From SNOMED CT Expressions to an FHIR RDF Representation: Exploring the Benefits of an Ontology-Based Approach

Mercedes ARGUELLO-CASTELEIRO[a], Catalina MARTÍNEZ-COSTA[b],
Julio DES-DIZ[c], Nava MAROTO[d], Maria Jesus FERNANDEZ-PRIETO[e]
and Robert STEVENS[a,1]

[a] *University of Manchester, UK*
[b] *Medical University of Graz, Austria*
[c] *Hospital do Salnés, Spain*
[d] *Universidad Politécnica de Madrid, Spain*
[e] *University of Salford, UK*

Abstract. We address the problem of semantic interoperability of HL7 standards and propose an ontology-based approach to convert smoothly HL7 C-CDA coded entries containing pre- and post-coordinated SNOMED CT expressions into HL7 FHIR resources in RDF. Our ontology-based approach is based on Content Ontology Design Patterns (ODPs) and seeks to: 1) constrain further the SNOMED CT post-coordination that otherwise can lead to an unmanageable number of possible SNOMED CT expressions; 2) minimise the variability when mapping between the structures and semantics of the HL7 FHIR specification to SNOMED CT expressions; and 3) smooth the transformation of SNOMED CT expressions to an FHIR RDF representation, leveraging on FHIR ShEx schemas to formally describe FHIR RDF data instances. To validate the proposal this study utilises 358 SNOMED CT expressions from 3 sections of anonymised consultation notes in HL7 C-CDA. Besides converting HL7 C-CDA document entries into FHIR RDF data instances, we explore the benefits of the Content ODPs to facilitate large-scale data analytics (e.g. cross-compare patients) and Natural Language Generation by generating text from the clinical coded data.

Keywords. Content ODPs, SNOMED CT, HL7 C-CDA, HL7 FHIR, NLG

1. Introduction

Initiatives at the country-level are acting as driving forces for the "digitalisation" of health care. There is interest from industry and government to achieve interoperability among Health Level 7 (HL7) primary healthcare standards [1]: Clinical Document Architecture (CDA) and Fast Healthcare Interoperability Resources (FHIR) and Continuous CDA (C-CDA). Among the challenges in achieving interoperability [1]:

- The value sets used in FHIR generally lack alignment with those in C-CDA.
- Negation used in C-CDA is quite different from the Core FHIR specification.

[1] Corresponding author, E-mail: Robert.Stevens@manchester.ac.uk.

- The level of granularity between C-CDA and FHIR is often different.

This paper addresses the problem of semantic interoperability of HL7 standards and proposes an ontology-based approach to convert smoothly C-CDA coded entries containing pre- and post-coordinated SNOMED CT expressions into FHIR resources represented in RDF [2]. Our ontology-based approach exploits Ontology Design Patterns (ODPs) that can "*encapsulate in a single named representation the semantics that require several statements in ontology languages*" [3]. There are different types of ODPs [4]. This study focuses on Content ODPs, which are domain-related ontology patterns [4], to formally represent healthcare data models and aid:

1. Constraining further the SNOMED CT post-coordination that otherwise can lead to an unmanageable variety of possible SNOMED CT expressions.
2. Minimising the variability when mapping between the structures and semantics of the HL7 FHIR specification to SNOMED CT expressions.
3. Smoothing the transformation of SNOMED CT expressions to FHIR resources in RDF by exploiting the RDF Shape Expressions language (ShEx) [5]; a language for RDF validation. We utilise the FHIR ShEx schemas from [6].

In the UK, the National Health Service (NHS) is moving towards a single terminology, SNOMED CT [7], across health and care settings [8]. Even the adoption of only one terminology by the NHS, like SNOMED CT, allows room for variance when coding clinical information. Table 1 shows the multiple options when coding the relatively simple clinical statement "cataract of left eye".

Table 1. Exemplifying SNOMED CT expressions that represent laterality

SNOMED CT expression	Close-to-user expression view
A nested expression with a laterality refinement	420123008| On examination - cataract|: 363698007|Finding site|=(78076003|Structure of lens of eye|: 272741003|Laterality|=7771000|Left|)
An expression with a refinement representing lateralisation	420123008| On examination – cataract|: 363698007|Finding site|= 88258005|Structure of lens of left eye|
Laterality pre-coordinated	418319009|On examination - cataract of left eye|

Content ODPs can narrow the available options and favour the systematic selection of favoured SNOMED CT expressions. In turn, a systematic post-coordination of SNOMED CT expressions enforced by Content ODPs can foster the retrieval of clinical characteristics that are easier to compare and may enable the discovery of clinical correlations of interest for both clinicians and researchers. Thus, a more effective secondary use of healthcare data.

Defining both the format (data model) and the content (standardised terminology) is an acknowledged necessity for health care data standardisation [9]. This study investigates if Content ODPs can contribute to achieve semantic interoperability of HL7 standards, and also explores if there are some other benefits to gain by adopting Content ODPs, such as: 1) facilitating large-scale data analytics; and 2) Natural Language Generation (NLG) [10] by generating text from the clinical coded data.

