*: https://en.wikipedia.org/wiki/Latex_allergy
- *Subclass of*: allergic disease

Penicillin Allergy
- *Definition*: a beta-lactam allergy triggered by penicillin
- *Subclass of*: beta-lactam allergy

Egg Allergy
- *Definition*: a food allergy that is an allergy or hypersensitivity to dietary substances from the yolk or whites of eggs, causing an overreaction of the immune system which may lead to severe physical symptoms
- *Has exact synonym*: allergy to eggs
- *Subclass of*: food allergy

Allergies, Adverse Reactions and Alerts		
Name	Latex	
Type	Propensity to adverse reactions to substance (disorder)	
Reactions	**Name**	Anaphylaxis
	Reaction severity	SEVERE
	Start date	2014
	End date	2014
Name	Penicillin	
Type	drug allergy	
Reactions	**Name**	Anaphylaxis
	Reaction severity	SEVERE
	Start date	2006
	End date	2006
Name	Eggs (edible)	
Type	Food allergy (disorder)	
Reactions	**Name**	Hives
	Reaction severity	moderate
	Start date	1998
	End date	1998

Anaphylaxis Shock
- *Definition*: an acute allergic reaction to an antigen to which the body has become hypersensitive
- *Link Wikipedia*: https://en.wikipedia.org/wiki/Anaphylaxis
- *Has exact synonym*: anaphylactic
- *Subclass of*: immune system symptom

Urticaria
- *Definition*: a skin and integumentary tissue symptom that is characterized by raised red skin wheals (welts) commonly caused by an allergic reaction with wheals that may vary in size from about 5 mm (0.2 inches) in diameter to the size of a dinner plate that are described as severely itchy a stingnig or burning sensation, and having a pale border
- *Link Wikipedia*: https://en.wikipedia.org/wiki/Hives
- *Has exact synonym*: hives
- *Subclass of*: skin and integumentary tissue symptom

Figure 6. Example of a patient's medical record with ID 444222222, showing data taken from three HL7 CDA files (centre) and HDO ontology (accessed on 10 November 2023).

6. Conclusions

EHR is an exceptional tool that allows citizens and clinicians to keep up with technology and obtain faster and easier-to-access results, offering a significant improvement in the efficacy and management of patient's health care services. To make these services more efficient, this work presents an original approach to improve the usability of medical records, reducing the complexity of the data and adding new and useful information to facilitate the doctor's work. Extensive testing have been carried out on actual medical data provided by the HL7 organisation, which successfully validate our work.

In currently common clinical settings, doctors often use EHRs with classic tools with which it is difficult to achieve a complete picture of a patients whole clinical record since each section is a separate CDA document based on potentially different keywords and dictionaries, strictly requiring (different) software tools to be managed or simply to be readable. In this context, this work proposes an improved scenario where our single tool is able to process all the EHR data and offer an easy-to-read, complete, and enriched medical record. Without our tool, several tools would have to be used independently of each other, leading to a considerable workload for the doctor as they are then required to have knowledge of and usage experience with all of them. In addition, they would need to spend a lot of time manually researching any correlations or associations with existing medical knowledge. They would be required to create special queries to manually search and read any available ontology data. Using our approach, our proposed single tool would allow for the effective presentation and navigation of the clinical information, already combined with any pertinent knowledge found in the matching ontologies, as a data aggregate which is immediately readable and clearly understandable for the doctors.

Author Contributions: Conceptualization, S.C., E.T. and G.V.; methodology, S.C., E.T. and G.V.; software, S.C. and G.V.; validation, A.C., E.T. and G.V.; writing—original draft preparation, A.C., E.T. and G.V.; writing—review and editing, A.C. and E.T.; supervision, A.C. and E.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors acknowledge the support of the University of Catania PIACERI 2020/22 Project "TEAMS" and project PO FESR Sicilia 2014–2020 Biospec3.

Conflicts of Interest: Salvatore Calcagno has been involved as consultant and expert in the field. He is an employee of the company Cyberetna. The authors declare no conflicts of interest.

