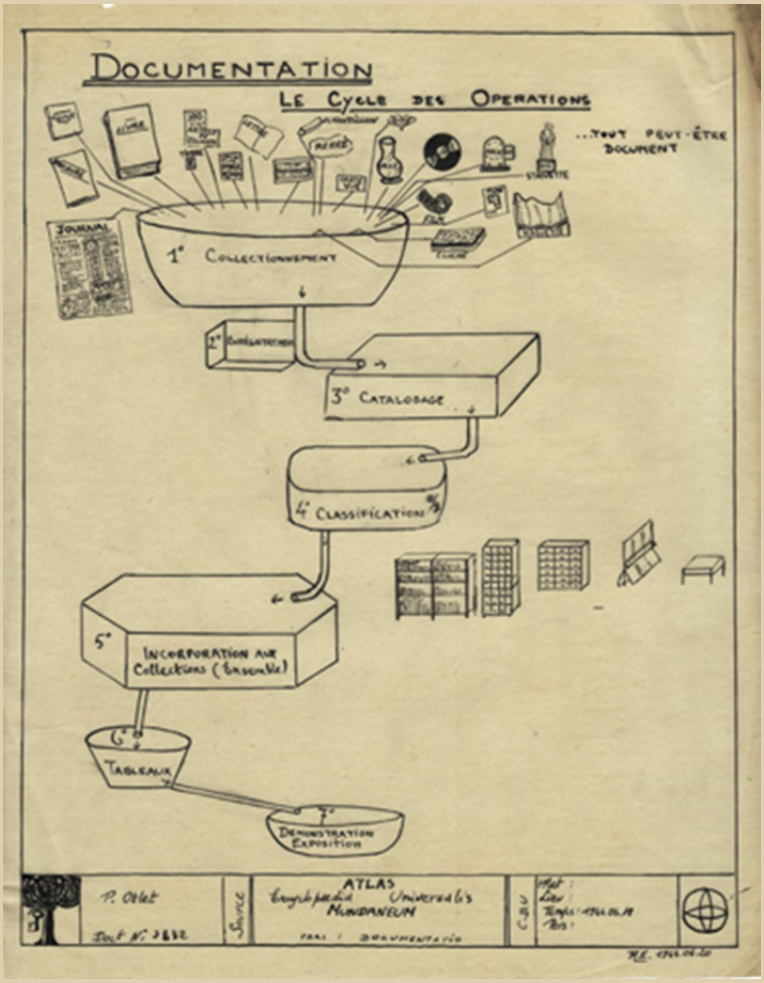
AIDAinformazioni

RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

NUMERO 3-4
ANNO 40
LUGLIO-DICEMBRE 2022



CACUCCI EDITORE
BARI

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

Fondata nel 1983 da Paolo Bisogno

Proprietario della rivista:
Università della Calabria

Direttore Scientifico:
Roberto Guarasci, *Università della Calabria*

Direttore Responsabile:
Fabrizia Flavia Sernia

Editrice: Cacucci Editore S.a.s.
Via D. Nicolai, 39 – 70122 Bari (BA)
www.cacucciedITore.it
e-mail: riviste@cacucciedITore.it
Telefono 080/5214220

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

«AIDAinformazioni» è una rivista scientifica che pubblica articoli inerenti le Scienze dell'Informazione, la Documentazione, la Gestione Documentale e l'Organizzazione della Conoscenza. È stata fondata nel 1983 quale rivista ufficiale dell'Associazione Italiana di Documentazione Avanzata e nel febbraio 2014 è stata acquisita dal Laboratorio di Documentazione dell'Università della Calabria. La rivista si propone di promuovere studi interdisciplinari oltre che la cooperazione e il dialogo tra profili professionali aventi competenze diverse, ma interdipendenti. I contributi possono riguardare topics quali Documentazione, Scienze dell'informazione e della comunicazione, Scienze del testo e del documento, Organizzazione e Gestione della conoscenza, Terminologia, Statistica testuale e Linguistica computazionale e possono illustrare studi sperimentali in domini specialistici, casi di studio, aspetti e risultati metodologici conseguiti in attività di ricerca applicata, presentazioni dello stato dell'arte, ecc.