Large-scale data analytics enabled by secondary use of healthcare data for clinical and translational research is expected to allow (among other goals):

- *Cross-compare patients* to identify at-risk patients or to obtain data-driven phenotypes based on clinical characteristics associated with clinical outcomes.

For example, establishing the clinical characteristics for a well-known medical condition like glaucoma.

- *Acquisition of clinical evidence* of known and/or unknown correlations among clinical/biomedical concepts. For example, performing cataract surgery prior to glaucoma surgery can bring multiple clinical benefits [11].

HL7 healthcare standards intend to make the clinical content machine processable (standardised content with rich shared semantics) while guaranteeing human readability. NLG can generate text from the coded clinical content, and this can help:

- Checking if the clinical meaning of the coded content is the same as the one intended in the human readable text provided. The text created by NLG can be cross-compared with the original human readable text.
- Lowering the current administrative/reporting burden of front-line clinicians. A recent retrospective cohort study of 142 family medicine physicians in Wisconsin (US) shows that "*primary care physicians spend nearly 2 hours on electronic health record (EHR) tasks per hour of direct patient care*" [12].

To validate our proposal, this study focuses on ophthalmology as "*a great number of systemic diseases can exhibit ocular manifestations during their evolution*" [13]. An ocular condition like glaucoma may cause symptoms like "nausea" or "vomiting". Diabetic eye disease - including diabetic retinopathy, diabetic macular edema, cataract, and glaucoma - are a group of eye conditions that affect people with diabetes [14]. Besides its significance, ophthalmology poses an intrinsic difficulty for SNOMED CT expressions. In ophthalmology there is the need to express locus (i.e. finding site) and laterality, which may hamper information retrieval if pre- and post-coordinated SNOMED CT expressions are not transformed into a common form [15].

2. An Ontology-Based Approach to Map SNOMED CT Expressions to RDF FHIR

Figure 1 shows an overview of the proposed approach. We start by describing the coded clinical data from C-CDA that needs to be converted into FHIR. Next, we discuss the *terminology binding* which is defined as: "*A key task is to define which codes are to be used where – to bind the terminology to the model of the medical record or message*" [16]. Finally, we provide details of the two conversions to be performed as highlighted in Figure 1.

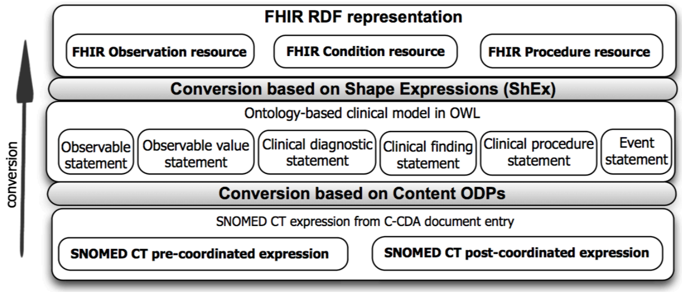
FHIR RDF representation
FHIR Observation resource
FHIR Condition resource
FHIR Procedure resource
Conversion based on Shape Expressions (ShEx)
Ontology-based clinical model in OWL
Observable statement
Observable value statement
Clinical diagnostic statement
Clinical finding statement
Clinical procedure statement
Event statement
Conversion based on Content ODPs
SNOMED CT expression from C-CDA document entry
SNOMED CT pre-coordinated expression
SNOMED CT post-coordinated expression
conversion

Figure 1. Approach overview to transform SNOMED CT expressions from C-CDA into FHIR RDF

2.1. Consultation Notes in C-CDA for the Ophthalmologic Domain of Red Eye

This study considers 358 coded entries with SNOMED CT expressions within three sections of de-identified consultation notes from a collection of 103 HL7 C-CDA documents. The three sections selected for this study (i.e. "Physical Exam", "Assessment", and "History Present Illness") were considered the most clinically significant. The sections "Physical Exam" and "Assessment" are required by C-CDA.

The 103 HL7 C-CDA documents belong to the ophthalmologic domain of *Red Eye*, which encompasses vision-threatening conditions like glaucoma and more benign eye conditions like conjunctivitis. The 358 coded entries are from 18 of the total 103 HL7 C-CDA documents. The document selection was made to cover as much as possible of the variety of medical conditions under the *Red Eye* domain.

2.2. General Pattern of SNOMED CT Expressions: Terminology Binding Graphically

A SNOMED CT expression can be pre-coordinated, with only one concept identifier; or post-coordinated with more than one concept identifier. Overall, a SNOMED CT expression consists of one or more concept identifiers, i.e. the focus concept(s), plus optional refinements [17].

There is a general pattern that applies to all SNOMED CT expressions [17]. According to this general pattern, there are: focus concepts; focus refinements; and nested refinements. The refinements may include any number of attributes [17].

