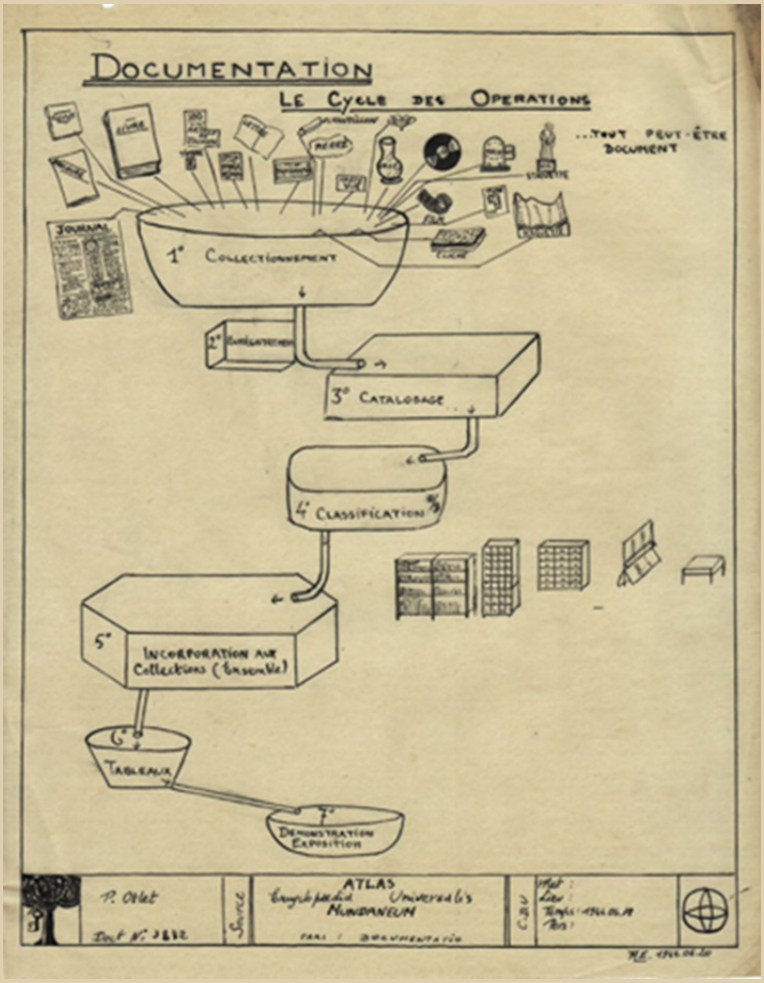
AIDAinformazioni

RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

NUMERO 3-4
ANNO 39
LUGLIO-DICEMBRE 2021



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 39

N. 3-4 – luglio-dicembre 2021

CACUCCI EDITORE

BARI

© 2021 Cacucci Editore – Bari
Via Nicolai, 39 – 70122 Bari – Tel. 080/5214220
http://www.cacucciedit ore.it e-mail: info@cacucci.it

Sommario

Contributi

Contributi in memoria di Maria Pia Carosella

Note e rubriche

Contributi

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-090-0
DOI 10.57574/596509002
pag. 31-44 (luglio-dicembre 2021)

Exploring clinical documents through advanced semantic analysis techniques

Antonietta Folino, Claudia Lanza, Erika Pasceri, Anna Perri*

Abstract: This paper aims at providing a formal structuring of paper-based clinical records through semantic annotation procedures. In particular, the approach adopted in this explorative study has been based on document management methodologies aiming at exploiting the texts from clinical documents through advanced document analysis, information extraction and modeling techniques, supported by language processing methods, which contribute to creating a comprehensive knowledge repository. The result will be a systematic organization of the most common categories proper to clinical record texts in the Italian language that could be used as a replication model for several health system spheres, with the perspective of providing a conceptual representation of the patients' clinical history over time.

Keywords: Clinical record, Digitization, Semantic annotation, Paper-based clinical document, Knowledge representation.

