JAMIA Open, 4(2), 2021, 1–9
doi: 10.1093/jamiaopen/ooab035
Research and Applications

Research and Applications

Aligning an interface terminology to the Logical Observation Identifiers Names and Codes (LOINC®)

Jean Noël Nikiema [1,2,3] Romain Griffier,[1,4] Vianney Jouhet [1,4] and Fleur Mougin [1]

[1]Univ. Bordeaux, Inserm, BPH, U1219, Team ERIAS, F-33000 Bordeaux, France, [2]Research Center, Centre hospitalier de l'Université de Montréal, Montréal, Québec, Canada, [3]Department of Management, Evaluation and Health Policy, School of Public Health, Université de Montréal, Montréal, Québec, Canada, and [4]CHU de Bordeaux, Pole de santé publique, Service d'information médicale, F-33000 Bordeaux, France

Jean Noël Nikiema, Vianney Jouhet, and Fleur Mougin contributed equally to this work.
Corresponding Author: Jean Noël Nikiema, Univ. Bordeaux, Inserm, BPH, U1219, team ERIAS, F-33000 Bordeaux, France; jean.nikiema@umontreal.ca

Received 11 November 2020; Revised 4 March 2021; Editorial Decision 26 March 2021; Accepted 15 April 2021

ABSTRACT

Objective: Our study consists in aligning the interface terminology of the Bordeaux university hospital (TLAB) to the Logical Observation Identifiers Names and Codes (LOINC). The objective was to facilitate the shared and integrated use of biological results with other health information systems.
Materials and Methods: We used an innovative approach based on a decomposition and re-composition of LOINC concepts according to the transversal relations that may be described between LOINC concepts and their definitional attributes. TLAB entities were first anchored to LOINC attributes and then aligned to LOINC concepts through the appropriate combination of definitional attributes. Finally, using laboratory results of the Bordeaux data-warehouse, an instance-based filtering process has been applied.
Results: We found a small overlap between the tokens constituting the labels of TLAB and LOINC. However, the TLAB entities have been easily aligned to LOINC attributes. Thus, 99.8% of TLAB entities have been related to a LOINC analyte and 61.0% to a LOINC system. A total of 55.4% of used TLAB entities in the hospital data-warehouse have been mapped to LOINC concepts. We performed a manual evaluation of all 1-1 mappings between TLAB entities and LOINC concepts and obtained a precision of 0.59.
Conclusion: We aligned TLAB and LOINC with reasonable performances, given the poor quality of TLAB labels. In terms of interoperability, the alignment of interface terminologies with LOINC could be improved through a more formal LOINC structure. This would allow queries on LOINC attributes rather than on LOINC concepts only.

Key words: LOINC, interface terminology, alignment process

OBJECTIVE

Interface terminologies are controlled vocabularies whose common definition in the biomedical domain is the following: "*a systematic collection of health care-related phrases (terms) that supports clinicians' entry of patient-related information into computer programs*".[1,2] Indeed, this type of terminologies is created for the specific use of certain healthcare structures. If the usability of interface terminologies is important for the health information systems in which they are developed, their use may be limited in an integrated way. For in-

Lay summary

Our study consisted in aligning the interface terminology of the Bordeaux university hospital (TLAB) to Logical Observation Identifiers Names and Codes (LOINC), making LOINC concepts the semantic support for sharing data encoded with TLAB. The alignment is based on an algorithm that links LOINC concepts and TLAB entities by highlighting their common definitional elements. The algorithm takes into account the difference in the granularity of definitions between LOINC and TLAB. The process points out that while LOINC can be useful for disambiguating information, its complexity can limit its alignment to interface terminologies. However, both resources need to be used together: interface terminologies for their usability, LOINC for interoperability. Querying results based on a specific set of LOINC parts may be more efficient than using LOINC concepts and their complex labels as such. Thus, The usability of LOINC as an interoperability tool can be improved by a more formal structure.

teroperability purposes, interface terminologies have to be aligned to reference terminologies,[1,3] that are consensual terminologies whose terms and structures are validated by the scientific community. Thus, aligning an interface terminology to a reference terminology is required for sharing data between different health information systems.[4,5] In the literature, many works have been concerned with the alignment of interface terminologies to reference terminologies.[2,6] From these works, it turns out that the ideal way to get an interface terminology aligned to a reference one is to directly create the interface terminology from a reference terminology.[7–9] Most of the time, this strategy cannot be applied because interface terminologies are usually created manually using items of historical paper forms.[4] Consequently, it is necessary to reuse techniques commonly proposed in the literature for finding correspondences between terminologies or ontologies (eg, morphosyntactic, structural and extensional techniques).[10] These techniques have to be adapted for dealing with issues related to interface terminologies, such as the presence of noisy labels and the lack of structure. At the Bordeaux university hospital, such an interface terminology is used for encoding and retrieving the results of biomedical analyses. This interface terminology is herein referred to by its French acronym TLAB for "Terminologie Locale d'Analyses Biomédicales" (ie, Local Terminology for Biomedical Analyses). Thus, as for other interface terminologies, aligning TLAB with a reference terminology describing laboratory observations is a requirement.

Many characteristics can induce the selection of a reference terminology as a support for sharing information. Some reference terminologies have been created and/or recommended by the World Health Organization (WHO). (Like the 10th revision of the International statistical Classification of Diseases and related health problems [ICD-10] that is used worldwide for epidemiology purpose.) Nevertheless, the novelty and the quality of some terminologies have imposed themselves as a reference in their sub-domain. The Logical Observation Identifiers Names & Codes (LOINC) is an example of such terminologies for recording laboratory observations that is used in many countries.[11,12] Containing consensual and validated terms of this domain, LOINC is a reference terminology. Thus, many works have been concerned by the alignment of interface terminologies to LOINC,[13–16] positioning LOINC as an international support terminology for sharing information about laboratory observations across different healthcare systems. The aim of this work was thus to align TLAB to LOINC because of its wide-scale adoption and use for representing biological analyses in a standardized way.

