HAL open science

Ensuring Interoperability of Laboratory Tests and Results: A Linguistic Approach for Mapping French Laboratory Terminologies with LOINC

Namrata Patel, Yoann Abel, Fleur Brun, Guilhem Mayoral

▶ **To cite this version:**

Namrata Patel, Yoann Abel, Fleur Brun, Guilhem Mayoral. Ensuring Interoperability of Laboratory Tests and Results: A Linguistic Approach for Mapping French Laboratory Terminologies with LOINC. 1st International Conference on Computer Science Protecting Human Society Against Epidemics (ANTICOVID), Jun 2021, Virtual, Poland. pp.14-22, 10.1007/978-3-030-86582-5_2 . hal-03746672

HAL Id: hal-03746672

https://inria.hal.science/hal-03746672

Submitted on 5 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This document is the original author manuscript of a paper submitted to an IFIP conference proceedings or other IFIP publication by Springer Nature. As such, there may be some differences in the official published version of the paper. Such differences, if any, are usually due to reformatting during preparation for publication or minor corrections made by the author(s) during final proofreading of the publication manuscript.

Ensuring interoperability of laboratory tests and results: A linguistic approach for mapping French laboratory terminologies with LOINC

Namrata Patel[1,2][0000−0002−7923−4840], Yoann Abel[1], Fleur Brun[1], and Guilhem Mayoral[1]

[1] Onaos, Montpellier, France
[namrata.patel|yoann.abel|fleur.brun|guilhem.mayoral]@onaos.com
http://www.onaos.com
[2] Université Paul Valéry Montpellier 3, équipe AMIS
namrata.patel@univ-montp3.fr

Abstract. With the increasing use of electronic patient records, interoperability of patient-related data is of primary concern and international standards have been developed to assure compatibility between health data management systems. The present study is a first step in this direction for laboratory tests written in french, following the French government's impetus for providing national health records for patients. To address the linguistic complexities inherent to this task, we adopt a natural language processing (NLP) methodology. Our pilot case study shows that computational linguistic assistance in aligning terminologies makes it 3 times faster for domain experts. The significant difference in performance between the existing state of the art and our tool reflects the impact of addressing the linguistic challenges involved in the mapping process, especially in the multilingual context.

Keywords: LOINC · terminology mapping · NLP

1 Background and Significance

Standardised terminologies are increasingly used for structuring computerised health data today [11, 5, 1], facilitating the semantic interoperability between IT systems that process Electronic Health Records (EHR) [16]. LOINC (Logical Observation Identifiers Names and Codes), along with SNOMED and RxNORM, is one of the three most widely used standardised terminologies [10]. This terminology ensures the unambiguous identification of laboratory exams as well as clinical and biometric observations such as those described in prescriptions for these exams. In France, the digital health agency has entrusted the maintenance of this terminology to a consortium comprising the Assistance Publique Hôpitaux de Paris (AP-HP), the companies Vidal and Mondeca, and the Société Française d'Informatique de Laboratoire (SFIL). The "Jeu de valeurs LOINC" is published once every six months. It is aligned with the "Nomenclature des Actes de Biologie Médicale (NABM)" since 2014. Its use is required in particular for :

- Producing and using medical laboratory test reports organised in accordance with the specifications of the "Medical laboratory test report" section of the"Cadre d'Interopérabilité des Systèmes d'Information de Santé (CI-SIS),
- guaranteeing the compatibility of these results and observations, when they come from different medical laboratories.

Unfortunately, as of now, medical laboratories use their own local terminologies to organise the biological data they produce [12]. This situation therefore leads to a growing need for a direct mapping between local terminologies and the standardised LOINC terminology [14, 15]. Semi-automatic approaches and tools, including the one distributed by Regenstrief, are proposed to assist domain experts [9], but the mapping of local terminologies with LOINC remains time-consuming and requires a lot of human expertise[8]. With the exponential advances in automatic language processing driven by increasingly efficient artificial intelligence algorithms, our study aims at proposing an innovative solution to the problem of mapping these local terminologies with the LOINC encoding. To this end, our approach is designed by a multidisciplinary team of researchers in medicine, biology and artificial intelligence.