«AIDAinformazioni» è riconosciuta dall'ANVUR come rivista di Classe A per l'Area 11 – Settore 11/A4 e censita per le Aree 10 – Scienze dell'antichità, filologico–letterarie e storico–artistiche; 11 – Scienze storiche, filosofiche, pedagogiche e psicologiche; 12 – Scienze giuridiche; 14 – Scienze politiche e sociali, così come dall'ARES (Agence d'évaluation de la recherche et de l'enseignement supérieur) che la annovera tra le riviste scientifiche dell'ambito delle Scienze dell'Informazione e della Comunicazione. La rivista è, inoltre, indicizzata in: ACNP – Catalogo Italiano dei Periodici; BASE –Bielefeld Academic Search Engine; ERIH PLUS – European Reference Index for the Humanities and Social Sciences – EZB – Elektronische Zeitschriftenbibliothek – Universitätsbibliothek Regensburg; Gateway Bayern; KVK – Karlsruhe Virtual Catalog; Letteratura Professionale Italiana – Associazione Italiana Biblioteche; The Library Catalog of Georgetown University; SBN – Italian union catalogue; Summon™ – by SerialsSolutions; Ulrich's; UniCat – Union Catalogue of Belgian Libraries; Union Catalog of Canada; LIBRIS – Union Catalogue of Swedish Libraries; Worldcat.

I contributi sono valutati seguendo il sistema del *double blind peer review*: gli articoli ricevuti dal comitato scientifico sono inviati in forma anonima a due referee, selezionati sulla base della loro comprovata esperienza nei topics specifici del contributo in valutazione.

AIDAinformazioni

Anno 40

N. 3-4 – luglio-dicembre 2022

CACUCCI EDITORE

BARI

Via Nicolai, 39 – 70122 Bari – Tel. 080/5214220
http://www.cacucciedi tore.it e-mail: info@cacucci.it

Sommario

Introduzione

Contributi

Note e rubriche

Contributi

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-220-1
DOI 10.57574/596522014
pag. 47-56 (luglio-dicembre 2022)

A proposal for a LOINC automatic mapping support tool

Maria Teresa Chiaravalloti*

Abstract: Clinical coding and classification systems are the *lingua franca* to overcome idiosyncrasies of the local information systems. Among those required into the Italian EHR, LOINC aims to uniquely identify both clinical document types indexed into EHR registries and tests into laboratory reports. Its use requires that local test codes are mapped to the codes of the standard so that equivalent concepts are aligned and can be easily understood and reused by other systems. This mapping process is cost and time-consuming as involves for many hours domain experts to find the right association. Therefore, this paper aims to make a proposal for an automatic mapping tool to support domain experts in finding the correct LOINC code to map to, specifically customized on the Italian language, combining textual similarity techniques with new NLP models and taking advantage of the so-called "wisdom of the crowd".

Keywords: LOINC, Mapping support tool, Text similarity, Natural Language Processing, Clinical coding systems.

1. Introduction and objective

Semantic interoperability of medical data used in clinical processes is necessary to exchange meaningful information among healthcare facilities. To pursue this aim, standardized coding and classification systems have to be used to uniquely identify clinical data. They are the *lingua franca* to overcome idiosyncrasies of the local informative systems.

In Italy, the newly published (Agid 2022) Guidelines on the Fascicolo Sanitario Elettronico (FSE is the acronym of the Italian national EHR) enforced the fundamental role played by interoperable data, as they are the key element to allow analysis through aggregation, information reuse, epidemiological surveillance and strategic planning. FSE infrastructure allows searching, retrieval and exchange of clinical documents among regional EHR systems, which are developed according to defined standards and protocols in order to be interoperable.

* Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Rende (CS), Italia. maria.chiaravalloti@iit.cnr.it.