Within the general pattern that applies to all SNOMED CT expressions, it is feasible to distinguish between the *clinical kernel* and the *context wrapper* [17], where the clinical kernel contains the focus concept(s). Figure 2 shows a focus expression (i.e. a close-to-user expression that refines a concept); and Figure 3 shows a nested close-to-user expression that includes context.

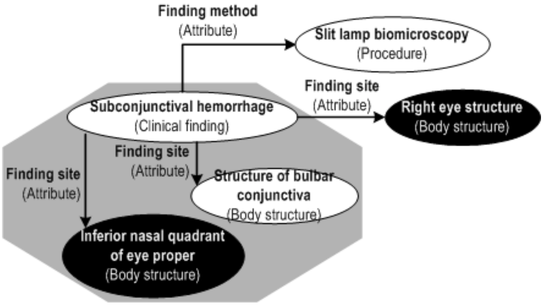

Figure 2. Common Close-To-User Expression Patterns: **Clinical finding present**. This is a focus expression.

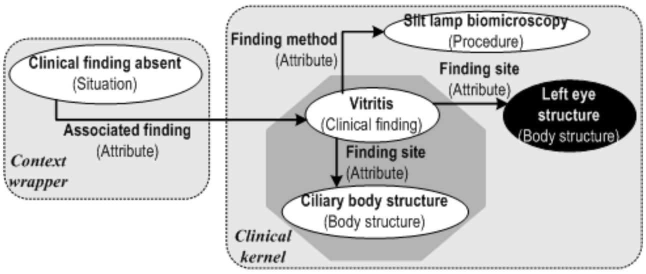

Figure 3. Common Close-To-User Expression Patterns: **Clinical finding absent.** This is a nested expression including context.

Figure 2 and 3 are also examples of SNOMED CT Common Close-To-User Expression Patterns [18] that exemplify the presence (Figure 2) or absence (Figure 3) of a clinical finding. Both diagrams contain an octagon that contains the part of a SNOMED CT expression that is coded in SNOMED CT within a C-CDA entry, while the arrows (attributes) and ellipses (concepts) just outside of the octagon represent the part of the SNOMED CT expression that overlaps in meaning with the data model. Therefore, Figure 2 and 3 can be interpreted as a graphical representation of the *terminology binding* while being detached from the concrete implementation details.

2.3. SNOMED CT Binding for FHIR Resources Observation and Condition

The SNOMED CT expressions from the *Physical Exam* section and *History Present Illness* section are mapped to the FHIR resource Observation [19], while the ones from the *Assessment* section are mapped to the FHIR resource Condition [20]. We found:

1. Representing laterality for body structure is non-trivial in FHIR [21].
2. Discrepancies exist between the rules for case patterns for the FHIR data elements code and value for the FHIR Observation in [19] and what appears stated as SNOMED CT concept domain bindings for the FHIR Observation in [22].
3. It is unclear how the SNOMED CT binding relates to SNOMED CT Common Close-To-User Expression Patterns [18].

To avoid the current difficulty of representing laterality in FHIR, this study uses systematically two SNOMED CT concepts that appear in Figure 4 as two black ellipses: "*8966001|Left eye structure*" and "*18944008|Right eye structure*". Both concepts are descendants of "*442083009|Anatomical or acquired body structure*", and thus, valid coded values for the data elements Observation.bodySite and Condition.bodySite in FHIR.

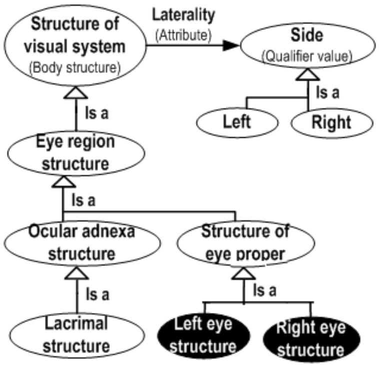

Figure 4. Overview of some SNOMED CT *Body Structure* concepts relevant for the *Red Eye* domain.

Table 2 shows some of the FHIR data elements (left column) and how they bind to SNOMED CT concepts (right column) applying exactly the mappings from [22,23] when they appear in bold. Typically, the coded values of the right column were expanded (e.g. including valid coded values for the Observation.code following case 3 pattern in [19]) or restricted (e.g. a more specific concept for the Observation.bodySite).

Table 3 illustrates our mapping of SNOMED CT Common Close-To-User Expression Patterns [18] (first column) to the patterns for the FHIR Observation resource [19] (last two columns) with the terminology binding proposed (Table 2). The absence of a clinical finding (i.e. negation) in FHIR is represented using the coded value ‘clinical-finding’ in the data element Observation.dataAbsentReason.