1. Introduction

The massive digitization of processes and sources could have a positive impact in those actions dealing with knowledge management. The concept of "digitization" itself could vary depending on the use of digitization techniques and also on its meaning which has evolved over time. As stated by Owen:

> An important difference between the early stages of computing and present-day information technology is to be found in something that was lacking in the initial stage, and has now become ubiquitous: textual and graphical information resources that convey meaning to human beings, playing a major role in

* Dipartimento di Culture, Educazione e Società, Università della Calabria, Rende (CS), Italy. antonietta.folino@unical.it; claudia.lanza@unical.it; erika.pasceri@unical.it; anna.perri@unical.it

Authors have equally contributed to this work, however Antonietta Folino particularly focused on 'Semantic analysis', Claudia Lanza on 'Semantic annotation' and 'Conclusion and future perspectives' sections, Erika Pasceri on 'Introduction' and 'State of the art' sections and Anna Perri reviewed the entire work.

> their exchange of knowledge, insights, opinions and ideas. The application of information technology is now predominantly aimed at creating, distributing, storing and accessing products of the mind, and at communication rather than at processing (Owen 2006, 2).

In addition, the concept of "digitization" is commonly used in librarian contexts with the goal of sharing knowledge with users and research communities, but very often the digital resources created refer to a format modification from analogue to digital in order to preserve a huge collection of books and to share them through the web. As we learnt from modern society, "digitization" is more than this, because it is a process that deeply modifies the concept that stands behind the actions. «Information is created directly in digital form, and in this sense the process of digitization is a behavioral transformation: the digital mode has become the common mode of expression, documentation and communication» (Owen 2006, 3). Those resources that are born analogue (paper) must (or should) be treated as such, within limits of course, strictly depending on their nature. Similarly, a resource that is born in digital format has all the potential – depending on its intrinsic characteristics – in presenting multiple uses. Modern techniques help to find a bridge with those different sources to effectively employ them in terms of reusability, preservation and knowledge extraction in an interoperable way, even when a format variation occurs (from analog to digital).

The research activity presented in this paper deals with the creation of annotated clinical record documentation starting from paper-based clinical documents. The work has been set in the Italian language and it can be replicated for further studies within this knowledge domain (Attardi, Cozza, and Sartiano 2015). Clinical documents have their own intrinsic specificity, indeed the relevant entities that could be identified often refer to *symptoms, drugs therapies, instrumental tests, patients' autonomy levels*. The identification of main (and recurring) entities is the starting point to proceed to text annotation and categorization tasks. The semantic categories that have been identified in this model in order to extract specific knowledge about the patients' clinical history have proved to be a significant baseline to establish a fixed and determined disposition of information. In this way, the paper underlines the benefits derived from a well-formed annotation corpus to treat clinical record documentation and run out semantic analysis. These latter can lead to distribution analysis of the health care information repeated in the clinical records and becoming, eventually, a future classification tool to automatize the inclusion of medical entities.

The semantic querying operations to annotate clinical electronic records can be applied by exploiting the features of pre-trained tools that work as extractors of determined categories set out by the users. This is the case of semi-automatic annotator tools like WebAnno (Yimam et al. 2013), or CAT-

MA (Meister et al. 2017) – which has been used for this research activity – that are trained to retrieve all the occurrences of specified categories pre-defined by the users in the texts given in input. From a methodological point of view, the added value of this study comes from the combination of different knowledge organization techniques and approaches, the digitization of native paper-handwritten clinical records and the production of semantic annotation models over their structures. The structure with which the clinical records are drafted is prefixed and this facilitated the semantic recognition process of different sections of the text associated with specific medical information to be tagged. From a scientific point of view, the contribution of this research project relies on the importance of extracting knowledge from those documents that contain a large amount of information that often remains "hidden", used only in clinical practice during patient care process and later archived, but which has the high potential to be reused and included in good practices to be shared with the reference community for subsequent usages. The model, projected to organize textual elements in a formalized annotated structure, could also be used for different medical purposes, when the lack of conceptual approaches and specific methodologies in managing data could cause loss of information. The main idea is, indeed, to add that fundamental information that comes from the daily clinical practice to those resources that are published everyday as scientific papers on main online databases.