BACKGROUND AND SIGNIFICANCE

In this section, we present the characteristics of TLAB and LOINC and expose the techniques that can be used for their alignment.

The terminologies to be aligned

TLAB

The interface terminology used at the Bordeaux university hospital for encoding data of medical test laboratories has been exported from the electronic health record system of the hospital. TLAB labels were recorded manually in French by healthcare professionals. The space limits in the recording step lead to non-conventional abbreviations of labels (eg, ***PCR.C.TRACHO/GENI*** for "Recherche par réaction en chaîne par polymérase de Chlamydia trachomatis au niveau génital" translatable as "genital Chlamydia trachomatis polymerase chain reaction (PCR) testing" in English). TLAB is a multi-hierarchical terminology composed of 8285 entities. These entities are hierarchically organized and rooted to 15 top-level entities (Anatomie et Cytologie Pathologiques [Pathological Anatomy and Cytology], Bactériologie [Bacteriology], Biochimie [Biochemistry], Immuno-hématologie EFS [Immunohematology], Génétique [Genetic], Hématologie [Hematology], Immunologie—Immunogénétique [Immunology—Immunogenetics], Mycologie—Parasitologie [Mycology—Parasitology], Hormonologie—Marqueurs tumoraux [Hormonology—Tumor markers], Biologie de la reproduction [Reproductive biology], Pharmacologie—Toxicologie [Pharmacology—Toxicology], Recherche [Research], Biologie des tumeurs [Tumor biology], Virologie [Virology], Hygiène hospitalière [Hospital hygiene]) that correspond to the different domains of biological analyses. As a result, the TLAB terminology corresponds to a set of 8300 entities. The absence of a formal definition for TLAB entities makes the Simple Knowledge Organization System (SKOS [https://www.w3.org/TR/skos-reference/]) format adequate to represent TLAB.[17] Thus, TLAB entities have been described each as a *skos: Concept* and their hierarchical relations have been defined through the *skos: broader* and *skos: narrower* relationships. Each entity corresponds to a unique alphanumeric code related to a label using the *skos: prefLabel* attribute (eg, a TLAB entity code: "syn-ana-vrku1" and its corresponding label: *PCR BK/urines*).

LOINC®

LOINC® is a reference terminology created and maintained by the Regenstrief Institute.[12] Published in 1995,[18] the first release of LOINC contained only codes related to laboratory testing. Nowadays, LOINC is a clinical terminology used for recording health measurements, observations, and documents.[15] The codes are being hereafter designated as "LOINC concepts".

The LOINC concepts belong to a specific class and are defined using the six major attributes (component/analyte, property, time, system, scale, method) and four minor attributes (challenge, adjustments, time modifier, super-system) designated as "LOINC attributes" (Figure 1 and details provided in Supplementary Appendix 1).

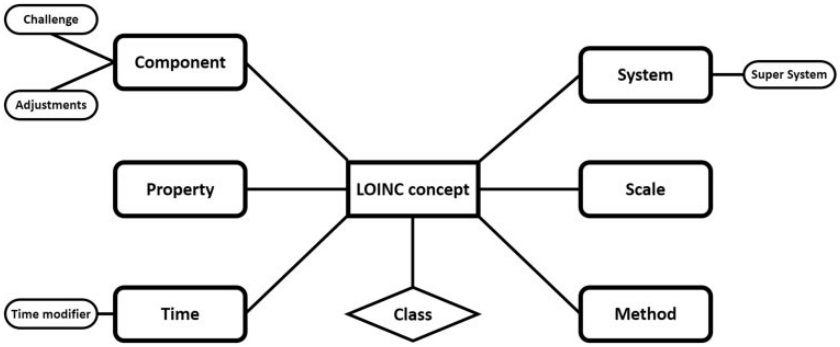

Figure 1. The description model of LOINC concepts. The model contains six mandatory attributes (rectangles with rounded corners): four optional attributes (ovals) to refine the description of three mandatory attributes (component, system and time) and a class (rhombus).

The labels of LOINC concepts were originally available in English. For an alignment to TLAB, it was necessary to focus on its available translated labels.

The alignment strategies

Aligning interface terminologies to reference terminologies is an important and time-consuming task, requiring automatic strategies.[13] The common automatic approaches used to perform alignment are lexical, structural, and instance-based.[19] However, the strategy to be implemented for the alignment of TLAB and LOINC had necessarily to deal with the differences between their structure and the absence of overlap between terms available in both terminologies, as well as the lack of quality that exists in interface terminology labels,[20,21] such as in TLAB labels.

Existing alignment approaches to LOINC

Many works have been described in the literature using LOINC as the reference terminology for the mapping of laboratory terms.[15,22–26] Three main strategies are generally followed to establish these mappings:

- The **manual alignment** of interface terminologies to LOINC,[22] which is a tedious task that is not reasonable to implement when dealing with large interface terminologies.
- The **use of the Regenstrief LOINC Mapping Assistant (RELMA)**,[15,24,26] that is an open access mapping tool provided by the Regenstrief Institute for the mapping of local terms (ie, terms available in interface terminologies or in corpora of documents) to LOINC concepts.[16] RELMA uses a morphosyntactic strategy with a manual correction of mappings, thus needing users' intervention.[13] In practice, the tool firstly proposes LOINC concepts as potential equivalences for local labels (one at a time). Then, a validation is requested from the users, or an alternative label entry is proposed when no LOINC concept is found.
- The **use of home-made algorithms**.[13,24,25] Like RELMA, the other mapping strategies are based on morphosyntactic approaches, sometimes combined with/improved by machine learning algorithms. Existing approaches were however deemed ineffective to deal with noisy labels. Indeed, authors that used home-made algorithms (including machine learning algorithms) and/or RELMA reported that the variation of local terms and the incompleteness of their description in interface terminologies are the main issues altering the quality of mappings. To compensate for these limitations, some of these authors cleaned and enhanced manually the terms in interface terminologies.[16,24]

All the applied strategies were designed rather to increase the number of obtained mappings than to obtain an optimal semantic quality of resulting mappings. Thus, erroneous mappings were not overcome by existing automatic processes.

