AIDAinformazioni
Rivista semestrale di Scienze dell'Informazione
Anno 42
N. 3-4 – luglio-dicembre 2024

Contributi

Monica Consolandi, Simone Magnolini, Mauro Dragoni
Risk Communication and Misunderstandings. The Healthcare Dominion Case

Caroline Djambian
Enhancing technoscientific heritage and introducing concepts into culture in the age of data, semantic web, and Artificial Intelligence. The ITinHeritage project

Erika Pasceri
I Knowledge Organization Systems come strumenti di sorveglianza sanitaria. L'applicazione della ICD in Italia

Grazia Serratore
La cartella clinica tra analogico e digitale: gestione e conservazione a norma

Salvatore Spina
Schiavitù e *invented archives*. Intelligenza Artificiale generativa nella costruzione del *Database on the Slave Trade*

Rubriche

Antonietta Folino
Recensione del volume *ClaG – Classificazione dei giochi per ludoteche e biblioteche*

Claudio Gnoli
Il macellaio pragmatista

mundaneum

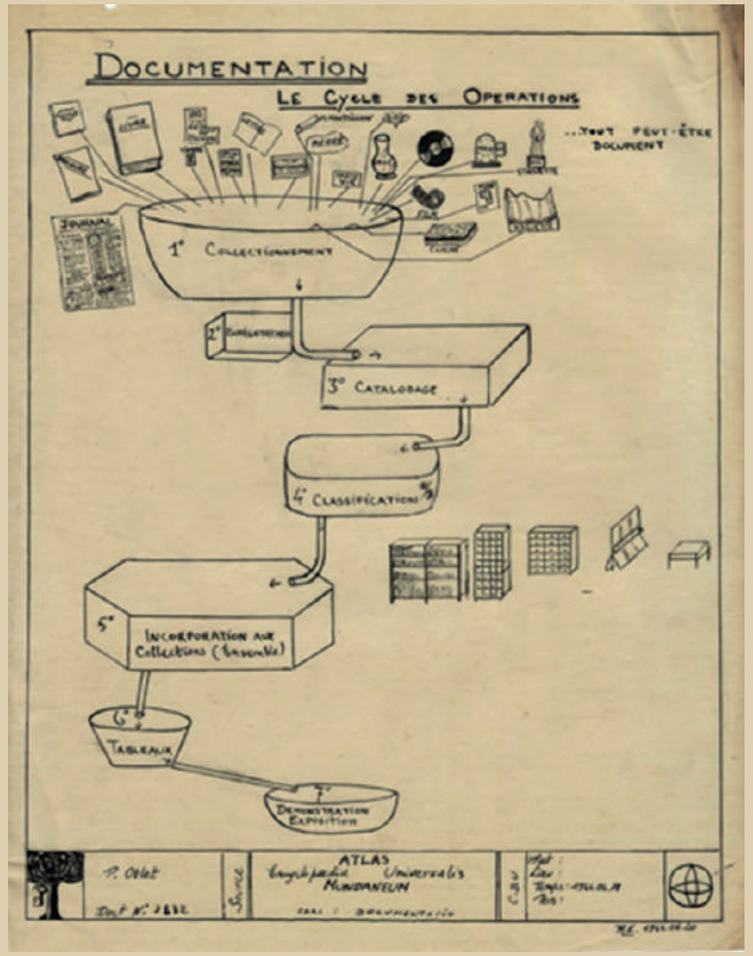
In copertina
Disegno di Paul Otlet, Collections Mundaneum, centre d'Archives, Mons (Belgique).

ISBN 979-12-5965-549-3
ISSN 1121-0095
9 791259 655493
9 770112 100950

AIDAinformazioni

RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

NUMERO 1-2
ANNO 43
GENNAIO-GIUGNO 2025



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

Comitato scientifico:
Anna Rovella, *Università della Calabria*;
Maria Guercio, *Sapienza Università di Roma*;
Giovanni Adamo, *Consiglio Nazionale delle Ricerche* †;
Claudio Gnoli, *Università degli Studi di Pavia*;
Ferruccio Diozzi, *Centro Italiano Ricerche Aerospaziali*;
Gino Roncaglia, *Università della Tuscia*;
Laurence Favier, *Université Charles-de-Gaulle Lille 3*;
Madjid Ihadjadene, *Université Vincennes-Saint-Dénis Paris 8*;
Maria Mirabelli, *Università della Calabria*;
Agustín Vivas Moreno, *Universidad de Extremadura*;
Douglas Tudhope, *University of South Wales*;
Christian Galinski, *International Information Centre for Terminology*;
Béatrice Daille, *Université de Nantes*;
Alexander Murzaku, *College of Saint Elizabeth*, USA;
Federico Valacchi, *Università di Macerata.*