1.1 LOINC: A Closer Look

The LOINC terminology is a list of concepts referenced by a unique identifier, defining (1) analyses results for medical biology examinations, as well as (2) clinical and biometric observations included in the prescriptions for biology examinations. Each LOINC concept comprises a "long label" and a unique identification code. Each long label is composed of seven parts, (component, magnitude, etc.) separated into columns. LOINC is a pre-coordinated terminology that specifies only those combinations of atomic concepts that are allowed. For example, the six-part combination "Cortisol", "Moles/Volume", "Ponctuel", "Sérum/Plasma", "Numérique" and "Immunoanalyse" is relevant for the diagnosis of Cushing's syndrome and Addison's disease and has thus been integrated into LOINC with a code (83090-1). For improved coverage and completeness, it is possible to include additional concepts (post-coordination) into LOINC. In particular, users or developers can suggest new combinations of atomic concepts (proposed LOINC concepts) to the organisations in charge of LOINC maintenance.

1.2 Related work

Existing approaches for mapping local terminologies with LOINC rely on extensive pre-processing of local terminologies to facilitate automated matching processing. Common pre-processing steps include harmonization of local terminologies[3], adoption of synonyms [6], resolution of abbreviations associated with an augmentation of local terms with definitions and annotations. Existing tools for mapping pre-processed local terminologies with LOINC typically compute concept similarities using the fully specified names. Using the example given above

(LOINC code 83090-1), these names would be: “Cortisol”, “Moles/Volume”, “Ponctuel”, “Sérum/Plasma”, “Numérique” and “Immunoanalyse”. Many of these studies use the Regenstrief LOINC Mapping Assistant (RELMA), which applies such a full name alignment strategy, based on exact and restrictive name matching [17]. Few studies have applied a linguistic alignment where a match can also be identified even for similar but slightly different concept names. Let us note, in particular, that such a full-name alignment strategy is linguistically dependant on the language in which the LOINC is written (i.e. English) Our approach, based on existing artificial intelligence techniques [7] combines both: (1) extensive pre-processing of local terminologies, (2) linguistic mapping using semantic similarity and (3) a structural mapping strategy by decomposing the local terminology into atomic concepts. It differs from other approaches by its use of linguistic processing techniques from the field of artificial intelligence. In the following, we will present our study followed by comparative results with the existing RELMA tool, revealing thereby the importance of taking linguistic considerations into the overall mapping stragegy.

2 Objective

In the interest of enhancing interoperability between French medical laboratory terminologies, we develop a linguistic algorithm for mapping these terminologies with “Jeu de valeurs LOINC”, which is a French terminology based on LOINC (Logical Observation Identifiers Names and Codes). This paper presents a pilot case study with data obtained from the Biochemistry laboratory at the University Hospital of Montpellier, Hôpital Lapeyronie. The purpose of this study is to analyse the impact of using linguistic techniques to increase mapping quality and reduce cognitive and manual effort.

To address the linguistic complexities inherent to this task, we adopt a natural language processing (NLP) methodology. This allows us to (1) identify the linguistic factors that contribute to the mapping process, (2) develop a linguistically-enhanced, unified database of LOINC data combining French and international LOINC terminologies with additional UMLS concepts, (3) define a mapping strategy that exploits this database with state-of-the-art text-indexing techniques and (4) evaluate our algorithm by comparing it with the official mapping assistant RELMA in our pilot case study.

3 Materials and Methods

There are several challenges inherent in developing a linguistically-aware algorithm for mapping local terminologies with LOINC:

1. How does one approach the problem of multilingual mapping?
2. How does one handle cases where atomic concepts from LOINC are represented by their synonyms in the local terminologies?

3. How does one account for the structural discrepancies between the local terminologies and LOINC?

To address each of these challenges, we adopt a computational linguistic methodology which has been used in salient linguistic processing approaches in the field of artificial intelligence [REFERENCES]. In accordance with this methodology we proceed as follows:

Step 1: Linguistic and structural analysis of the input data
Step 2: Identification of linguistic markers that will facilitate mapping with LOINC
Step 3: Using these markers to define a semantically enriched LOINC knowledge base
Step 4 : Develop a mapping strategy based on structural and linguistic processing of local terminologies.

We evaluate our approach on data obtained from the biochemistry laboratory of the CHU of Montpellier.

The rest of this section is devoted to detailing each of these steps individually.