In practice, no results were obtained when using RELMA for the alignment of TLAB labels to LOINC. We believed that the use of the structure of LOINC labels as a validation element of the mappings could be a solution to address these issues. Thus, the goal of our work was to implement a specific and semi-automatic process for the alignment of TLAB to LOINC by using a TLAB label correction step and taking into account the structure of LOINC for the validation of mappings.

Pre-processing of TLAB labels

The morphosyntactic approach is the common initial step of all automatic mapping processes. Such approaches are limited for interface terminologies such as TLAB due to the poor quality of their labels.[20] Pre-processing is therefore necessary to improve the efficiency of mapping strategies. For interface terminologies, it is sometimes possible to find guidelines describing the naming conventions of their labels.[21] If such guidelines are not available (which is the case for TLAB), strategies developed for processing texts available in fora, social networks, and Short Message Systems (SMS) can be used to improve the quality of local labels.[27–29] These strategies, including the detection and correction of non-standard-words, are described in detail in ref.30 and have been applied to TLAB labels.

MATERIALS

The graph model of LOINC in SKOS containing French labels

For the alignment process, we used the 2.65 version of LOINC (https://loinc.org/). This release contains CSV format tables for the description of each LOINC part, the linguistic variants of LOINC labels and the multi-axial hierarchy of LOINC. "*The atomic elements that make up each LOINC term name are called Parts.*"[12] LOINC parts mainly correspond to LOINC attributes to which an identifier is assigned.

Using the structure of LOINC induced the necessity to describe it in a computational language. Exploring the constructed structure of LOINC in the state-of-art, we found constructed SKOS structures in Bioloinc (https://bioloinc.fr/bioloinc/KB/#Group:uri=http://aphp.fr/Bioloinc/JDV_LOINC_Biologie;tab=props) and BioPortal (https://bioportal.bioontology.org/ontologies/LOINC). However, both constructions are based on the LOINC part table (LPT) that describes

LOINC concepts and their related parts, including ambiguous descriptions of LOINC parts within LOINC concepts related to many attributes of the same type (eg, 13505-3-*Herpes simplex virus 1 + 2 Ab pattern [Interpretation] in serum* is related to two components, being LP14822-8-*Herpes simplex virus 1 + 2* and LP40415-9-*Herpes simplex virus 1 + 2 Ab pattern*[31]) In addition, the attributes described within Bioloinc correspond to a simple tokenization of LOINC labels (with labels as the identifiers of LOINC attributes). For these reasons, we have chosen to build our own SKOS format of LOINC whose structure contained:

- non-ambiguous relations between LOINC concepts and LOINC attributes,
- attributes described with all the French variant labels available in the release.

The strategy used to construct this structure was detailed in ref.31

The ServoMap tool

ServoMap is a mapping tool developed by Diallo.[32] It is a highly configurable large scale ontology matching system, which is able to process large terminologies. ServoMap is based on Lucene, measures morphosyntactic similarity and provides equivalence mappings between entities of two terminologies.[33] We used the latest version of ServoMap in our alignment process.

METHODS

To realize the alignment of TLAB to LOINC, we have developed a method based on the LOINC structure, leveraging:

- the ability to decompose a LOINC concept into its constitutive attributes,
- the hierarchical structure of LOINC.

The following three steps have been performed:

1. The mapping of tokens constituting the labels of concepts in both terminologies,
2. The anchoring step for identifying: (i) the mappings between TLAB entities and the attributes of LOINC concepts, and (ii) the mappings between TLAB entities and LOINC concepts, and
3. The instance-based filtering of the obtained mappings: a data-driven validation process.

The mapping of tokens

The tokenization process consisted in splitting the labels of TLAB and LOINC according to white-spaces and punctuation. Stop-words were removed using an existing list of French stop-words (https://github.com/stopwords-iso/stopwords-fr). As a result, we obtained a set of tokens linked to TLAB entities on the one hand, and to LOINC attributes on the other.

We then used the ServoMap tool in order to map tokens that were extracted (Figure 2). In this frame, the cardinality of mappings between TLAB and LOINC tokens has been computed.

The anchoring step

The anchoring step was 2-fold: (i) the anchoring of TLAB entities to LOINC attributes (Figure 3a), followed by (ii) the anchoring of TLAB entities to LOINC concepts (Figure 3b).

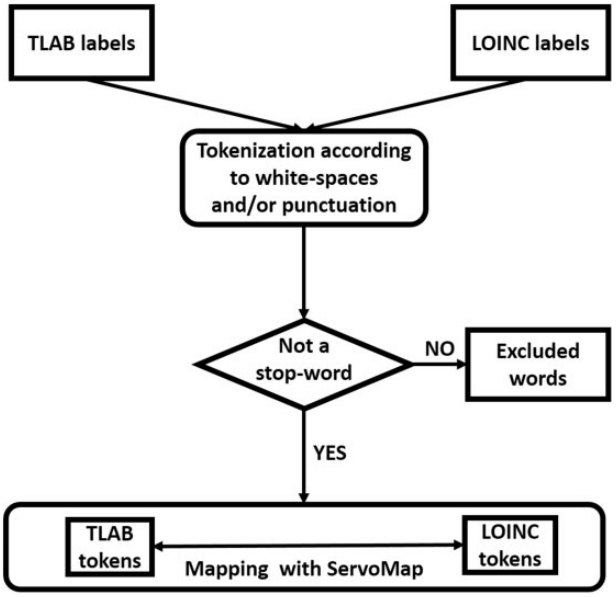

Figure 2. Mapping of TLAB and LOINC tokens. Tokens of TLAB and LOINC are words, excluding stop-words, which are found using a tokenization process applied to their labels based on white-spaces and punctuation.