Comitato di redazione:
Antonietta Folino, *Università della Calabria*;
Erika Pasceri, *Università della Calabria*;
Maria Taverniti, *Consiglio Nazionale delle Ricerche*;
Maria Teresa Chiaravalloti, *Consiglio Nazionale delle Ricerche*;
Assunta Caruso, *Università della Calabria*;
Claudia Lanza, *Università della Calabria.*

Segreteria di Redazione:
Valeria Rovella, *Università della Calabria*

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

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

«AIDAinformazioni» è una rivista scientifica che pubblica articoli inerenti alle Scienze dell'Informazione, alla Documentazione, all'Archivistica, alla Gestione Documentale e all'Organizzazione della Conoscenza ma amplia i suoi confini in ulteriori campi di ricerca affini quali la Terminologia, la Linguistica Computazionale, la Statistica Testuale, ecc. 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 pubblicati affrontano questioni teoriche, metodologie adottate e risultati ottenuti in attività di ricerca o progettuali, definizione di approcci metodologici originali e innovativi, analisi dello stato dell'arte, ecc.

«AIDAinformazioni» è riconosciuta dall'ANVUR come rivista di Classe A per l'Area 11 – Gruppo Scientifico Disciplinare 11/HIST-04 – Scienze del libro, del documento e storico-religiose e come rivista scientifica 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. È anche annoverata dall'ARES (Agence d'évaluation de la recherche et de l'enseignement supérieur) 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; The Library Catalog of Georgetown University; SBN – Italian union catalogue; Ulrich's; 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 sono inviati in forma anonima a due referee, selezionati sulla base della loro comprovata esperienza nei topics specifici del contributo in valutazione.

AIDAinformazioni

Anno 43

N. 1-2 – gennaio-giugno 2025

CACUCCI EDITORE

BARI

PROPRIETÀ LETTERARIA RISERVATA

© 2025 Cacucci Editore – Bari
Via Nicolai, 39 – 70122 Bari – Tel. 080/5214220
http://www.cacuccieditore.it e-mail: info@cacucci.it

Ai sensi della legge sui diritti d'Autore e del codice civile è vietata la riproduzione di questo libro o di parte di esso con qualsiasi mezzo, elettronico, meccanico, per mezzo di fotocopie, microfilms, registrazioni o altro, senza il consenso dell'autore e dell'editore.

Sommario

Contributi

Monica Consolandi, Simone Magnolini, Mauro Dragoni, Risk Communication and Misunderstandings. The Healthcare Dominion Case 9

Caroline Djambian, Enhancing technoscientific heritage and introducing concepts into culture in the age of data, semantic web, and Artificial Intelligence. The ITinHeritage project 23

Erika Pasceri, I Knowledge Organization Systems come strumenti di sorveglianza sanitaria. L'applicazione della ICD in Italia 45

Grazia Serratore, La cartella clinica tra analogico e digitale: gestione e conservazione a norma 63

Salvatore Spina, *Schiavitù* e *invented archives*. Intelligenza Artificiale generativa nella costruzione del *Database on the Slave Trade* 85

Rubriche

Antonietta Folino, Recensione del volume *ClaG – Classificazione dei giochi per ludoteche e biblioteche* 123

Claudio Gnoli, Il macellaio pragmatista 127

Contributi

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-549-3
DOI 10.57574/59655493XX
pag. 9-22 (gennaio-giugno 2025)
Open access article licensed under CC-BY-NC-SA