3.1 Input data: structural and linguistic analyses

Following the afore-mentioned automatic mapping methodology, we first proceed with a fine-grained linguistic and structural analysis of the input data, in order to identify the elements that will serve as key points in the development of our mapping algorithm. We now present some of the results of our analyses; those that would fail when applying a full-name mapping strategy:

- Local terminology contains atomic concepts that need to be broken down into two different LOINC concepts (e.g. "Albumine plasmatique" becomes LOINC "composant" *Albumine* + LOINC "milieu" *Serum/Plasma*).
- Local terminology describes units such as *g/mol* while LOINC describes these as "grandeur" such as *Masse/Volume*.
- Local terminologies describe atomic concepts under a synonym of those used by the LOINC.
- Local terminology describes ratios using the word 'Ratio' (e.g. Ratio Sodium Créatinine) while these are represented using the symbol '/' in LOINC (e.g. Sodium/Creatinine).

3.2 Linguistic markers: the contributing factor in the mapping process

The results of the linguistic and structural analyses allow us to define useful linguistic markers for the development of the final mapping algorithm. These markers are defined in the form of a list of linguistic rules that will be systematically applied to the whole dataset, in a pre-processing phase. For example, the adjective "fluoré (fluorinated)" in the local terminology is a linguistic marker

that maps the input data with the atomic concept "Serum/Plasma" in LOINC. The identification of these linguistic markers is an essential step in the automatic mapping of the data because they allow the expert to avoid a manual step of defining correspondences as requested by the RELMA tool. It remains to be evaluated to what extent these linguistic markers are applicable to data sets from other laboratories.

3.3 Enriched Linguistic Model for Mapping LOINC Data

The official LOINC terminology is organised in a multi-column table, accessible in an Excel-type spreadsheet. The order of the entries does not reflect any structural organisation of the concepts, making it impossible to browse the data quickly. In order to overcome this difficulty, we develop an "enriched" LOINC database for our mapping algorithm. This database is defined using advanced data management technology, allowing :

- Hierarchical structuring of data,
- powerful text indexing for rapid identification of any keyword in the database, despite structural complexity.

Thanks to the functionalities provided by this technology, we were able to define the LOINC atomic concept pertaining to a "component" in a more refined way:
Component = [name] + follwing properties :

1. Analyte,
2. primary class,
3. synonyms,
4. semantic type,
5. syntactic type (ratio, challenge, etc.).

3.4 Mapping Strategy

The final step in the development of our LOINC mapping algorithm is to combine the previous steps into a mapping strategy. The strategy optimises the use of (1) our prior analyses and (2) our semantically enriched LOINC knowledge base and is as follows:

1. Identification of relevant columns in the input data,
2. prepossessing the data with the expert rules identified during the linguistic and structural analysis,
3. mapping in stages (multi-part matching) by addressing atomic concepts individually,
4. return the top 20 LOINC candidates for each input entry

4 Evaluation

Our approach has been evaluated as a proof of concept in a pilot study at the Lapeyronie University Hospital of Montpellier. Our results are promising and reflect the necessity of using a linguistic approach for mapping.

However, we note that the results presented in this section do not suffice for evaluating our apporoach, as they do not represent a representative sample of data.

4.1 Experimental Setup

We evaluate our approach by comparing our performance with that of the official LOINC mapping assistant RELMA, based on a gold-standard provided by our medical experts.

The local terminology contains 792 distinct entries. The gold-standard mapping between this local terminology and LOINC was established manually by two independent medical experts, both familiar with the laboratory nomenclature. The accuracy and consistency of inter-annotator mapping was manually ensured in case of discrepancies: the two experts were invited to discuss jointly to decide on a consensus for the disputed cases.

4.2 Results

Of the 792 entries contained in the input dataset, 86 (10.85%) were unmappable because they were incomplete or incompatible with LOINC. 706 entries were therefore processed by our mapping algorithm. We present our results with respect to the gold standard in Table 1.

Table 1. Mapping algorithm: results.

Total distinct entries	792
Total incomplete/inconsistent entries	86 (10.9%)
Total mapped entries	706 (89.1%)
Correctly mapped entries	593 (84%)
Unmapped entries	113 (16%)

Among the correct alignments proposed by our solution, 271 were processed by the RELMA tool officially proposed with the LOINC. Our comparative results are summarised in Table 2. Of the top 20 codes provided by each tool, the correct alignment code was contained in 100% of our results and 50.8% of those of RELMA.