The anchoring to LOINC attributes

The objective of this stage was to identify definitional attributes for TLAB entities. The mapped tokens constituted bridges between TLAB entities and LOINC concepts' attributes. For each type of attributes, when a TLAB entity was mapped to multiple LOINC attributes, we chose the attribute(s) having the highest number of tokens in common with the description of this TLAB entity. Then, the attribute(s) related to a TLAB entity were propagated to all its descendants.

The anchoring to LOINC concepts

In this stage, we firstly identified the candidate anchors that correspond to LOINC concepts and TLAB entities sharing the same analyte. Then, we filtered these correspondences according to classes, systems, and methods' hierarchies. Thus, the mappings involving entities that belonged to distinct classes, systems or methods were excluded. For the last step, when a TLAB entity was not related to any LOINC method, we validated only the anchored LOINC concept(s) that did not exhibit any method attribute.

The filtering of anchors and validation of mappings

To remove erroneous mappings, we conducted an instance-based filtering process by using lab test results coming from the Bordeaux university hospital's data warehouse. The Bordeaux university hospital uses a health data warehouse (based on i2b2 [https://www.i2-b2.org/]), which gathers various structured and unstructured data (clinical data, prescriptions and administration data of medicinal products, biological data, medical imaging data, anatomopathological data, and administrative data) for patients who have been visiting the hospital at least once since 2010. At the May 31, 2019, the collected data concerned 1 591 272 patients corresponding to 11 637 437 visits and 1 152 516 900 observations. Biological data represented 29.3% of all available data (337 860 938 observations). Among these biological results, 279 065 808 (82.6%) were encoded with the TLAB terminology (the remaining observations being de-

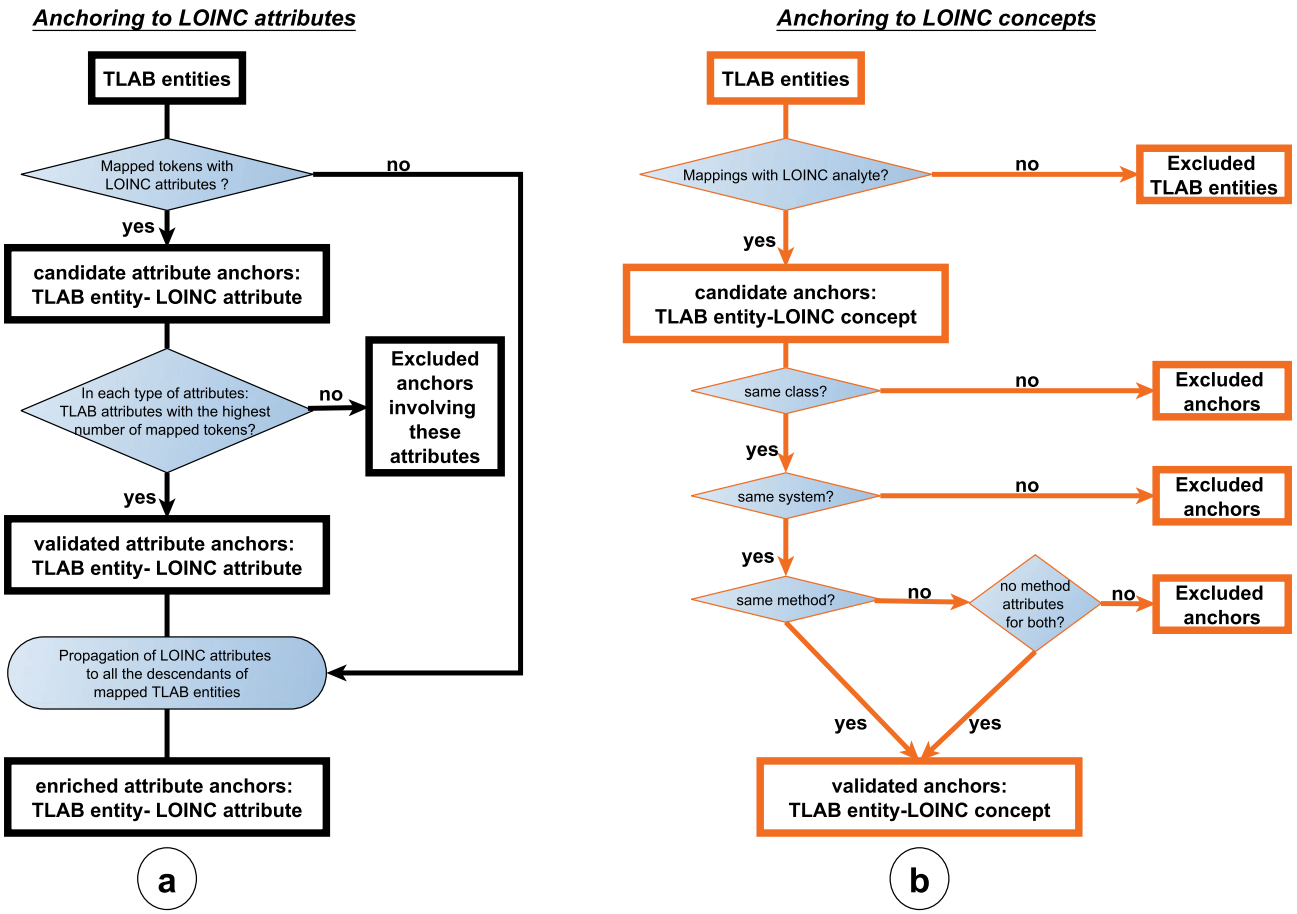

Figure 3. Anchoring of TLAB entities to: (a) LOINC attributes, and (b) LOINC concepts.

scribed in other interface terminologies for specific needs, being out of the scope of this paper).