Table 2. SNOMED CT concept domain binding proposed in this study to FHIR data elements

FHIR terminology bindings	Descendant of SNOMED CT concept
Observation.code	**363787002|Observable entity (observable entity)** OR **386053000|Evaluation procedure (procedure)** OR 404684003|Clinical finding (finding) OR 272379006|Event (event)
Observation.value	**441742003|Evaluation finding (finding)** [OR numeric values]
Observation.method	386053000|Evaluation procedure (procedure)
Observation.bodySite	442083009|Anatomical or acquired body structure (body structure)
Observation.specimen	**123038009|Specimen (specimen)**
Condition.code	64572001|Disease (disorder)
Condition.bodySite	**442083009|Anatomical or acquired body structure (body structure)**

Table 3. SNOMED CT Common Close-To-User Expression Patterns map to FHIR Observation patterns

SNOMED CT Common Close-To-User Expression Patterns	**Observation.code** Descendant of SNOMED CT concept that is the focus concept(s) [17]	**Observation.value** [It can be omitted]
History of: Patient-reported symptom	404684003|Clinical finding	[SNOMED CT refinements]
History of: Patient-reported event	272379006|Event	[SNOMED CT refinements]
Observable + value: Clinician-examination	363787002|Observable entity	{Numeric; nominal; ordinal; coded result; measurement}
Clinical finding present Clinician-examination	SCT:404684003|Clinical finding	[SNOMED CT refinements]
Clinical finding absent Clinician-examination	SCT:404684003|Clinical finding	[SNOMED CT refinements]

2.4. Creating Content ODPs for SNOMED CT Common Close-To-User Expressions

The SNOMED CT Common Close-To-User Expression Patterns [18], which appear in the first column in Table 3, can be represented formally using Content ODPs [4] in OWL [24]. Table 4 illustrates some of the Content ODPs written using Manchester OWL Syntax [25]. An earlier version of the Content ODPs was presented in [26].

Table 4. Exemplifying Content ODPs in OWL for SNOMED CT Common Close-To-User expressions

Common Close-To-User Expression Patterns	**Content ODPs in Manchester OWL Syntax**
History of: (Present Illness) Patient-reported symptom	SymptomReportedStatement and ('has part' some 'Known present (qualifier value)')
History of: (Present Illness) Patient-reported event	'information object' and ('has part' some InformationAboutSubjectOfInformation) and ('is outcome of' some 'Event (event)')
Observable + value: Clinician-examination	ClinicalStatement and ('is outcome of' some 'Evaluation procedure (procedure)') and (represents some ObservableResultValue)
Clinical finding present Clinician-examination	ClinicalFindingStatement and ('has part' some 'Known present (qualifier value)')
Clinical finding absent Clinician-examination	ClinicalFindingStatement and ('has part' some 'Known absent (qualifier value)')
Clinical finding present Clinician-asserted diagnosis	ClinicalStatement and ('is outcome of' some 'Diagnostic procedure (procedure)') and (represents some ClinicalLifePhase)

The Content ODPs in Table 4 use OWL constructs from:

- SNOMED CT in OWL. For example the OWL Class for the SNOMED CT concept "*272379006|Event (event)*", which is a top-level concept.
- BioTopLite2 (BTL2) [27], a biomedical top-level ontology that provides a rich set of constraining axioms to enforce the consistency of ontologies modeled thereunder. For example all clinical statements are defined as subclasses of 'Information Object' in BTL2.
- Clinical model reuses SNOMED CT in OWL and BioTopLite2, an ontology that is populated with instances generated according to the Content ODPs. For example, the definition of the OWL Class "ClinicalDiagnosticStatement" appears in the last row of Table 4.

2.5. Converting OWL Individuals from Content ODPs to FHIR RDF Using FHIR ShEx

The FHIR resources can be represented in the Turtle format [28]. FHIR has made available ShEx schemas in [6] that describe the RDF format and represent the Turtle schema [29]. Solbrig et al. [30] showed that "*ShEx is a good candidate for formally describing the FHIR metamodel when it is realized as RDF data instances*".

Table 5 partially reproduces the FHIR ShEx schemas used in this study:

- The first row corresponds to the ShEx schema "*measurements and simple assertions*" from [6] that is typically mapped to Content ODPs defined in the clinical model with the superclass "*ClinicalStatement*".
- The second row represents the ShEx schema "*detailed information about conditions, problems or diagnoses*" from [6] that is typically mapped to the Content ODP "*ClinicalDiagnosticStatement*" defined in the clinical model.
- The third row corresponds to the ShEx schema "*concept - reference to a terminology or just text*" from [6] that intends representing concepts and applies to the Coding data type [29].
- The fourth row represents the ShEx schema "*a measured or measurable amount*" from [6] that represents measurements and is typically mapped to the Content ODP "*MeasurementResultValue*" in the clinical model.