The first part of this paper covers the digitization process of clinical documents, starting from the description of the corpus compilation and of the handwritten-text recognition procedures. In the second section, the semantic annotation of the digitized clinical records is presented, specifically the anamnesis contents, to identify and classify the more relevant information and make the retrieval processes of documents easily executable and, thus, creating a discoverable and linked knowledge base on which to run several queries in the clinical history of the patients.

2. State of the art

Databases in the biomedical field available online represent a key tool for both scientific research and for supporting the clinicians in customized clinical patient management. In this particular domain some of them are general, others relate to specific sectors, some index primary information sources, other systematic reviews. The available databases include primary and secondary databases where the first ones draw their sources from primary medical literature, i.e., articles published in leading biomedical journals. Among these latter, there are:

- MEDLINE, the most important open-source electronic database produced by the National Library of Medicine, available through PubMed

portal, an online platform for almost all the international biomedical literature;
- EMBASE (The Excerpta Medica Database), a non-open-source bibliographic database, produced by Elsevier Science, specializing in medical literature, mainly European, and specifically refers to pharmacology and toxicology.
- The CINAHL Database (Cumulative Index to Nursing and Allied Health Literature), non-open-source database dedicated to nursing sciences. It also includes references on biomedicine and behavioral sciences.
- The Cochrane Library, a monthly updated electronic publication which collects works from the Cochrane Collaboration. It is composed of different databases including systematic reviews, registers of randomized clinical trials referring to drug treatments, diagnosis and screening, health promotion and organization of service. It represents a useful tool for the clinical choices, continuous education and services organization.

Among the secondary sources, there are documentations that are generally collected under the term *gray literature*, meaning unconventional documents not disseminated through normal commercial publication channels and, thus, difficult to identify and access. Dissertations, reports, conference proceedings, unpublished results, informal communications and so on, are part of the gray literature. This scenario demonstrates how the main repositories, which allow the sharing and exchange of specialized knowledge in the biomedical domain, consist of documents which do not include primary clinical documents. Therefore, in this sense, our project idea could represent a novelty with a high potential interest for clinicians and for researchers in the biomedical field.

The documents constituting the source corpus of clinical digitized records have undergone an annotation task. Semantic annotation represents the process of attaching to a text document, or other unstructured content, metadata about concepts (e.g., people, places, organizations, products or topics) which result relevant. Text Analysis and Knowledge Extraction System (cTAKES) (Savova 2010) and MetaMap (Stewart, von Maltzahn, and Raza Abidi 2012) are two well-known systems to semi-automatically annotate the biomedical corpora of clinical record documents. cTAKES is based, as is CATMA, on Unstructured Information Management Architecture (UIMA) and OpenNLP frameworks, while MetaMap works in a way that it maps the elements included in the medical source documents to the corresponding concepts within formal medical lexicons, such as the Unified Medical Language System (UMLS) Metathesaurus. Searle (2019) describes the interface MedCATTrainer to annotate specific biomedical texts by giving in input a model of Named Entity

Recognition and Linking (NER+L) that collects the salient parts of texts in compliance with specialist domain knowledge requirements. In the Italian framework, the work carried out by (Attardi, Cozza, and Sartiano 2015) highlights methods to retrieve technical information from texts through regular expression matching structures to map relevant elements in clinical records, such as, symptoms, treatments or drugs, creating the basis for the patients' medical status descriptions in the semantic post-processing tasks. The patterns batteries considered by the authors include the possible variants of the expressions usually employed by physicians. These authors also provide an in-depth analysis of the negative-phrase patterns constructions to detect the corresponding positive statements standing out as pathology or disorder triggers in patients. In the French context, Campillos (2017) proposed a semantic annotation procedure for clinical corpus-documentation with the aid of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT) addressing several entities, attributes and relations. Oronoz et al. (2013) also studied the advantages offered by the automatic semantic annotation process of clinical records through the employment of existing ontologies like SNOMED CT: in this way, the system the authors explored automatically runs to retrieve the medical entities, such as disease, drug and substance names in given input documents. The task of detecting important medical information through the use of Natural Language Processing (NLP) techniques, specifically concerning the automatic annotation methodology, is addressed by (Tou et al. 2017) who proposed a detection of the infections starting from predictive categories, i.e., *personal information, admission note, vital signs, diagnose test results and medical image diagnose,* or by (Hassanzadeh, Nguyen, and Koopman 2016) in retrieving from electronic health records or free-text. The authors used four annotation systems providing a more formal lexicon representation to be used as a comparative measure with the annotation structures, i.e., MetaMap, NCBO annotator, Ontoserver, and QuickUMLS, aiming at collecting the main concepts starting from clinical free-text documents.