Infrastructural and technical interoperability of the FSE has been widely tested (Chiaravalloti et al. 2015) and it is working as well as the use of a shared reference standard for the syntactic structuration of the clinical documents (namely the Clinical Document Architecture release 2 produced by the association Health Level 7). Therefore, the Guidelines identify semantics as the focus of the FSE 2.0.

The Prime Minister Decree n. 178/2015 is the first Italian regulation on FSE and, among the various aspects that are disciplined in it, there is the prescription for using four specific clinical coding systems: i) the International Classification of Diseases – Clinical Modification, 9th revision (ICD-9 CM) to identify diseases and other medical conditions; ii) the Logical Observation Identifiers Names and Codes (LOINC) to identify both clinical document types indexed into the FSE registries and tests into laboratory reports; iii) the Anatomical Therapeutic Chemical classification system (ATC) to encode the active ingredients of drugs; and iv) the Marketing Authorisation (namely Autorizzazione all'Immissione in Commercio or AIC), which is the Italian classification to specifically identify drugs' package type. Among them, LOINC is the newest standard in the Italian healthcare context. It was created in 1994 at Regenstrief Institute of Indianapolis and it counts more than 99,000 codes in its last release (2.73 in August 2022) and a community of hundreds of thousands of users in the world. Fostering its wider adoption is fundamental because i) laboratory reports are the clinical document type most indexed in the FSE registries; ii) the high level of detail of the laboratory medicine and the granularity of LOINC make easier the unique identification of the single items; iii) the pandemic made clear the need to have comparable and aggregable data, especially for measurable observations that could reveal critical situations for public health.

LOINC use requires that local codes are mapped to the codes of the standard so that equivalent concepts are aligned and can be easily understood and reused by other systems. This mapping process is costly and time-consuming as involves for many hours domain experts to find the right association. This is especially true for laboratories, which usually have large size local catalogues of tests to map. Inheriting mappings from other laboratories is always risky because tests that laboratory A and laboratory B call the same may be not equivalent in the substance and vice versa.

Frassica (2003) stated that a small subset of LOINC codes covers over the 90% of the routinely tests performed by a laboratory. This means that there is a great amount of observations commonly performed by laboratories that could be with a high probability mapped to the same LOINC code. Moreover, our experience as LOINC Italia working group[1] showed that two laboratories

[1] The LOINC Italia working group was founded and is operating into the Institute of Informatics and Telematics of the Italian National Research Council (IIT-CNR) and it is officially

perform over the 50% of identical tests of a given specialty. Considering this, mapping could be a non-onerous process if performed taking into account other successful experiences. LOINC implementation in Italy is currently performed manually, because the standard has been applied to limited laboratory specialties, so the required effort was not computationally high. Nonetheless, in the near future there will be the pressing need to have an automatic support for wider mapping operations of local catalogues to LOINC. Therefore, this paper aims to make a proposal for an automatic mapping tool to support domain experts in finding the correct LOINC code to map to.

2. Background

Many mapping support tools and methodologies had been developed overtime trying to make easy and speed up this task and proving the great attention focused on the mapping phase, as it is fundamental to foster the standard adoption. Among them, it is worth to firstly mention the software released by Regenstrief Institute, which is the LOINC Standard Development Organization. RELMA (REgenstrief LOINC Mapping Assistant) is a free program which offers different functionalities to help users in finding the right LOINC code, such as a test frequency rank (Regenstrief Institute, n.d.), the Intelligent Mapper (Vreeman and McDonald 2005) and the Community Mappings repository (Dixon, Hook and Vreeman 2015). The goodness of the Intelligent Mapper was then tested in a comparative study (Vreeman and McDonald 2006) with a vector space model-based program for mapping local diagnostic radiology terms to LOINC, and in a German study (Zunner et al. 2013) aimed at proving its validity also on non-English languages. Zollo and Huff (2000) proposed a mapping approach that aims to map local catalogues to LOINC by cross-referencing the matching codes. They demonstrated that if two local tests match in all the information they carried, and one of them is correctly mapped to LOINC, then this mapping could be inherited by the others. Fidahussein and Vreeman (2014) approach is based on supervised machine learning and information retrieval using Apache's Maxent and Lucene to show that the collective knowledge contained in a complete dataset of local terms mapped to LOINC can be used to support the mapping new local terms. Lau et al. (2000) presented a methodology study for automatically mapping local terms to LOINC by using parsing, logic rules, synonyms, attribute relationships and the frequency of mapping to a specific LOINC code. Parr et al. (2018) developed an automated machine learning pipeline that leverages noisy labels to map laboratory data to LOINC codes. They obtained highly