Table 5. Partial reproduction of FHIR ShEx schemas from [6]

FHIR ShEx schemas
<Observation> CLOSED { a [fhir:Observation]; … fhir:Observation.category @<CodeableConcept>*; fhir:Observation.code @<CodeableConcept>; fhir:Observation.subject @<Reference>?; … (fhir:Observation.valueQuantity @<Quantity> fhir:Observation.valueCodeableConcept @<CodeableConcept> | fhir:Observation.valueString @<string> | fhir:Observation.valueInteger @<integer> | …) fhir:Observation.dataAbsentReason @<CodeableConcept>?; … fhir:Observation.bodySite @<CodeableConcept>?; fhir:Observation.method @<CodeableConcept>?; …}
<Condition> CLOSED { a [fhir:Condition]; … fhir:Condition.category @<CodeableConcept>*; fhir:Condition.severity @<CodeableConcept>?; fhir:Condition.code @<CodeableConcept>?; fhir:Condition.bodySite @<CodeableConcept>*; fhir:Condition.subject @<Reference>; … }
<CodeableConcept> CLOSED { a NONLITERAL*; … fhir:CodeableConcept.coding @<Coding>*; fhir:CodeableConcept.text @<string>?; … }
<Quantity> CLOSED {… fhir:Quantity.value @<decimal>?; fhir:Quantity.unit @<string>?; fhir:Quantity.system @<uri>?; fhir:Quantity.code @<code>?; …}

In our approach, the values for the data elements in the FHIR ShEx schemas are:

- Static across HL7 resources (e.g. the value for the data element subject).
- Static and dependent on the Content ODPs (e.g. the value for category).
- Dynamic and dependent on the Content ODPs and SNOMED CT expression.

3. Applying the Proposed Ontology-Based Approach

Table 6 shows the number of OWL Individuals created by applying the Content ODPs to the 358 C-CDA coded entries with SNOMED CT expressions. The number of OWL Individuals created per C-CDA document differs. The number of OWL Individuals for one of the 18 C-CDA documents is 33 and the DL expressivity is ALE(D).

Table 6. Number of OWL Individuals created applying the Content ODPs introduced

Coded entry from HL7 C-CDA section	OWL Class Content ODP definition	Number of OWL individuals
History Present Illness	SymptomReportedPresentStatement	74
History Present Illness	SymptomReportedAbsentStatement	3
History Present Illness	EventStatement	4
Physical Exam	ClinicalFindingPresentStatement	127
Physical Exam	ClinicalFindingAbsentStatement	19
Physical Exam	ClinicalFindingUnknownStatement	2
Physical Exam	ObservableValueStatement	102
Assessment	ClinicalDiagnosticStatement	27

Table 7 illustrates some FHIR RDF data instances according to the FHIR ShEx schemas used in this study (see subsection 2.5 for details).

The first and second row in Table 7 show the values for the data elements that are static across the HL7 resources Observation and Condition in RDF. The values showed are static as the C-CDA documents are de-identified consultation notes.

Table 7. FHIR RDF representation leveraging on the FHIR ShEx schemas from [6]

FHIR RDF representation in Turtle format
fhir:Condition.subject [fhir:link <http://hl7.org/fhir/Patient/example>; fhir:Reference.reference [fhir:value "Patient/example"]];
fhir:Observation.subject [fhir:link <http://hl7.org/fhir/Patient/example>; fhir:Reference.reference [fhir:value "Patient/example"]];

fhir:Observation.category [fhir:index 0;
fhir:CodeableConcept.text [fhir:value "Signs and Symptoms"]];

fhir:Observation.category [fhir:index 0; fhir:CodeableConcept.coding [fhir:index 0;
fhir:Coding.system [fhir:value "http://terminology.hl7.org/CodeSystem/observation-category"];
fhir:Coding.code [fhir:value "exam"]; fhir:Coding.display [fhir:value "Exam"]]];

fhir:Condition.category [fhir:index 0; fhir:CodeableConcept.coding [fhir:index 0;
a sct:439401001; fhir:Coding.system [fhir:value "http://snomed.info/sct"];
fhir:Coding.code [fhir:value "439401001"]; fhir:Coding.display [fhir:value "diagnosis"]]];

fhir:Observation.code [fhir:CodeableConcept.coding [fhir:index 0; a sct:373428006;
fhir:Coding.system [fhir:value "http://snomed.info/sct"]; fhir:Coding.code [fhir:value "373428006"];
fhir:Coding.display [fhir:value "Corneal epithelial edema (finding)"]];
fhir:CodeableConcept.text [fhir:value "Corneal epithelial edema"]];

fhir:Condition.code [fhir:CodeableConcept.coding [fhir:index 0; a sct:392291006;
fhir:Coding.system [fhir:value "http://snomed.info/sct"]; fhir:Coding.code [fhir:value "392291006"];
fhir:Coding.display [fhir:value "Angle-closure glaucoma (disorder)"]]];

fhir:Observation.bodySite [fhir:CodeableConcept.coding [fhir:index 0; a sct:18944008;
fhir:Coding.system [fhir:value "http://snomed.info/sct"]; fhir:Coding.code [fhir:value "18944008"];
fhir:Coding.display [fhir:value "Right eye structure (body structure)"]];

```
      fhir:CodeableConcept.text [ fhir:value "Right eye" ] ];
  fhir:Observation.valueQuantity [ fhir:Quantity.value [ fhir:value "12"^^xsd:decimal ];
     fhir:Quantity.unit [ fhir:value "mmHg" ]; fhir:Quantity.system [ fhir:value "http://unitsofmeasure.org" ];
      fhir:Quantity.code [ fhir:value "mm[Hg]" ] ];
```