For this study, a similar semantic annotation method proposed by (Attardi, Cozza, and Sartiano 2015), (Tou et al. 2017) and (Hassanzadeh, Nguyen, and Koopman 2016) to retrieve the information from digitized clinical records has been pursued. The novelty brought with this research activity is based on the source documents, i.e., the digitization process of Italian clinical records that have been then treated to extract medical-oriented information encapsulated in the predetermined domain-specific semantic categories.

3. Semantic analysis

In this section the documentation analysis will be addressed starting from the corpus compilation and going further to the processing of the information

extracted from the group of documents and to the tasks related to the semi-automatic semantic annotation procedure.

3.1. *Corpus compilation*

The activity takes its ground from the construction of the corpus from which to start to run the semi-automatic semantic annotation process. For this task the documents which have been selected are part of Italian paper-handwritten clinical record collections randomly chosen from several medical areas of specialization, e.g., *nephrology, neurogenetics, radiology, biology,* etc., particularly isolating the anamnesis sections. Attention has been paid in selecting clinical records showing a regular structure in the disposition of the semantic contents in order to facilitate the process of identification of recurrent patterns that should provide the information to be semantically normalized as to become the main semantic categories/tags. These latter had to be selected to constitute the reference classification of the medical categories to be retrieved throughout the medical documentation compiled as the source corpus. Indeed, the clinical records analyzed for this study resulted to be characterized by a repetitive configuration in the information provided: the initial set of data refers to the patients' personal data (*age, date and place of birth, city of residence, marital status, date of death*), followed by the anamnestic documentation, physical examination, and the clinical diary for the disease follow-up, including the drugs therapy. Moreover, the clinical record includes the diagnostic tests which have been performed by the patient. This semantic prototype is meant to be shared within the scientific community of physicians belonging to specific medical environments and can be considered an analytical platform from which to begin to implement real-time data in a computer-assisted application designed with appropriate categories according to each applicative medical sphere.

3.2. *Analysis of documentary corpus*

As above mentioned, the clinical records have been digitized and processed. This has been possible through Transkribus[1], a software for handwritten text recognition in order to make the textual information shareable for the automatic medical entity detection (signs, symptoms, etc.).

The identification of the recurrent physicians' text-structures follows a phase of semi-automatic training within the selected software that performs text recognition and transcription of specific documents. This phase addressed the management of the clinical records corpus, grouping documents according to the multiple medical handwriting styles as well as the evaluation of the

[1] <https://readcoop.eu/transkribus/> (last consultation: 15/09/2021).

contents' structure, i.e., identification of the information to be semantically annotated (e.g., anamnesis/psychology, instrumental tests). During this task, the involvement of a domain expert has been necessary. Indeed, it is thanks to the support of the experts that the knowledge organization tools become more accurate and reliable and a means to help the community of physicians in retrieving precise and domain-oriented information for their practical uses (Kos, Kosar, and Mernik 2012). Therefore, the manual scanning of clinical records has been performed in order to carry out the clinical data digitization process to be imported in the text recognition software. The scanned textual documents in output have been then imported into the tool for the semi-automated document recognition and transcription. As previously mentioned, the selected software to run these operations is Transkribus, a platform to automatically recognize texts with any kind of writing. It is used by many other institutions[2] to perform text recognition tasks specifically referring to archival contexts and historical documentation where handwritten systems are most frequent. The software elaborates the information starting from a very precise collection of documents where texts which are likely to present the same structure are included. This constitutes a baseline from which to start to identify the text regions of the documents which will be automatically recognized once having trained the tool with the correct graphically located place of the words in the records by matching them with the corresponding transcribed forms. The model created is able to recall the learned operations to recognize the text regions and it improves the quality of the text identification progressively when dealing with large collections in an automatic way. That means to train the software to interpret the handwritten styles to automatically detect the text of each new clinical record that will be further integrated in the source corpus.