To carry out this process, we applied a pragmatic filtering of the resulting mappings. Among all TLAB entities, we eliminated those that were not used at all to encode lab results within the data warehouse, considering that those entities were useless. For the remaining TLAB entities (those that were effectively used), we extracted their related property and measurement scale from the lab results (that were not available in the TLAB terminology).

The following three steps were performed for this process (Figure 4):

1. We first annotated the values and units of measure available in the laboratory data with the Unified Code for Units of Measure (UCUM [https://unitsofmeasure.org/ucum.html]) codes. This annotation was realized using a simple morphosyntactic technique for mapping automatically the UCUM codes and the units of measure found in the lab results.
2. Each UCUM code is related to a property describing the type of measure. We then manually mapped these UCUM properties to the property attributes of LOINC. Thus, this mapping led to a description of TLAB entities used in the laboratory data according to some validated LOINC properties.
3. Finally, for each TLAB entity, we validated the anchored LOINC concepts that exhibited the same LOINC property.

Based on these results, we manually curated all the 1-1 mappings (1 TLAB entity mapped to 1 LOINC concept) and computed the precision of results. The validation was realized in a consensual way by two medical doctors (R.G. and J.N.N.), having both medical and knowledge representation backgrounds. The experts searched for equivalences between TLAB entities and LOINC concepts or determined if a hierarchical relation existed between them.

RESULTS

The mapping of tokens

The tokenization process resulted in 4735 and 12 737 unique tokens for TLAB and LOINC, respectively. The mapping process identified 2346 (49.5%) TLAB tokens mapped to 2410 (18.9%) LOINC tokens. Table 1 describes the cardinality of mappings between TLAB and LOINC tokens.

The anchoring step

From the mapping of tokens, we inferred triplets that are composed of a TLAB entity, an attribute relation and a LOINC attribute. The first inference corresponded to 9 217 089 triplets (7808 TLAB entities related to 39 929 LOINC attributes). These triplets have been reduced to 1 365 129 (7808 TLAB entities related to 39 152 LOINC attributes) after considering, for the same type of attribute, the LOINC attribute(s) that shared the highest number of tokens with TLAB entities. As an example, for the TLAB entity syn-ana-vtal1-*PCR Adéno/LCR* (an alternative label created by the pre-processing step was "réaction en chaîne par polymérase Adéno/liquide céphalo-

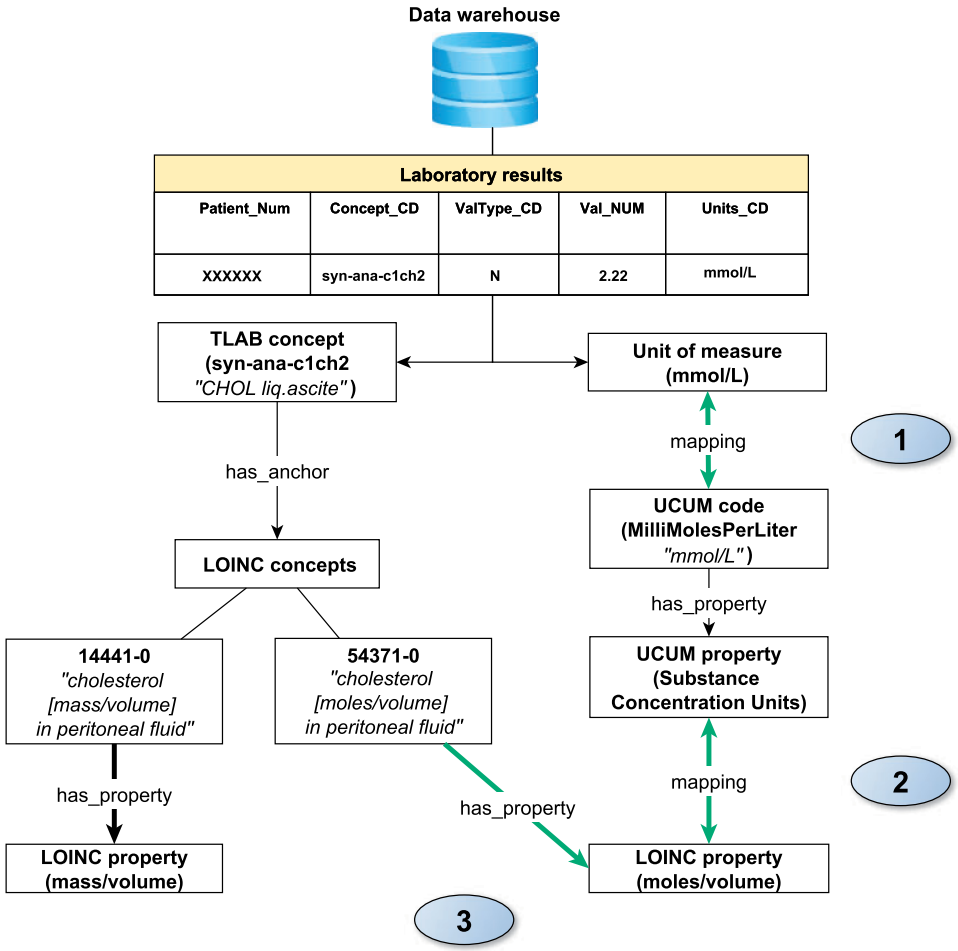

Figure 4. Instance-based filtering: instantiation of laboratory results using mapped LOINC concepts. (1) Mapping of units of measure in the data warehouse to UCUM codes, (2) mapping of UCUM codes' properties to LOINC properties, (3) validation of the mappings between anchored concepts of LOINC and TLAB that share the same LOINC property. For example, syn-ana-c1ch2-cholesterol dans le liquide d'ascite (ie, cholesterol in ascites fluid) is used in the data warehouse with "mmol/L" as a unit of measure. Consequently, only the LOINC concept, 54371-0-cholesterol (moles/volume) in peritoneal fluid, which shares the appropriate LOINC property with this unit of measure, can be used to instantiate the results.