Risk Communication and Misunderstandings

The Healthcare Dominion Case

Monica Consolandi, Simone Magnolini, Mauro Dragoni*

Abstract: Risk communication constitutes a complex area of discourse within the healthcare sector, defined by the inherent risk of misunderstandings between healthcare providers and patients and the challenges associated with effectively conveying information about risks. A significant barrier to improving this communication lies in enhancing healthcare providers' awareness of the implicit nuances embedded in pre-operative risk discussions. This study seeks to address this gap in the existing literature by exploring the issue through the framework of the philosophy of language, with a particular focus on pragmatic analysis as a tool to uncover implicit meanings in doctor-patient interactions. By applying this approach to examine instances of risk evaluation of cardiac surgery, the study provides an empirical analysis of collected data, highlighting the limitations of contemporary models in detecting misunderstandings. Finally, the study outlines potential directions for future research, emphasizing the critical need to advance strategies for improving risk communication in medical contexts.

Keywords: Risk communication, Doctor-Patient Interaction, Misunderstandings, LLMs, Artificial Intelligence.

1. Introduction

Risk communication is a complex and heavily debated topic within the context of doctor-patient interactions (Consolandi 2023). Although medicine adheres to evidence-based principles (Timmermans and Angeli 2001; Tanenbaum 1999), it remains inherently imprecise, characterized by uncertainty and probabilistic reasoning (Curi 2017; Diamond-Brown 2016; Sadegh 2012; Whitby 1951). In formal interactions, physicians are tasked with conveying information regarding treatment success rates, potential complications, and side effects. This process inherently involves uncertainty, as medical knowled-

* Fondazione Bruno Kessler, Intelligent Digital Agents Unit at Digital Health and Well Being Center, Trento, Italy. mconsolandi/magnolini/dragoni@fbk.eu. ORCID Monica Consolandi: 0000-0002-1516-1953. ORCID Simone Magnolini: 0000-0003-0170-3472. ORCID Mauro Dragoni: 0000-0003-0380-6571.

ge is predominantly probabilistic rather than deterministic. Risks in medicine - such as treatment failure, surgical complications, or post-operative recovery challenges - are unavoidable and necessitate careful communication (Zipkin et al. 2014; Visschers et al. 2009).

The assessment of risks must balance the patient's best interests, prioritizing the communication of high-stakes risks. For example, while comparing treatment options for minor conditions versus surgical procedures, the risk-benefit ratio becomes more critical in the latter. Several efforts have sought to categorize uncertainty in relation to medical risks. Uncertainty may stem from its source, such as incomplete information, inadequate understanding, or equally viable alternatives, or from its issue, namely the specific outcomes, situations, or alternatives to which the uncertainty applies (Lipshitz and Strauss 1997; Han et al. 2011).

Weinfurt (2008) expands on Miller and Joffe's typology of uncertainty by identifying additional sources, including uncertainty about causal agents, the validity of surrogate endpoints, the generalizability of findings (e.g., using past data to predict outcomes in new cases), estimation errors, and the population-level nature of statistical estimates, which may not directly apply to individual patients. In phase 1 oncology trials, Miller and Joffe (2008) emphasize that effective informed consent requires clinicians to clearly communicate the probability, magnitude, and duration of potential benefits and risks, along with the associated uncertainties. As Weinfurt highlights, these tasks are not straightforward. Physicians must possess a robust understanding of both the benefits and risks of medical interventions, which is complicated by the inherent uncertainties. Simultaneously, patients must comprehend the information provided by their physicians, a process that Gigerenzer and Edwards (2003) identifies as critical. This dual requirement underscores the importance of both the content and the manner in which information is conveyed. Effective risk communication necessitates specific skills, as patients are often confronted with sensitive topics, including risks of mortality. This dimension has garnered significant interest among psychologists, as poor communication can exacerbate emotional distress (Wells and Kaptchuk 2012). Conversely, clear and empathetic communication fosters trust between doctors, patients, and their families (Wei et al. 2020; Riva et al. 2012; Heyland et al. 2002). Studies have shown that effective risk communication directly correlates with higher levels of trust in the doctor-patient relationship (Engdahl and Lidskog 2014; Duffy et al. 2004).