For this activity a number of 243 pages of clinical records has been taken into account to constitute the training model. The percentage of Character Error Rate (CER) - [(i + s + d) / n]*100, where number of characters (n), minimum number of insertions (i), substitutions (s) and deletion (d) - computed is not really low, but it stands at 13% on the training set and on 12% on the test set, meaning that they are equally distributed in the recognition of words, as depicted in Fig. 1.

[2] For example, *The National Archive of Finland, Amsterdam City Archives*.

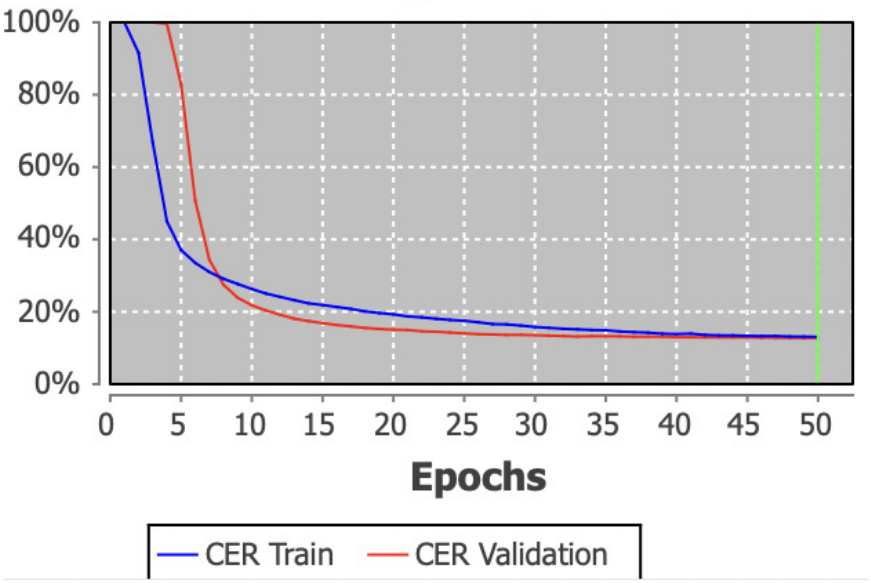


Figure 1: Transkribus accuracy measurement.

The transcription of clinical records represents the digitized source corpus to process with subsequent text mining tasks, which include the semi-automatic tagging operations on these documents. The transcription export allows the selection of the preferred formats that can be computable for the following entities recognition operations, thus fostering interoperability between information systems.

To sum up, the first phases of this research activity dealt with (i) the acquisition of the clinical information by treating the paper-handwritten collection of documents through a scanning procedure; (ii) the manual transcription of the training set of documents; (iii) the automatic recognition of the text regions with Transkribus; (iv) the export of the output of the digitized clinical records to be used as new source corpus to execute the semantic annotation operations; (v) the selection of the main tagsets; (vi) the annotation of the digitized clinical records.

4. Semantic annotation

Anamnesis and test sets have been semantically annotated in order to identify and classify the most relevant information from the documents. In this phase, the tasks addressed the designation and isolation of the representative parts of clinical non-structured texts in order to retrieve salient health information to be associated with medical history of patients (Pinal et al. 2018).

The creation of these patterns' clusters is strictly associated with the tagging framework which provides a statistical representation of the main elements to be notified in defining or reporting medical conditions about patients. The analysis of clinical records aims at providing as output significant informative results from medical records, and, as consequence, annotate through medical entities the clinical reports datasets (Abbas 2021).