Table 1. Distribution of TLAB and LOINC tokens according to the cardinality of resulting mappings

		TLAB tokens	LOINC tokens
Cardinality	1-0	2389 (50.5%)	10 327 (81.1%)
	1-1	2226 (47.0%)	2347 (18.4%)
	1-N	120 (2.5%)	63 (0.5%)
Total		4735 (100.0%)	12 737 (100.0%)

Note: 1-0 represent tokens without mappings; 1-1 mappings represent tokens having only one mapping; and 1-N represent tokens having multiple mappings.

Table 2. Distribution of TLAB entities according to their anchored LOINC attributes

Anchored attributes	Direct anchors[a]	Extended anchors[b]
Component	7688	8285
Challenge	3362	4561
Adjustment	348	1093
Property	1656	2374
Time	462	794
Time modifier	0	0
System	3371	5065
Super system	502	767
Scale	139	211
Method	4949	7944
Class	2262	7137

[a]Direct anchors correspond to entities that were directly mapped to LOINC attributes because their tokens were mapped by ServoMap.

[b]Extended anchors correspond to direct anchors with additional anchors of TLAB entities related to the LOINC attributes of their ancestor(s).

rachidien"), the algorithm selected SYST1723-*liquide céphalorachidien* rather than SYST1533-*liquide vitrée* because the TLAB entity shared two tokens (ie, *liquide* and *céphalorachidien*) with the first LOINC system attribute, whereas it shared only one token with the second LOINC system (ie, *liquide*).

Anchoring to LOINC attributes

Table 2 describes the distribution of TLAB entities according to the LOINC attributes they have been anchored to. By propagating the LOINC attributes associated with each TLAB entity to all their corresponding descendants, almost all the 8285 TLAB entities have

Table 3. Distribution of anchored TLAB entities at the different steps (consecutive steps being listed from left to right) of the filtering process

		Initial	Filtering by class	Filtering by system	Filtering by method
Cardinality of anchors	1-0	5	6	52	409
	1-1	10	99	354	1011
	1-N	8285	8195	7894	6880

Note: The distribution is refined according to the cardinality of the related LOINC concepts (1-0 if no related LOINC concept; 1-1 if only one related LOINC concept; and 1-N if multiple related LOINC concepts).

been related to an analyte. The number of TLAB entities that have been related to a LOINC system, a LOINC method or a LOINC class was multiplied by 1.5 (from 3371 to 5065 entities), 1.6 (from 4949 to 7944 entities) and 3.2 (2262 to 7137 entities), respectively.

Anchoring to LOINC concepts

Table 3 describes the distribution of TLAB entities according to the LOINC concepts they have been anchored to, at each step of the filtering. Among the 8300 TLAB entities, 8295 (99.9%) were mapped to at least one LOINC concept. However, 8285 were related to multiple LOINC concepts, thus denoting a significant number of irrelevant mappings at the initial step. The filtering steps based on the LOINC classes, systems and methods reduced the number of LOINC concepts mapped to a TLAB entity, thus resulting in more 1-1 mappings and fewer 1-N mappings. At the end of the filtering process, 7891 (95.0%) TLAB entities were still related to at least one LOINC concept. The median cardinality of mappings was reduced from 324 to 14 LOINC concepts and the maximum cardinality from 24 017 to 5254 LOINC concepts through the filtering process.

The filtering of anchors and validation of mappings

Among the 8300 TLAB entities, 2144 (25.0%) were effectively used within the data warehouse. As stated before, these entities represented 279 065 808 laboratory results. Hence, the instance-based filtering process was performed for these 2144 TLAB entities. We were able to relate 1942 TLAB entities to 92 units of measure (corresponding to 279 065 424 laboratory results). Of these 92 units of measure, 57 have been mapped to UCUM codes and through these mappings, 1187 TLAB entities could be related to UCUM codes. The 57 UCUM codes corresponded to 24 UCUM properties that were manually mapped to 77 LOINC properties. The 1187 TLAB entities were mapped to 23 273 LOINC concepts before the instance-based validation process. By eliminating mappings with LOINC concepts that did not share the same LOINC property, the 1187 over 2144 (55.0%) TLAB entities have finally been mapped to 8455 LOINC concepts. The median cardinality of mappings for these TLAB entities was reduced from 20 to 5 LOINC concepts and the maximum cardinality from 5254 to 1227 LOINC concepts.

The 1187 TLAB entities covered 152 159 025 laboratory results (54.5%). The manual evaluation concerned 197 TLAB entities (being those having a 1-1 mapping to a LOINC concept), of which 92 were deemed equivalent and 25 corresponded to a subsumption relation between TLAB entities and LOINC concepts. That resulted in a precision of 0.59.

DISCUSSION

Findings

To align TLAB and LOINC, we used a more gradual approach than what is generally used in the literature, ie, a morphosyntactic similarity between the labels of concepts supplemented by a hierarchical similarity.[19] Indeed, our strategy consisted in using LOINC attributes to create definitional features for TLAB entities in order to support semantic alignment. Next, the LOINC attributes and their transversal relations with LOINC concepts were used as a support to query the appropriate LOINC concepts for the alignment (example provided in Supplementary Appendix 2). Finally, data from the Bordeaux university hospital were used to find additional knowledge for TLAB and thus improve the mapping results. The latter were acceptable with a precision of 0.59. As pointed out in ref.,[34] LOINC as a flat list may limit its usability as an interoperability tool. However, our results confirm that LOINC attributes can be more easily related to local terminology labels (eg, 99.8% and 61.0% of TLAB entities related to LOINC components and systems respectively). Then, we believe that LOINC attributes, rather than LOINC concepts as unique codes for laboratory results, can be used to more accurately anchor and query laboratory result data around the world. This will facilitate the combined use of local terminologies and LOINC benefiting from: (a) the unambiguity of LOINC for interoperability purpose, and (b) the usability of local terminologies.[1]