References

1. Henderson, M.L.; Dayhoff, R.E.; Titton, C.P.; Casertano, A. Using IHE and HL7 conformance to specify consistent PACS interoperability for a large multi-center enterprise. *J. Healthc. Inf. Manag.* **2006**, *20*, 47.
2. Noumeir, R. Active learning of the HL7 medical standard. *J. Digit. Imaging* **2019**, *32*, 354–361. [CrossRef] [PubMed]
3. Al-Aswadi, F.N.; Chan, H.Y.; Gan, K.H. Automatic ontology construction from text: A review from shallow to deep learning trend. *Artif. Intell. Rev.* **2020**, *53*, 3901–3928. [CrossRef]
4. Dissanayake, P.I.; Colicchio, T.K.; Cimino, J.J. Using clinical reasoning ontologies to make smarter clinical decision support systems: A systematic review and data synthesis. *J. Am. Med Inform. Assoc.* **2020**, *27*, 159–174. [CrossRef]
5. Schriml, L.M.; Mitraka, E.; Munro, J.; Tauber, B.; Schor, M.; Nickle, L.; Felix, V.; Jeng, L.; Bearer, C.; Lichenstein, R.; et al. Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Res.* **2019**, *47*, D955–D962. [CrossRef]
6. Shaker, N.; Patel, A.; Tozbikian, G.; Parwani, A. Anastomosing hemangioma: A case report of a benign tumor often misdiagnosed as a malignant epithelioid angiosarcoma. *Urol. Case Rep.* **2022**, *42*, 102023. [CrossRef] [PubMed]
7. Chen, J.J.; Kardon, R.H. Avoiding clinical misinterpretation and artifacts of optical coherence tomography analysis of the optic nerve, retinal nerve fiber layer, and ganglion cell layer. *J. Neuro-Ophthalmol.* **2016**, *36*, 417. [CrossRef]
8. Beckmann, C.L.; Keuchel, D.; Soleman, W.O.I.A.; Nürnberg, S.; Böckmann, B. Semantic Integration of BPMN Models and FHIR Data to Enable Personalized Decision Support for Malignant Melanoma. *Information* **2023**, *14*, 649. [CrossRef]
9. Torres-Silva, E.A.; Rúa, S.; Giraldo-Forero, A.F.; Durango, M.C.; Flórez-Arango, J.F.; Orozco-Duque, A. Classification of Severe Maternal Morbidity from Electronic Health Records Written in Spanish Using Natural Language Processing. *Appl. Sci.* **2023**, *13*, 10725. [CrossRef]
10. Rosenbloom, S.T.; Denny, J.C.; Xu, H.; Lorenzi, N.; Stead, W.W.; Johnson, K.B. Data from clinical notes: A perspective on the tension between structure and flexible documentation. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 181–186. [CrossRef]
11. Rousseau, J.F.; Ip, I.K.; Raja, A.S.; Valtchinov, V.I.; Cochon, L.; Schuur, J.D.; Khorasani, R. Can automated retrieval of data from emergency department physician notes enhance the imaging order entry process? *Appl. Clin. Inform.* **2019**, *10*, 189–198. [CrossRef]
12. Rousseau, J.F.; Ip, I.K.; Raja, A.S.; Schuur, J.D.; Khorasani, R. Can emergency department provider notes help to achieve more dynamic clinical decision support? *J. Am. Coll. Emerg. Physicians Open* **2020**, *1*, 1269–1277. [CrossRef] [PubMed]
13. Abdelgalil, L.; Mejri, M. HealthBlock: A Framework for a Collaborative Sharing of Electronic Health Records Based on Blockchain. *Future Internet* **2023**, *15*, 87. [CrossRef]
14. Lin, B.; Chen, Y.; Chen, X.; Yu, Y. Comparison between JSON and XML in Applications Based on AJAX. In Proceedings of the International Conference on Computer Science and Service System, Nanjing, China, 11–13 August 2012; pp. 1174–1177.
15. Rousseau, J.F.; Oliveira, E.; Tierney, W.M.; Khurshid, A. Methods for development and application of data standards in an ontology-driven information model for measuring, managing, and computing social determinants of health for individuals, households, and communities evaluated through an example of asthma. *J. Biomed. Inform.* **2022**, *136*, 104241. [CrossRef]
16. Rinner, C.; Duftschmid, G. Bridging the gap between HL7 CDA and HL7 FHIR: A JSON based mapping. In Proceedings of the eHealth, Vienna, Austria, 24–25 May 2016; pp. 100–106.
17. Haendel, M.A.; Chute, C.G.; Robinson, P.N. Classification, ontology, and precision medicine. *N. Engl. J. Med.* **2018**, *379*, 1452–1462. [CrossRef] [PubMed]
18. Grimm, S.; Hitzler, P.; Abecker, A. Knowledge representation and ontologies. In *Semantic Web Services*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 51–105.
19. Gai, K.; Qiu, M.; Chen, L.C.; Liu, M. Electronic health record error prevention approach using ontology in big data. In Proceedings of the Conference on High Performance Computing and Communications, Symposium on Cyberspace Safety and Security, and Conference on Embedded Software and Systems, New York, NY, USA, 24–26 August 2015; pp. 752–757.
20. Abdelreheim, M.; Soliman, T.H.A.; Klan, F. A Personalized Ontology Recommendation System to Effectively Support Ontology Development by Reuse. *Future Internet* **2023**, *15*, 331. [CrossRef]
21. Mahmoodi, S.A.; Mirzaie, K.; Mahmoudi, S.M. A new algorithm to extract hidden rules of gastric cancer data based on ontology. *SpringerPlus* **2016**, *5*, 312. [CrossRef]
22. Reyes-Ortíz, J.A.; Jiménez, A.L.; Cater, J.; Meléndez, C.A.; Márquez, P.B.; García, M. Ontology-based Knowledge Representation for Supporting Medical Decisions. *Res. Comput. Sci.* **2013**, *68*, 127–136. [CrossRef]
23. Sow, A.; Guissé, A.; Niang, O. Enrichment of Medical Ontologies from Textual Clinical Reports: Towards Improving Linking Human Diseases and Signs. In Proceedings of the International Conference on Innovations and Interdisciplinary Solutions for Underserved Areas, Cairo, Egypt, 14–15 February 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 104–115.