The transcriptions have been then semantically annotated in a prototype model using a specific linguistic semantic tool, CATMA annotator, developed for the purpose of setting out an iterated system for fixed expressions retrieval in documentary collections. Since clinical records are characterized by a structured, regular textual outline, the employment of regular expressions referring to text segmentations has supported in a consistent way the automatic association of anamnesis elements association elements for each patient. The annotation process, depicted in Fig. 2, has been carried out in a first instance by isolating the most remarkable information in the *clinical records* that could offer important medical pathways to collect semantic data about the diseases, or better the patients' health history (*anamnesis*), and the treatments associated (*tests*). In this way, multiple clusters of categories, i.e., *tagsets*, representing the selected medical information have been set out. The objective of this activity is the creation of a group of tagsets organized in a sub-tags system meant to create a framework of semantic annotations, i.e., e-dictionaries or formal grammars, to be applied on the clinical records textual structure to automatically detect the most frequent and relevant linguistic elements observed in the source corpus documents. Once these tagsets have been validated by medical professionals, they have been connected to typical and recurrent phrases used by doctors in their diagnosis within the digitized clinical texts. These latter have been semantically normalized in order to be associated to specific *tags* within the structure of the digitized clinical records in a cyclic manner. In this way, the identification of structured information, coming from an informative representation format under the lens of natural language, provides the basis from which to start to implement an inferential systematization about clinical patients' histories.

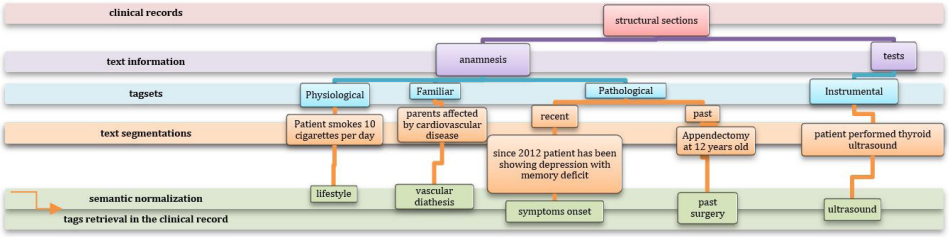


Figure 2: Outline of the annotation system.

Indeed, the output in terms of clinical trends retrieved with the aid of the tags, identified through the aforementioned process, are the results of the informative flow distribution marked from the semantic structures commonly employed by the reference doctors in drafting the clinical records (Fig. 3).

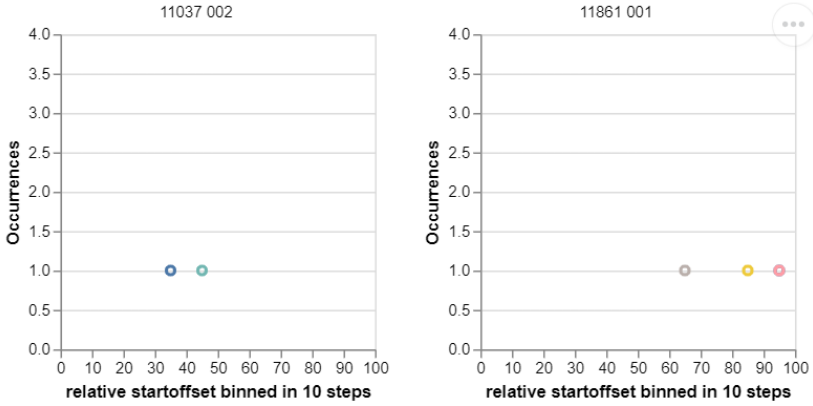


Figure 3: Distribution of tagsets in two clinical records.

More in detail, the annotation process started with the definition of the main categories selected in accordance with the physicians and the existing works on the semantic tags identification for the medical domain. These tags are meant to integrate the texts with supplementary information that relate the writer's fixed expressions usages through distance reading connective operations (Horstmann 2020). Some of the categories/tagsets established in the CATMA tool are the following: *anamnesis, instrumental_tests, autonomy_level, drugs_therapy,* and they are characterized by salient contextual information, as shown in Fig. 4, that forges the clinical frameworks about the disease typicality.