recognized by the Regenstrief Institute as Italian partner for LOINC translation and related activities through a Memorandum of Understanding signed into 2014.

accurate mappings for 85% of the tests in the sample. None of the abovementioned solutions has been tested on the Italian language yet. Previously, we worked on a similar topic by developing a NooJ application[2] to speed up the validation of the already performed mappings (Parisi and Chiaravalloti 2018).

This rich literature proved that the research topic is not new, but it is still promising as it is based on two components that are alive and continuously evolving: language and technology. None of the automatic or semi-automatic cited tools dispensed with manual mapping work, even only for double-checking, and so it still remains the gold standard for linking local codes to LOINC codes. Moreover, the major part of these tools is tested on the English language, which has not such peculiarities as other languages like the Italian for example.

3. A proposal for a LOINC mapping support tool

This study arose from the verified need to have a support in the operations of mapping local laboratory codes to LOINC codes. As LOINC Italia working group, we frequently receive requests for LOINC education and training sessions, assistance in mapping efforts, and validation of already performed mappings. Both physicians and decision-makers always ask for an easy way to approach the standard and to speed up its adoption. Nevertheless, laboratory tests have so granular differences that it is not possible to adopt a top down approach providing a big catalog already mapped to LOINC which is surely valid for all the Italian laboratories. As Regenstrief Institute suggests and we experienced in our activities as LOINC Italia working group, a bottom up approached mapping is successful because it takes advantage from the knowledge of the domain of the laboratory physicians as mappers. Another possible approach is to make first the effort to unify local catalogues, such those of the laboratories of the same Region, and then mapping to LOINC. Whichever approach is chosen, the same need remains: to reduce the time required by this activity and at the same time to prevent potential errors in mapping the same semantic content by different domain experts.

Cited prior studies presented automatic or semi-automatic mapping support tools. They are based on different techniques and methodologies, but all of them require expert review to double check computer-generated results. The proposal presented in this paper aims to invert this trend by realizing a

[2] NooJ is a linguistic development environment software as well as a corpus processor constructed by Max Silberztein. NooJ allows linguists to construct the four classes of the Chomsky-Schützenberger hierarchy of generative grammars: Finite-State Grammars, Context-Free Grammars, Context-Sensitive Grammars as well as Unrestricted Grammars, using either a text editor (e.g. to write down regular expressions), or a Graph editor (Wikipedia 2022).

solution which requires as little human intervention as possible, being able to count on a high degree of reliability of the automatically created mappings. In our view, these associations between source language, represented by the local catalogues' tests, and target language, which is the standard LOINC, can gain consistency from different elements: i) the use of text similarity algorithms; ii) the iterative learning from already validated mappings; and iii) the experimentation of machine learning models such as the transformers.

Text similarity is about to calculate how two phrases are close to each other. It could be lexical, if based on the syntactic analysis of the words, or semantic, if it considers also the meaning of the words and their interrelations. In a previous work (Parisi and Chiaravalloti 2018), we evaluated mapping correctness by comparing descriptive elements of both local laboratory tests and LOINC terms. We constructed dictionaries to describe the morphological properties of local linguistic forms and their related flexional grammars, considering each local test like a sentence constituted by fixed elements in fixed positions. Text similarity applied to laboratory observations could, in fact, benefit from the absence of textual context and the presence of recurrent structure in tests descriptions. So, formal grammars recognition rules were applied exactly to the correct category (e.g. the analyte, the type of sample, etc.), being sure to find it in a specific position. The added value given by the software NooJ (Silberztein 2015) was the possibility to associate local linguistic forms and standardized tags from LOINC, by using the semantic properties associated to the lexical units. This allowed also the domain experts' knowledge formalization, which is the hardest task when working with natural language. Starting from the results of this study, the proposed methodology could be applied to the development of a mapping support tool.