Rows 3 to 5 in Table 7 illustrate the values for the data elements that are static and dependent on the Content ODPs. The third row shows a value for the data element Observation.category indicating that the FHIR Observation resource is mapped to a coded entry from the HL7 C-CDA section *History Present Illness*. Likewise, the fourth row shows a value for the data element Observation.category indicating that the FHIR Observation resource is mapped to a coded entry from the HL7 C-CDA section *Physical Exam*. In the same vein, the fifth row shows a value for the data element Condition.category indicating that the FHIR Condition resource is mapped to a coded entry from the HL7 C-CDA section *Assessment*, i.e. clinician-asserted diagnosis.

Rows 6 to 9 in Table 7 illustrate the values for the data elements that are dynamic and dependent on the Content ODPs and SNOMED CT expression. The row 6 and 7 follow the ShEx schema "*concept - reference to a terminology or just text*" from [6], which appears in the third row of Table 5. Row 8 shows how the two SNOMED CT concepts "*8966001|Left eye structure*" and "*18944008|Right eye structure*" are systematically used in this study as valid coded values for the data elements Observation.bodySite and Condition.bodySite in FHIR (see 2.3 for details). Row 9 follows the ShEx schema "*a measured or measurable amount*" from [6].

4. Exploring the Benefits of an Ontology-Based Approach

4.1. One SPARQL Query to Retrieve OWL Individuals Created with Content ODPs

The first row of Table 8 shows an SPARQL [31] SELECT query that can retrieve the OWL Individuals created with the Content ODPs. For the query to work, it is necessary to specify the values for the three variables appearing in bold within a constraint expressed by the keyword FILTER (i.e. the values for **?x1** and **?x2** and **?C3**). The rows 2 to 5 show how the FILTER can handle four different Content ODPs (e.g. in row 2 the FILTER works with the Content ODP ClinicalFindingPresentStatement).

Table 8. An SPARQL SELECT query to retrieve OWL Individuals created with different Content ODPs

SPARQL SELECT query and exemplifying the use of the variables within the FILTER constraint
SELECT DISTINCT ?D ?E WHERE { ?y rdf:type owl:NamedIndividual; rdf:type ?C1 ; btl2:represents ?D . ?C1 rdf:type owl:Class ; owl:intersectionOf/rdf:rest*/rdf:first ?x1 . ?C1 rdf:type owl:Class ; owl:intersectionOf [rdf:rest*/rdf:rest ?z1]. ?z1 rdf:first ?w1 . ?w1 a owl:Restriction ; owl:onProperty ?p1 ; owl:someValuesFrom ?C2 . ?C2 rdf:type owl:Class ; owl:intersectionOf/rdf:rest*/rdf:first ?x2 . ?C2 rdf:type owl:Class ; owl:intersectionOf [rdf:rest*/rdf:rest ?z2]. ?z2 rdf:first ?w2 . ?w2 a owl:Restriction ; owl:onProperty ?p2 ; owl:someValuesFrom ?C3 . OPTIONAL { ?D rdf:type owl:Class ; skos:prefLabel ?E } . FILTER (**?x1 =** && **?x2 =** && **?C3 =**) }
FILTER (?x1 = cm:ClinicalFindingPresentStatement && ?x2 = sct:55468007 && ?C3 = sct:18944008)
FILTER (?x1 = cm:ClinicalFindingAbsentStatement && ?x2 = sct:55468007 && ?C3 = sct:18944008)
FILTER (?x1 = cm:ClinicalDiagnosticStatement && ?x2 = sct:103693007 && ?C3 = sct:18944008)

FILTER (?x1 = cm:ObservableValueStatement && ?x2 = sct:252832004 && ?C3 = sct:18944008)

The SPARQL query introduced uses as value for **?x1** within the FILTER constraint the OWL Classes from the clinical model in OWL (prefix cm) where the Content ODPs are created. The value for **?x2** is an OWL Class from the hierarchy *Procedure* (i.e. the SNOMED CT top-level concept "*71388002| Procedure (procedure)*") from the SNOMED CT ontology in OWL (prefix sct). Finally, the value for **?C3** can be any pre- or post-coordinated SNOMED CT expression that can be assigned to the SNOMED CT attribute "363704007|Procedure site (attribute)". Hence, the query does not lack generalisation, although in this study, only two pre-coordinated expressions are typically considered as values for **?C3**, i.e. two OWL Classes corresponding to the SNOMED CT concepts "8966001|Left eye structure" and "18944008|Right eye structure" (see 2.3 for details).