Hierarchical relations also played an important role through its combination with transversal relations. Indeed, propagating the related LOINC attributes of a TLAB entity to all its descendants gave the possibility to overcome the inconsistencies of some labels. To illustrate this last situation, syn-ana-cy301-*soit* ("*soit*" being the meaningless label) has been correctly anchored to 48432-9-*fructose (molar amount) in unspecified time semen*, thanks to its hierarchical relation with syn-ana-csfru-*FRUCTOSE SPERME*. Conversely, with the same inaccurate label, the other TLAB entity syn-ana-cy133-*soit* has been correctly anchored to 50193-2-*cholesterol in ldl.narrow density (mass/volume) in serum or plasma*, thanks to its hierarchical relation with syn-ana-cldl-*CHOLESTEROL LDL*. Thus, these mappings have been successfully established between TLAB entities and LOINC concepts although they did not share the same label or the same attributes (these TLAB entities cannot be related to LOINC attributes).

Finally, a sustainable finding was the benefit of using encoded data in the alignment process. Indeed, for a TLAB entity, we were able to validate only the mapped LOINC concepts that were instantiated by its related biological test results. Thus, these test results played the role of support knowledge to help validate the mappings obtained after the preliminary alignment. They provided information that was not accessible via TLAB labels. As an example, for the TLAB entity syn-ana-i261c-*c261-pholcodine*, the LOINC concept 73720-5-*pholcodine ige ab (units/volume) in serum* was selected as the appropriate anchor rather than 81971-4-*pholcodine IgE Ab RAST class (Presence) in Serum* because the results encoded with syn-ana-i261c were presented with the "***kUA/L***" unit of measure. Indeed, unlike *Presence*, *units/volume* and *kUA/L* could be linked to the same UCUM property "*Arbitrary concentration units*."

Alignment limitations and perspectives

Our process is based on the structure of the involved terminologies. However, some characteristics used in the description of LOINC labels may not be found in an interface terminology label. The main characteristics that may be identified in a TLAB label are the analyte, the system and sometimes the technique (Table 2). For this reason, only these attributes were used in the mapping process. In addition, the difference of granularity between TLAB entities and LOINC concepts induced multiple mappings for some TLAB entities. For example, syn-ana-i202f-*f202 noix cajou* was anchored to 6718-1-*cashew nut ige ab (units/volume) in serum* and 7183-7-*cashew nut igg ab (units/volume) in serum*. Using the original version of LOINC, previous work described the use of the LOINC group structure for seeking the parent concept of the anchored LOINC concepts.[13] However, as it is the case of our previous example, this parent does not always exist. Thus, a more formal structure for LOINC (like in the Web Ontology Language [OWL; https://www.w3.org/TR/owl-features/]) can help improve this strategy. Thus, in continuity with previous works,[35–37] an appropriate format, which integrates all linguistic variants and all parts and groups of LOINC as well as the hierarchical structure (pre-existing or automatically created), could allow to better disambiguate the multiple anchors by choosing those that involve the most general LOINC concept.

Finally, some authors used machine learning algorithms to deal with noisy labels for the annotation of laboratory results with LOINC concepts.[13,38] However, the labels of TLAB cannot be used to build a corpus for this purpose. A more controllable process by correcting TLAB labels and using the semantics of the LOINC structure was sufficient to obtain some good results. As a perspective, the step consisting in a "lexical mapping of tokens followed by the validation of mappings between the labels sharing the largest number of tokens in common" could be enhanced by machine learning algorithms.

CONCLUSION

In order to perform an alignment between TLAB and LOINC, our study used enhanced TLAB labels and a SKOS structure of LOINC. Based on the obtained structure, we anchored TLAB with LOINC with reasonable performances. However, our process presented some limitations. Perspectives for its improvement are the creation of a more formal structure of LOINC and the use of machine learning methods to improve the natural language processing of noisy labels.

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

J.N.N. developed the methodology, conducted the analyses, interpreted the results and wrote the first draft of the manuscript. R.G. and J.N.N. performed the evaluation of the obtained mappings. V.J. and F.M. supervised the work and were actively involved in defining the methodology, participating in the analyses and interpreting the results. All authors have reviewed, contributed to the writing and accepted the submitted paper.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST

The authors have no competing interests to declare.

DATA AVAILABILITY

The data and code underlying this article are available in GitHub at https://github.com/JNnikiema/LOINCTOTLAB.