Semi-automatic mapping times were also measured, with all 792 entries processed in 4 hours and 40 minutes by a medical expert, i.e. 170 entries mapped per hour. Although few studies mention local terminology mapping times with LOINC, we can compare this time with that achieved during an evaluation of

Table 2. Mapping algorithm vs RELMA.

	Our algorithm	**RELMA**
Mapped entries	271	271
Correctly mapped	271 (100%)	138 (50.9%)
Incorrectly mapped	0	133 (49.1%)

RELMA by a German team, which found an average of 60 entries mapped per hour [17]. Our system is thus 2.8 times faster than RELMA.

4.3 Discussion

The heterogeneity of field data and the often long history with which laboratory databases have been designed and maintained, imposes an effort to pre-process the data even before applying algorithms to align the data with the LOINC repository. This pre-processing step can only be effective if the data owner has a good understanding of the LOINC nomenclature data structure and what types of relevant information are to be considered in their own data, to ensure the best performance of the automatic alignment. This work requires training time for the domain expert medical biologists involved[4]. The question of the interoperability of LOINC-coded data deserves to be asked concerning LOINC codes relating to measurements for which the technique is not specified, even though it constitutes a specificity to be taken into consideration for the correct interpretation of the associated biological result. One example is "creatinine", for which it is recommended in France to use an "enzymatic" (as opposed to "colorimetric") technique to estimate the glomerular filtration rate according to the CKD-EPI equation[13]. Thus two laboratories measuring creatinine in serum and expressed in moles/volume, but where one laboratory uses an enzymatic assay and the other a colorimetric assay, would share the same LOINC code, raising the question of the comparability of the result for the same patient visiting each of these two laboratories. Thus, other teams have suggested the need to associate additional information with the LOINC format in order to correctly interpret the associated biological result[2] This additional information relates to the characteristics of the analyser (technique, reagent) or to specific comments on the result that should be communicated in relation to the test results. All coding systems have weaknesses and, in particular, the "uncoded" system (currently used in France) of using highly ambiguous and interchangeable names for tests also has weaknesses. Thus, despite these limitations, we believe that the structuring of biological data according to the LOINC reference system has the merit of giving medical laboratories a structuring framework, which will enable them to make better secondary use of the data they produce. Based on our positive results, which are satisfactory for the expert biologists who use it, we propose to make sustained efforts to support medical laboratories in the correct use of this data format. These efforts must involve: experts in the field (hospital and private medical biologists), government agencies such as the

Agence du Numérique en Santé, and digital health manufacturers. The latter can facilitate the adoption of this format by proposing solutions for automatic transcoding of field data

5 Conclusion

This paper presents a pilot case study evaluated upon data obtained from the Biochemistry laboratory at the University Hospital of Montpellier, Hôpital Lapeyronie. We analyse the impact of using linguistic techniques to increase mapping quality and reduce cognitive and manual effort.

Our pilot case study shows that computational assistance in aligning terminologies makes it 3 times faster for domain experts. The significant difference in performance between RELMA and our tool reflects the impact of addressing the linguistic challenges involved in the mapping process. RELMA, developed for the English language, drops significantly in performance levels when compared with a mapping algorithm tailored to the French language.