Table 9 shows the results of the SPARQL query over the OWL Individuals mapped to the coded entries of a single C-CDA document (one of the 18 C-CDA documents used in this study). The first column has the values for the three variables within the FILTER constraint. The second and third columns are the values for the two variables **?D** and **?E** from the SELECT clause (see first row of Table 8).

The last two columns in Table 9 contain the results of the query using ARQ [32], which is a query engine for Jena that supports the SPARQL query language. The value for **?D** in the last row of Table 9 contains the prefix cn meaning consultation note. The prefix sctpost that appear in some values of **?D** (second column) indicates that it is a SNOMED CT post-coordinated expression, while the prefix sct in a value of **?D** means it is a SNOMED CT pre-coordinated expression.

Table 9. Results of the SPARQL query when executed over the OWL instances mapped to coded entries included in C-CDA sections of a single C-CDA document

Variables from FILTER constraint	Variable from SELECT ?D	Variable from SELECT ?E
?x1 = cm:ClinicalFindingPresentStatement ?x2 = sct:55468007 ?C3 = sct:18944008	sctpost:SCTexp_85 sct:193894004 sctpost:SCTexp_84	"2+ Tyndall" "conjunctival hyperemia" "2+ limbal injection"
?x1 = cm:ClinicalFindingAbsentStatement ?x2 = sct:55468007 ?C3 = sct:18944008	sctpost:SCTexp_89 sct:87807004 sct:246998009 sct:78778007	"anterior vitritis" "hypopyon" "keratic precipitates" "synechiae of iris"
?x1 = cm:ClinicalDiagnosticStatement ?x2 = sct:103693007 ?C3 = sct:18944008	sctpost:SCTexp_231	"Acute anterior uveitis"
?x1 = cm:ObservableValueStatement ?x2 = sct:252832004 ?C3 = sct:18944008	cn:MeasureResult_12mmHg	

The SPARQL SELECT query presented can be used to query OWL Individuals mapped to the clinical statements coded, which would normally be included in C-CDA sections. Although we have illustrated the results of the query over OWL Individuals that have one C-CDA document as source (see Table 9), the query can be run over a larger number of OWL individuals mapped to the coded entries of a set of C-CDA documents. Furthermore, the execution of this query over a large number of OWL Individuals is guaranteed as SPARQL 1.1 has a federated query extension [33], which allows executing queries distributed over different SPARQL endpoints.

Other SPARQL SELECT queries can be created; we have just illustrated how one query can be used across Content ODPs, while the query scalability is guaranteed.

4.2. Natural Language Generation Exploiting Content ODPs

NLG systems can generate texts in English from computer-accessible data [10], which overall implies two major tasks [10]: 1) content determination; and 2) text-planning, i.e. organising the content to be communicated into a rhetorically coherent structure.

In this study, content determination is achieved by SPARQL queries like the one presented in the previous subsection. For text-planning, two subtasks are being done:

- Sentence planning – leveraging on the labels assigned to pre- and post-coordinated SNOMED CT expressions as the key clinical content to be included in the different types of sentences that need to be generated.
- Realisation – to generate the individual sentences, a set of fill-in-the-blank templates was created. Table 10 shows three of the fill-in-the-blank templates.

Table 10. Exemplifying the definition and use of three of the fill-in-the-blank templates created

Definition of the fill-in-the-blank templates	**Sentence planning and realisation applying the templates**
[Descendant of "71388002| Procedure (procedure)"] of [Value of "363704007|Procedure site (attribute)"] revealed [OWL Individuals of cm:ClinicalFindingPresentStatement] [OWL Individuals of cm:ClinicalFindingAbsentStatement].	Slit lamp biomicroscopy of right eye revealed conjunctival hyperemia, 2+ limbal injection, 2+ Tyndall, no keratic precipitates, no hypopyon, no anterior vitritis, and no synechiae of iris.
The established diagnosis is [Value of "263502005|Clinical course (attribute)"] [Descendant of "64572001|Disease (disorder)"] in [Value of "363704007|Procedure site (attribute)"].	The established diagnosis is acute anterior uveitis in right eye.
[Descendant of "386053000|Evaluation procedure (procedure)"]: StringSplit([OWL Individuals of cm:ObservableResultValue]) [Value of "363704007|Procedure site (attribute)"].	Intraocular pressure: 12mmHg right eye.

The templates defined in Table 10 (left column) use the results of the SPARQL SELECT query that appear in Table 9 (last column). The template in the first row of Table 10 used the values of the variable **?E** appearing in the first and second row of Table 9. Likewise, the template in the second row of Table 10 used the values of the variable **?E** appearing in the third row of Table 9. Finally, the template in the last row of Table 10 used the values of the variable **?D** from the last row of Table 9.