The possibility to count on the "wisdom of the crowd", intended as looking at mapping choices already made by other LOINC users in the world, is one of the functionalities offered by RELMA (Dixon, Hook and Vreeman 2015). Embedding this knowledge base into a mapping support tool, with a weighting system to balance any errors or discrepancies and combined with text similarity scores, would allow domain experts to be confident that automatically proposed mappings are correct, so they can be skipped without reviewing them. The more mappings of local tests named alike, or by synonyms, to the same LOINC code, the more this helps to increase the consistency of that association. Differently from the solution adopted by Regenstrief, the proposed mapping support tool should consider only mappings validated by experts[3] so that the reliability of the "wisdom of the crowd" is higher. Dictio-

[3] Among the activities of the LOINC Italia working group, there is also the validation of the local tests mapped to LOINC. Mappings are double checked one by one by LOINC Italia experts, who make code suggestions, ask for more information or manage new LOINC codes submissions, if needed.

naries and grammars should be applied also in this phase to manage the strong idiosyncrasy that local test names could have by reconducting them to the same LOINC concept even if their graphic designations are different. Using Knowledge Organization Systems (KOS) specific of the reference domain can be helpful in managing the strong linguistic variability of the Italian language (rich in synonyms and acronyms, especially in the medical domain). Analyzed solutions, in fact, were focalized more on the informatics aspects of the process rather than on the linguistics one. For this reason, we believe that using a linguistic development environment software (e.g. NooJ) could make the difference and reveal as an innovative aspect. Figure 1 shows an overview of these components of the proposed tool.

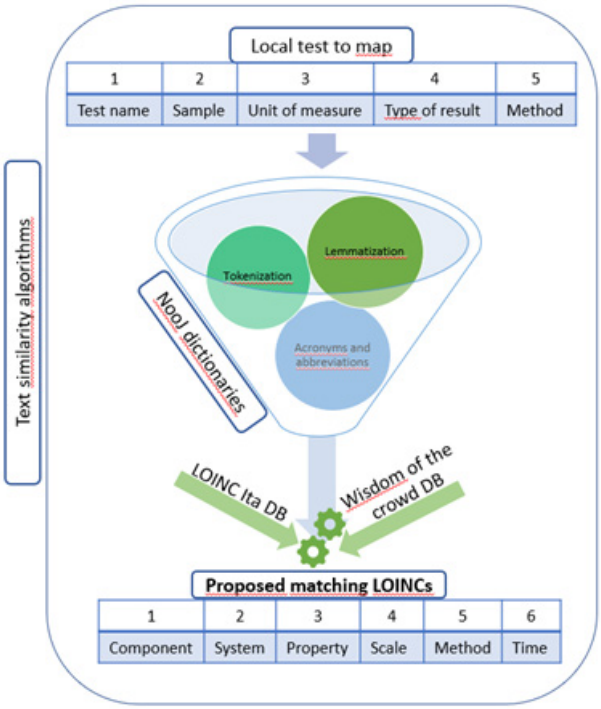


Figure 1. Text similarity algorithms of the proposed LOINC mapping support tool.

The new element of this proposal, if compared to those previously mentioned, is the experimentation of the use of transformers. They are a deep-learning model introduced in 2017 by a Google Brain research team[4] and are now a leading model for Natural Language Processing (NLP) problems. Transformers consider each part of the input data as a space vector and differen-

[4] Google Brain is a deep learning artificial intelligence research team under the umbrella of Google AI, a research division at Google dedicated to artificial intelligence.

tially weight their significance in accordance with the context. They are able to predict the occurrence of a certain word in a certain position of the input data and, considering the context, to determine the semantic similarity of two phrases. This methodology is successfully applied to translation and text summarization and so we believe it could bring good results with laboratory tests, which are highly structured reported with the same type of information always in the same position of the phrase. Figure 2 shows an overview of the expected functionalities of the application of NLP techniques like transformers into the proposed tool.