24. Tramontana, E.; Verga, G. Ontology Enrichment with Text Extracted from Wikipedia. In Proceedings of the International Conference on Software Engineering and Information Management (ICSIM), online, 21–23 January 2022; pp. 113–117.
25. Tramontana, E.; Verga, G. Keeping Researchers Updated by Automatically Enriching an Ontology in the Medical Field. In Proceedings of the International Conference on Computer and Communications Management (ICCCM), Okayama, Japan, 29–31 July 2022; pp. 257–262.
26. Kalra, D. Electronic health record standards. *Yearb. Med. Inform.* **2006**, *15*, 136–144. [CrossRef]
27. Bologna, S.; Bellavista, A.; Corso, P.P.; Zangara, G. Electronic Health Record in Italy and Personal Data Protection. *Eur. J. Health Law* **2016**, *23*, 265–277. [CrossRef] [PubMed]
28. Sachdeva, S.; Bhalla, S. Using Knowledge Graph Structures for Semantic Interoperability in Electronic Health Records Data Exchanges. *Information* **2022**, *13*, 52. [CrossRef]
29. Kalra, D.; Beale, T.; Heard, S. The openEHR foundation. *Stud. Health Technol. Inform.* **2005**, *115*, 153–173. [PubMed]
30. Muñoz, P.; Trigo, J.D.; Martínez, I.; Muñoz, A.; Escayola, J.; García, J. The ISO/EN 13606 standard for the interoperable exchange of electronic health records. *J. Healthc. Eng.* **2011**, *2*, 1–24. [CrossRef]
31. Begoyan, A. An overview of interoperability standards for electronic health records. In Proceedings of the Integrated Design and Process Technology, Antalya, Turkey, 3–8 June 2007.
32. Bender, D.; Sartipi, K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In Proceedings of the Symposium on Computer-Based Medical Systems, Porto, Portugal, 20–22 June 2013; pp. 326–331.
33. Nagy, M.M. HL7-Based Data Exchange in EHR Systems. In Proceedings of the Ph.D. Conference, Jizerka, Czechia, 29 September–1 October 2008.
34. Dolin, R.H.; Alschuler, L.; Beebe, C.; Biron, P.V.; Boyer, S.L.; Essin, D.; Kimber, E.; Lincoln, T.; Mattison, J.E. The HL7 clinical document architecture. *J. Am. Med. Inform. Assoc.* **2001**, *8*, 552–569. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.