Keyword in context

Left Context	Keyword	Right Context
DOPP. T.S.A : "	riscontra le placche " .	
/02/2016 minuto	Controllo ECG : RS Freq.	Prosegue trattament.
/ la sera. /	Deve fare ECG di controll	dopo 7-10 g
modesta comp. c.v.	Deve ripetere ECODOPP	+ OMOCISTEINA. Co..
B12 Folati. OMOCIST...	ECG /	bradicardia sinusale .
(Dic. 2016)	ECG : BRADICARDIA	sinusale (Recente) .
Visite di Follow-up	ECG: freq. 54/minuto	ECO Dopp. T.S.A :
sinusale (Recente) .	ECO DOPP. T.S.A	: " riscontra le placche
freq. 54/minuto	ECO Dopp. T.S.A[...]nto in	, poi OK- 14
BLUNORM RET 1/die	ECODOPP. T.S.A OK	- 02/05/

Figure 4: Isolation of instrumental_tests tagset in phrases.

Since the clinical records are characterized by non-structured language, the process of semantic annotation has also implied a post-editing of the fixed expressions in order to make the task of tags' recognition easier to perform and to create a more standardized end-vocabulary meant to be exploited. This method refers to the normalization activity that has as its main objective that of providing a standardization of the documents' concepts translated in categories that embed unique terms. Therefore, the standardization process of the information obtained from the tagging task has mapped each expression in the clinical texts with a corresponding normalized form both from the linguistic point of view (e.g., memory dis., dis. of memory, dis. Memory → memory disorders) and from a semantic one (e.g., the patient has incontinence → genitourinary disorders) that can be inferred by the variants included in the anamnesis schemas' expressions present in the clinical record documents constituting the source corpus. This matching matrix is obtained following a selection process of each tagset within the source corpus and allows the launch of a quantitative analysis over the sample of clinical records under study.

The tagging system will give as output a data model (e.g., xml file, text file, matrix of normalized symptoms that are connected with the linguistic form within the source documents – 'the patient gets lost on the road'/spatial disorientation) that is meant to support the implementation of early detection algorithms for patients' diseases.

5. Conclusion and future perspectives

The purpose of this research activity has been based on the organization of the informative contents to automatically annotate upcoming input clinical records by relying on the accepted semantic categories to be identified. Indeed, the huge amount of clinical data produced and stored everyday by health providers needs to be collected and managed in a precise way (Cardillo, Chiaravalloti, and Pasceri 2016), even when some of this knowledge lay in clinical records and some information could be lost. In such a clinical sub-domain there is some information essentially in narrative form that needs to be formalized in order to extract "tacit" knowledge from it (e.g. the good success of pharmacological treatment, some details associated to early onset of a disease, etc.).

The approach adopted in this project is essentially based on document management methodologies, but it can be considered interdisciplinary, because, for the fulfillment of some objectives, it will require cooperative interaction among knowledge organization experts, natural language technicians and medical experts. The result of this work is an accurate prototype of the semantic iterative structures analysis that constitute the clinical record documentation sections. The creation of such a tag system could be used as main

clinical categories to which to refer and that can be expanded each time information about a patient's clinical history changes.

The underlining innovative side of this paper leans on the digitization process of paper-based free-texts in the Italian language referring to several spheres of health systems. This specific task has represented the foundation phase before treating the documents with NLP techniques and setting a reference tagset.

Another interesting output provided by this semantic annotation process has referred to the exploitation of a software able to recognize the several handwriting styles of the doctors responsible for writing clinical records. This process should be iterative for all the doctors' handwriting styles in a future perspective in the logic of augmenting the text corpus of clinical records by using stylometry techniques. The distance reading methods applied to the recognition of medical text styles can constitute the basis for an automatic association of the medical figures' prototypical examinations that could filter out the information depending on authorial attributions.

Moreover, as previously mentioned, the semantic annotated structure developed can represent a key model to be integrated in a new database or to develop new health apps which could possibly be introduced within medical institutions.