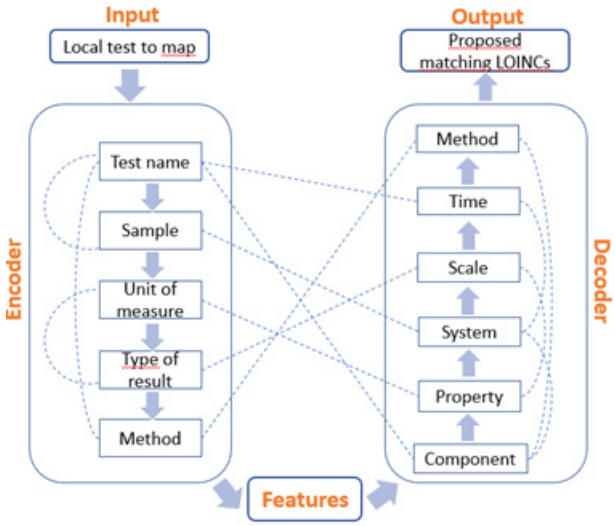


Figure 2. Application of NLP techniques into the LOINC mapping support tool.

The overall objective of this work is to make a proposal of an automatic mapping tool to support domain experts in finding the correct LOINC code to map their local tests. Considering previous works in the same field and the described innovative elements that would make this solution different from the others and more customized on the Italian language, this proposal would be developed according to the following steps: i) creation of a comparison matrix to identify and analyze strengths and weaknesses of the existing similar solutions; ii) evaluation of the usability of the methodology developed for our previous work (Parisi and Chiaravalloti 2018) on mappings correctness evaluation through NooJ software; iii) design of the tool according to the above described features; iv) development and testing of the tool.

4. Discussion

A LOINC mapping support tool, like the one proposed in this paper, is expected to be greatly required in the near future. In fact, the strong input coming from the Italian government to increase the use of coding systems within the FSE undeniably raises the question of how to interface these standards with the local systems currently in use. Supporting the transcoding processes will speed up the process, as it will make the approach to these coding systems less difficult. In this scenario, developing a mapping support tool specifically customized on the Italian language would have a strong impact on the national scene because there are few solutions of this type applied to this specific domain and because it aims to combine textual similarity techniques with new NLP models, taking advantage of the so-called "wisdom of the crowd". Successful implementations of the LOINC standard followed an incremental approach of the mapping effort, starting from easier laboratory specialties in order to give the chance to familiarize with the coding system. The same could be done to train the system on sectoral specificities.

Functionalities that can constitute added values of the proposed tool are: i) the automatic mapping of all the associations with a top-level similarity score between source and target terminologies, so allowing time saving and error reductions by skipping them; ii) the chance to get knowledge from validated mappings already performed by the others and iii) the application of the newest NLP techniques to an highly structured language as that of laboratory observations is. Moreover, the tool would allow the creation of a repository of tests mapped to LOINC, which could be used for secondary research and government purposes. For example, if all the laboratories of the same Regions share their mappings to LOINC, then by means of the same code of the standard a regional catalog of mappings can be created and local tests mapped to the same LOINC code can be compared for economic, clinical and government purposes.

The strong push towards digitalization supported by the European digital agenda and by funding from the Italian Piano Nazionale di Ripresa e Resilienza (PNRR) emphasizes the importance of semantic interoperability, without which technologies would not be able to fully exploit their potential. The use of standard coding systems in the clinical field is not yet uniform, despite the crucial importance of conveying detailed information in this sector. Therefore, the creation of automatic support tools that can facilitate the use of these solutions in the daily practice of clinical stakeholders, also facilitating their work, is felt as an urgent need. LOINC plays a leading role as a standard for the unambiguous identification of laboratory tests, however, due to the high level of granularity of its organization, its use is not immediate like that of other medical standards which can rely on descriptions closer to natural lan-

guage. Creating support tools for mapping could therefore foster its correct use. They should be based on both the management of the linguistic variability and the exploitation of the knowledge created by the mapping experience of the others. This proposal discussed expected features of a LOINC mapping support tool based on the knowledge of the standard and the experience of the domain needs in Italy gained from the activities of the LOINC Italia working group. Next steps will be putting around the same table all the competencies needed to support the design and development of the proposed tool.