The SPARQL SELECT query (first row of Table 8) has a constraint with the keyword OPTIONAL that uses the annotation property skos:prefLabel. This property is used to stored terms that are the preferred option for reporting. For example, the SNOMED CT concept “246993000|Anterior chamber cells (finding)" is also known as “Tyndall effect” or “Tyndall” for short when reporting. Another example is the post-coordinated expression depicted in Figure 3, the part of the SNOMED CT expression that appears in the octagon within the *clinical kernel* is known as “anterior vitritis”.

5. Conclusion

This paper has presented the preliminary results of an on-going investigation that utilises Content ODPs and FHIR ShEx schemas for converting HL7 C-CDA document entries into FHIR RDF data instances. This paper concentrates only on the HL7 C-CDA document entries (from three document sections) that are typically the clinical statements coded in SNOMED CT. We have demonstrated the viability of our ontology-based approach with 18 of 103 HL7 C-CDA documents related to the *Red Eye* domain. Future work intends to improve Content ODPs presented in Table 4.

Having SNOMED CT expressions underpinned by Content ODPs allows the creation of one SPARQL SELECT query that can be executed over a myriad of OWL Individuals and multiple Content ODPs. The query can facilitate both large-scale data analytics and NLG. A tangible benefit of a "systematic" codification of clinical information by Content ODPs is that SNOMED CT expressions become predictable, reducing the different types of sentences that need to be generated, and therefore, making realisation (a NLG task) computationally simpler and cheaper.

Acknowledgements

Partially funded by Precise4Q (GA:777107) H2020-SC1-2017-CNECT-2.

References

[1] HL7 C-CDA on FHIR Implementation Guide. http://www.hl7.org/fhir/us/ccda/
[2] RDF. https://www.w3.org/TR/rdf11-concepts/
[3] Egaña, M., et al., Applying ontology design patterns in bio-ontologies. In: EKAW (2008), 7-16.
[4] Falbo, R., et al., Ontology patterns: clarifying concepts and terminology. In: WOP (2013), 14-26.
[5] Shape Expressions Language. http://shex.io/
[6] FHIR ShEx Schemas. https://www.hl7.org/fhir/downloads.html
[7] SNOMED CT. https://confluence.ihtsdotools.org/display/DOCSTART/
[8] NHS 2020. https://www.gov.uk/government/publications/personalised-health-and-care-2020
[9] Kruse, C.S., et al., Challenges and opportunities of big data in health care: a systematic review. JMIR medical informatics, **4** (2016), 38.
[10] Reiter, E. and Dale, R., Building natural language generation systems. Cambridge Univ. press. 2000.
[11] Kung, J.S., et al., Cataract surgery in the glaucoma patient. Middle East African journal of ophthalmology, **22** (2015), 10.
[12] Arndt, B.G., et al., Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. The Annals of Family Medicine, **15** (2017), 419-426.
[13] Andonegui Navarro J., Ocular manifestations of systemic diseases. Anales del Sistema Sanitario de Navarra, ISSN 1137-6627 (2009), 7-11.
[14] Facts About Diabetic Eye Disease. https://nei.nih.gov/health/diabetic/retinopathy
[15] SNOMED CT Terminology Services Guide. https://confluence.ihtsdotools.org/display/DOCTSG/
[16] Rector, A., Qamar, R. and Marley, T., Binding ontologies & coding systems to electronic health records and messages. In: Formal Biomedical Knowledge Representation (2006).
[17] SNOMED CT expressions. https://confluence.ihtsdotools.org/display/DOCTSG/12.2+Expression+Parts
[18] Bhattacharyya, S.B., SNOMED CT Expressions. In Introduction to SNOMED CT (2016), 95-129.
[19] FHIR resource Observation. https://www.hl7.org/fhir/observation.html
[20] FHIR resource Condition. https://www.hl7.org/fhir/condition.html
[21] FHIR resource Body Structure. https://www.hl7.org/fhir/bodystructure.html
[22] FHIR Observation mappings. https://www.hl7.org/fhir/observation-mappings.html
[23] FHIR Condition mappings. https://www.hl7.org/fhir/condition-mappings.html
[24] OWL. https://www.w3.org/TR/owl2-syntax/
[25] Horridge, M., et al., The Manchester OWL syntax. In: OWLed (2006).

[26] M. Arguello Casteleiro, et al., Experiments to create ontology-based disease models for diabetic retinopathy from different biomedical resources. In: SWAT4HCLS (2017).

[27] Schulz S, Boeker M, Martinez-Costa C., The BioTop Family of Upper Level Ontological Resources for Biomedicine. Stud Health Technol Inform **235** (2017), 441-445.

[28] Turtle. https://www.w3.org/TR/turtle/

[29] FHIR RDF representation. https://www.hl7.org/fhir/rdf.html

[30] Solbrig, H.R., et al., Modeling and validating HL7 FHIR profiles using semantic web Shape Expressions (ShEx). J Biomed Inform., **67** (2017), 90-100.

[31] SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/

[32] Apache Jena ARQ. https://jena.apache.org/documentation/query/index.html

[33] SPARQL 1.1 Federated Query. https://www.w3.org/TR/sparql11-federated-query/