REFERENCES

1. Rosenbloom ST, Miller RA, Johnson KB, *et al*. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *J Am Med Inform Assoc* 2006; 13 (3): 277–88.
2. Juvé-Udina ME. What patients' problems do nurses e-chart? Longitudinal study to evaluate the usability of an interface terminology. *Int J Nurs Stud* 2013; 50 (12): 1698–710.
3. Daniel C, Booker D, Beckwith B, *et al*. Standards and specifications in pathology: image management, report management and terminology. *Stud Health Technol Inform* 2012; 179: 105–22.
4. Griffon N. *Modélisation, création et évaluation de flux de terminologies et de terminologies d'interface: application à la production d'examens complémentaires de biologie et d'imagerie médicale*. 2013. https://tel.archives-ouvertes.fr/tel-00877697/file/these.pdf (accessed 3 March 2021).
5. Rosenbloom ST, Brown SH, Froehling D, *et al*. Using SNOMED CT to represent two interface terminologies. *J Am Med Inform Assoc* 2009; 16 (1): 81–8.
6. Wade G, Rosenbloom ST. Experiences mapping a legacy interface terminology to SNOMED CT. *BMC Med Inform Decis Mak* 2008; 8 (S1): S3.
7. Oluoch T, de Keizer N, Langat P, *et al*. A structured approach to recording AIDS-defining illnesses in Kenya: A SNOMED CT based solution. *J Biomed Inform* 2015; 56: 387–94.
8. Bakhshi-Raiez F, Ahmadian L, Cornet R, *et al*. Construction of an interface terminology on SNOMED CT: generic approach and its application in intensive care. *Methods Inf Med* 2010; 49 (4): 349–59.
9. Griffon N, Savoye-Collet C, Massari P, *et al*. An interface terminology for medical imaging ordering purposes. *AMIA Annu Symp Proc AMIA Proc* 2012; 2012: 1237–43.
10. Shvaiko P, Euzenat J. Ontology matching: state of the art and future challenges. *IEEE Trans Knowl Data Eng* 2013; 25 (1): 158–76.
11. McDonald CJ, Huff SM, Suico JG, *et al*. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem* 2003; 49 (4): 624–33.
12. Bodenreider O, Cornet R, Vreeman D. Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform* 2018; 27 (1): 129–39.
13. Parr SK, Shotwell MS, Jeffery AD, *et al*. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. *J Am Med Inform Assoc* 2018; 25 (10): 1292–300.
14. Georgy K. Mapping Russian laboratory terms to LOINC. *Stud Health Technol Inform* 2015; 210: 379–83. doi:10.3233/978-1-61499-512-8-379.
15. Lau LM, Johnson K, Monson K, *et al*. A method for the automated mapping of laboratory results to LOINC. *Proc AMIA Symp* 2000; 2000: 472–6.
16. Zunner C, Bürkle T, Prokosch H-U, *et al*. Mapping local laboratory interface terms to LOINC at a German university hospital using RELMA V.5: a semi-automated approach. *J Am Med Inform Assoc* 2013; 20 (2): 293–7.

17. Baker T, Bechhofer S, Isaac A, *et al*. Key choices in the design of Simple Knowledge Organization System (SKOS). *J Web Semant* 2013; 20: 35–49.
18. Forrey AW, McDonald CJ, DeMoor G, *et al*. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem* 1996; 42 (1): 81–90.
19. Nikiema JN, Jouhet V, Mougin F. Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. *J Biomed Inform* 2017; 74: 46–58.
20. Schulz S, Rodrigues J-M, Rector A, *et al*. Interface terminologies, reference terminologies and aggregation terminologies: a strategy for better integration. *Stud Health Technol Inform* 2017; 245: 940–4.
21. Wang Y, Patrick J, Miller G, *et al*. A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. *BMC Med Inform Decis Mak* 2008; 8: S5.
22. Baorto DM, Cimino JJ, Parvin CA, *et al*. Combining laboratory data sets from multiple institutions using the logical observation identifier names and codes (LOINC). *Int J Med Inf* 1998; 51 (1): 29–37.
23. Khan AN, Griffith SP, Moore C, *et al*. Standardizing Laboratory Data by Mapping to LOINC. *J Am Med Inform Assoc* 2006; 13 (3): 353–5.
24. Kim H, El-Kareh R, Goel A, *et al*. An approach to improve LOINC mapping through augmentation of local test names. *J Biomed Inform* 2012; 45 (4): 651–7.
25. Lee L-H, Groß A, Hartung M, *et al*. A multi-part matching strategy for mapping LOINC with laboratory terminologies. *J Am Med Inform Assoc* 2014; 21 (5): 792–800.
26. Kopanitsa G. Application of a Regenstrief RELMA V.6.6 to Map Russian Laboratory Terms to LOINC. *Methods Inf Med* 2016; 55 (2): 177–81.
27. Kukich K. Technique for automatically correcting words in text. *ACM Comput Surv* 1992; 24 (4): 377–439.
28. Sproat R, Black AW, Chen S, *et al*. Normalization of non-standard words. *Comput Speech Lang* 2001; 15 (3): 287–333.
29. Gadde P, Subramaniam LV, Faruquie TA. Adapting a WSJ trained part-of-speech tagger to noisy text: preliminary results. In: *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*. Beijing, China: ACM Press 2011. 5: 1–5: 8. doi:10.1145/2034617.2034623.
30. Nikiema JN, Mougin F, Jouhet V. Processus de prétraitement des libellés d'une terminologie d'interface. In: *4e édition du Symposium sur l'Ingénierie de l'Information Médicale*. Toulouse, France: 2017. 95–103. https://www.irit.fr/SIIM/2017/actes_siim2017.pdf.
31. Nikiema JN, Mougin F, Jouhet V. Building a Graph Representation of LOINC to Facilitate its Alignment to French Terminologies (in press). *AMIA Annu Symp Proc* 2020.
32. Diallo G. An effective method of large scale ontology matching. *J Biomed Sem* 2014; 5 (1): 44.
33. McCandless M, Hatcher E, Gospodnetić O, *et al*. *Lucene in Action*. 2nd ed. Stamford Conn: Manning Pub; 2010.
34. Carter AB, de Baca ME, Luu HS, *et al*. Use of LOINC for interoperability between organisations poses a risk to safety. *Lancet Digit Health* 2020; 2 (11): e569.
35. Bodenreider O. Issues in mapping LOINC laboratory tests to SNOMED CT. *AMIA Annu Symp Proc* 2008; 2008: 51–5.
36. Cora D, Petra D-H, Josef I. Aggregation and visualization of laboratory data by using ontological tools based on LOINC and SNOMED CT. *Stud Health Technol Inform* 2019; 264: 108–12. doi:10.3233/SHTI190193
37. Adamusiak T, Bodenreider O. Quality assurance in LOINC using Description Logic. *AMIA Annu Symp Proc* 2012; 2012: 1099–108.
38. Fidahussein M, Vreeman DJ. A corpus-based approach for automated LOINC mapping. *J Am Med Inform Assoc JAMIA* 2014; 21 (1): 64–72.