References

1. Amith, M.F., He, Z., Bian, J., Lossio-Ventura, J.A., Tao, C.: Assessing the Practice of Biomedical Ontology Evaluation: Gaps and Opportunities. Journal of biomedical informatics **80**, 1–13 (Apr 2018). https://doi.org/10.1016/j.jbi.2018.02.010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5882531/
2. Drenkhahna, C.: The LOINC Content Model and Its Limitations of Usage in the Laboratory Domain. Digital Personalized Health and Medicine p. 6
3. Fidahussein, M., Vreeman, D.J.: A corpus-based approach for automated LOINC mapping. Journal of the American Medical Informatics Association : JAMIA **21**(1), 64–72 (Jan 2014). https://doi.org/10.1136/amiajnl-2012-001159, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3912728/
4. Fiebeck, J., Gietzelt, M., Ballout, S., Christmann, M., Fradziak, M., Laser, H., Ruppel, J., Schönfeld, N., Teppner, S., Gerbel, S.: Implementing LOINC – Current Status and Ongoing Work at a Medical University p. 7
5. Gkoutos, G.V., Schofield, P.N., Hoehndorf, R.: The anatomy of phenotype ontologies: principles, properties and applications. Briefings in Bioinformatics **19**(5), 1008–1021 (Apr 2017). https://doi.org/10.1093/bib/bbx035, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6169674/
6. Jung, B.K., Kim, J., Cho, C.H., Kim, J.Y., Nam, M.H., Shin, B.K., Rho, E.Y., Kim, S., Sung, H., Kim, S., Ki, C.S., Park, M.J., Lee, K.N., Yoon, S.Y.: Report on the Project for Establishment of the Standardized Korean Laboratory Terminology Database, 2015. Journal of Korean Medical Science **32**(4), 695–699 (Apr 2017). https://doi.org/10.3346/jkms.2017.32.4.695, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5334171/
7. Kaci, S., Patel, N., Prince, V.: From nl preference expressions to comparative preference statements: A preliminary study in eliciting preferences for customised decision support. In: 2014 IEEE 26th International Conference on Tools with Artificial Intelligence. pp. 591–598. IEEE (2014)

8. Kim, H., El-Kareh, R., Goel, A., Vineet, F., Chapman, W.W.: An Approach to Improve LOINC Mapping through Augmentation of Local Test Names. Journal of biomedical informatics **45**(4), 651–657 (Aug 2012). https://doi.org/10.1016/j.jbi.2011.12.004, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3340474/
9. Kopanitsa, G.: Application of a Regenstrief RELMA V.6.6 to Map Russian Laboratory Terms to LOINC. Methods of Information in Medicine **55**(02), 177–181 (2016). https://doi.org/10.3414/ME15-01-0068, http://www.thieme-connect.de/DOI/DOI?10.3414/ME15-01-0068
10. McDonald, C.J., Huff, S.M., Suico, J.G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., Williams, W., Case, J., Maloney, P.: LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clinical Chemistry **49**(4), 624–633 (Apr 2003). https://doi.org/10.1373/49.4.624
11. Poli, R., Healy, M., Kameas, A. (eds.): Theory and Applications of Ontology: Computer Applications. Springer Netherlands (2010). https://doi.org/10.1007/978-90-481-8847-5, https://www.springer.com/gp/book/9789048188468
12. Rosenbloom, S.T., Miller, R.A., Johnson, K.B., Elkin, P.L., Brown, S.H.: Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. Journal of the American Medical Informatics Association : JAMIA **13**(3), 277–288 (2006). https://doi.org/10.1197/jamia.M1957, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1513664/
13. de Santé, H.A.: Évaluation du débit de filtration glomérulaire et du dosage de la créatininémie dans le diagnostic de la maladie rénale chronique chez l'adulte. Bio tribune magazine **41**(1), 6–9 (Dec 2011). https://doi.org/10.1007/s11834-011-0067-3, http://link.springer.com/10.1007/s11834-011-0067-3
14. Uchegbu, C., Jing, X.: The potential adoption benefits and challenges of LOINC codes in a laboratory department: a case study. Health Information Science and Systems **5**(1) (Oct 2017). https://doi.org/10.1007/s13755-017-0027-8, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5636728/
15. Wade, G., Rosenbloom, S.T.: Experiences mapping a legacy interface terminology to SNOMED CT. BMC Medical Informatics and Decision Making **8**(Suppl 1), S3 (Oct 2008). https://doi.org/10.1186/1472-6947-8-S1-S3, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2582790/
16. Wang, K.C.: Standard Lexicons, Coding Systems and Ontologies for Interoperability and Semantic Computation in Imaging. Journal of Digital Imaging **31**(3), 353–360 (Jun 2018). https://doi.org/10.1007/s10278-018-0069-8, http://link.springer.com/10.1007/s10278-018-0069-8
17. Zunner, C., Bürkle, T., Prokosch, H.U., Ganslandt, T.: Mapping local laboratory interface terms to LOINC at a German university hospital using RELMA V.5: a semi-automated approach. Journal of the American Medical Informatics Association **20**(2), 293–297 (Mar 2013). https://doi.org/10.1136/amiajnl-2012-001063, https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2012-001063