Bibliographic References

Agenzia per l'Italia Digitale. 2022. "Fascicolo Sanitario Elettronico 2.0: pubblicate le Linee Guida per l'attuazione." https://www.agid.gov.it/it/agenzia/stampa-e-comunicazione/notizie/2022/07/19/fascicolo-sanitario-elettronico-20-pubblicate-linee-guida-lattuazione.

Chiaravalloti, Maria Teresa, Mario Ciampi, Erika Pasceri, Mario Sicuranza, Giuseppe De Pietro, and Roberto Guarasci. 2015. "A model for realizing interoperable EHR systems in Italy." In *15th International HL7 Interoperability Conference (IHIC 2015) Proceedings*, 13–22, http://ihic2015.hl7cr.eu/Proceedings-web.pdf.

Dixon, Brian E., John Hook, and Daniel J. Vreeman. 2015. "Learning from the crowd in terminology mapping: the LOINC experience." *Lab Medicine* 46 (2): 168–74.

Fidahussein, Mustafa, and Daniel J. Vreeman. 2014. "A corpus-based approach for automated LOINC mapping." *Journal of the American Medical Informatics Association* 21 (1): 64–72.

Frassica, Joseph J. 2003. "Frequency of laboratory test utilization in the intensive care unit and its implications for large scale data collection efforts." In *AMIA Annual Symposium Proceedings*, 844.

Lau, Lee Min, Kate Johnson, Kent Monson, Siew H. Lam, and Stanley M. Huff. 2000. "A method for the automated mapping of laboratory results to LOINC." In *AMIA Annual Symposium Proceedings*, 472–76.

Parisi, Francesca, and Maria Teresa Chiaravalloti. 2018. "A set of NooJ Grammars to verify laboratory data correctness." In *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications*, edited by Ignazio Mauro Mirto, Mario Monteleone, Max Silberztein, 155–66. Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-10868-7_14.

Parr, Sharidan K., Matthew S. Shotwell, Alvin D. Jeffery, Thomas A. Lasko, and Michael E. Matheny. 2018. "Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database." *Journal of the American Medical Informatics Association* 25 (10): 1292–300. DOI: 10.1093/jamia/ocy110.

Regenstrief Institute, Inc. n.d. "LOINC Code Rankings." Accessed November 25, 2022. http://loinc.org/usage/obs.

Silberztein, Max. 2015. *La formalisation des langues, l'approche de NooJ*. London : ISTE.

Vreeman, Daniel J., and Clement J. McDonald. 2005. "Automated mapping of local radiology terms to LOINC." In *AMIA Annual Symposium Proceedings*, 769–73.

Vreeman, Daniel J., and Clement J. McDonald. 2006. "A comparison of Intelligent Mapper and document similarity scores for mapping local radiology terms to LOINC." In *AMIA Annual Symposium Proceedings*, 809–13.

Wikipedia. 2022. "NooJ." Last modified January 6, 2022. https://en.wikipedia.org/wiki/NooJ.

Zollo, Kenneth A., and Stanley M. Huff. 2000. "Automated mapping of observation codes using extensional definitions." *Journal of the American Medical Informatics Association* 7 (6): 586–92.

Zunner, Christian, Thomas Bürkle, Hans-Ulrich Prokosch, and Thomas Ganslandt. 2013. "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): 293–97.

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

Anno 40

N. 1-2 – gennaio-giugno 2022

mundaneum

In copertina
Disegno di Paul Otlet, Collections Mundaneum, centre d'Archives, Mons (Belgique).

ISBN 979-12-5965-220-1

9 791259 652201

ISSN 1121-0095

9 770112 100950