References

Abbas, Asim, Muhammad Afzal, Jamil Hussain, Taqdir Ali, Hafiz S.M. Bilal, Sungyoung Lee, and Seokhee Jeon. 2021. "Clinical Concept Extraction with Lexical Semantics to Support Automatic Annotation" *International Journal of Environmental Research and Public Health* 18, no. 20: 10564. <https://doi.org/10.3390/ijerph182010564>.

Attardi, Giuseppe, Vittoria Cozza, and Daniele Sartiano. 2015. "Annotation and Extraction of Relations from Italian Medical Records." In *Proceedings of 6th Italian Information Retrieval Workshop*, Cagliari: CEUR Workshop Proceedings.

Campillos, Leonardo, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. "A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)." *Language Resources and Evaluation* 52, no. 2: 571-601.

Cardillo Elena, Maria Teresa Chiaravalloti, and Erika Pasceri. 2016. "Healthcare Terminology Management and Integration in Italy: Where we are and What we need for Semantic Interoperability." *European Journal for Biomedical Informatics* 22, no.1: 2.

Gius, Evelyn, Christoph Meister Jan, Meister Malte, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, and Jan Horstmann. 2021. CATMA 6 (Version 6.3), <https://zenodo.org/record/4353618>.

Hassanzadeh, Hamed, Anthony Nguyen, and Bevan Koopman. 2016. "Evaluation of medical concept annotation systems on clinical records." In *Proceedings of the Australasian Language Technology Association Workshop*, 15-24.

Horstmann, Jan. 2020. "Undogmatic Literary Annotation with CATMA". In *Annotations in Scholarly Editions and Research: Functions, Differentiation, Systematization* edited by Julia Nantke and Frederik Schlupkothen, 157-176. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110689112-008>.

Kos, Tomaž, Tomaž Kosar, and Marjan Mernik. 2012. "Development of data acquisition systems by using a domain-specific modeling language." *Computers in Industry* 63 (3), doi:10.1016/j.compind.2011.09.004.

Meister, Jan Christoph, Evelyn Gius, Jan Horstmann, Janina Jacke, and Marco Petris. 2017. "CATMA 5.0 Tutorial.", <https://dh-abstracts.library.cmu.edu/works/4195>.

Muhie Seid, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. "Webanno: A flexible, web-based and visually supported system for distributed annotations." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1-6. Sofia: Association for Computational Linguistics.

Oronoz, Maite, Arantza Casillas, Koldo Gojenola, and Alicia Perez. 2013. "Automatic annotation of medical records in Spanish with disease, drug and substance names." In *Iberoamerican Congress on Pattern Recognition*, 536-543, Berlin: Springer.

Owen, John M. 2006. "The digitization of information resources." In *The Scientific Article in the Age of Digitization*. *Information Science and Knowledge Management*, vol. 11. Dordrecht: Springer, <https://doi.org/10.1007/1-4020-5340-1_4>.

Pinal, Patel, Disha Davey, Vishal Panchal, and Parth Pathak. 2018. "Annotation of a large clinical entity corpus." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2033-2042.

Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." *Journal of the American Medical Informatics Association* 17, no. 5: 507-513.

Searle, Thomas, Zeljko Kraljevic, Rebecca Bendayan, Daniel Bean, and Richard Dobson. 2019. "MedCATTrainer: A biomedical free text annotation interface with active learning and research use case specific customisation." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Hong Kong, China, November 2009*, 139-144.

Stewart, Samuel Alan, Maia Elizabeth von Maltzahn, and Syed Sibte Raza Abidi. 2012. "Comparing Metamap to MGrep as a Tool for Mapping Free Text to Formal Medical Lexions." In *KECSM@ ISWC*, 63-77.

Tou, Huaixiao, Lu Yao, Zhongyu Wei, Xiahai Zhuang, and Bo Zhang. 2018. "Automatic infection detection based on electronic medical records." *BMC bioinformatics* 19, no. 5: 117.

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

Anno 39

N. 3-4 – luglio-dicembre 2021

Contributi

Contributi in memoria di Maria Pia Carosella

Note e Rubriche

mundaneum

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

ISBN 979-12-5965-090-0
9 791259 650900

ISSN 1121-0095
9 770112 100950