Risk communication is also explored in broader discussions, including its role in informed consent (General Medical Council 1998), patient rights (Ibrahim et al. 2018), and the physician's ethical obligation to disclose the truth (Jackson 1991). These issues have been examined extensively in clinical ethics and bioethics (Chakrabarty et al. 2012; Gillett 2006), as well as through

the lens of the philosophy of language (Weinfurt et al. 2003; Weinfurt 2018). Philosophical analyses of language explore the implications of verbal expressions, particularly their ability to clarify or obscure meaning (Sbisà 2007). When applied to doctor-patient interactions, such analyses have the potential to enhance speakers' awareness, improving overall communication during discussions of risk.

This paper seeks to address the challenge of detecting misunderstandings in risk communication dialogues within healthcare. Despite its importance, this research topic has received limited attention in the literature due to the scarcity of resources and the complexity of the task (Ardissono et al. 1997). We begin by examining this issue from a philosophical perspective, focusing on the implicit dimensions of dialogue and the role of trust. We then present use cases drawn from a dataset, detailing a preliminary classification of misunderstandings observed in analyzed dialogues. The classification framework used for annotating misunderstandings follows the codebook developed by Rossi and Macagno (2020), which specifically addresses risk communication in healthcare. Finally, we assess the capabilities of state-of-the-art large language models (LLMs), such as OpenAI's GPT-4, in recognizing and analyzing misunderstandings. This assessment is conducted through a qualitative analysis of a single dialogue, shedding light on the limitations and opportunities for improving automated approaches to this task.

2. Language and its dimensions

The implicit dimension of discourse carries profound meaning, and philosophers of language have long sought to uncover its nuances. Grice (1975) was the first to introduce the concept of implicatures to describe these implicit aspects of communication. According to the *Stanford Encyclopedia of Philosophy*, implicature refers to either «(i) the act of meaning or implying one thing by saying something else, or (ii) the object of that act». For example, when one asks, "Do you know what time it is?" the expected response is not a mere "Yes", but an indication of the actual time, such as "Noon" or "It's late". Implicitly, the question communicates more than its literal meaning: it may express a lack of knowledge about the time or, alternatively, inquire whether someone is punctual. The same wording thus conveys varied implications depending on the context and intent.

Gricean implicatures reveal that communication relies not only on explicit content but also on the unspoken meanings that arise through what remains unsaid. Implicatures can be categorized into two types: (i) those that are conventional and inevitable, and (ii) those resulting from deliberate choices by the speaker. Both types are context-dependent and essential for effective communication. The first category – conventional implicatures – comprises those

that are essential for facilitating communication. For instance, when a doctor offers a patient "some water", it is implicitly understood that the water will be provided in a drinkable form and served in a suitable container. Such conventions enable smooth interactions by omitting unnecessary details. However, these conventions are culturally contingent and may vary significantly across societies (Austin 1962). Misunderstandings can arise in cross-cultural contexts, as exemplified by the differing meanings of a nod in Greece versus other cultures. These variations highlight the concept of "felicity" (Austin 1962), which governs whether a conversational convention is successfully applied in a given context. The second category – conversational implicatures – is more intricate, as it arises from intentional choices by the speaker. These implicatures depend on factors such as clarity, complexity, and context. For example, when a dermatologist says "Bring me your exams tomorrow", the speaker does not explicitly state the time, location, or type of exams. The implicature assumes that the exams should pertain to skin pathology, be delivered the next day, and be brought to the doctor's office. Unlike conventional implicatures, conversational implicatures are context-specific and rely on the interlocutors' shared understanding.

While Blečić (2017) suggests that both doctors and patients should avoid conversational implicatures, we contend that implicatures are an unavoidable component of communication. However, we agree with Blečić that both parties could benefit from enhanced skills in recognizing and interpreting implicatures. The philosophy of language seeks to systematize such interactions by illuminating implicit meanings, identifying gaps, and correcting misunderstandings. By analyzing implicatures in doctor-patient communication, we can map the interaction more comprehensively and address specific challenges arising from misinterpreted or missed meanings.

Grice's cooperative principle underpins effective communication, emphasizing that understanding in dialogue depends on the mutual recognition of a shared goal. This principle aligns with Wittgenstein's (1953) theory of language games, which posits that each speaker brings their own implicit "rules" to an interaction. These unspoken rules shape communication and can generate unease when participants' language games differ. Wittgenstein suggests that a shared "third language game" must emerge, comprising a compromise between interlocutors' distinct rules (Malherbe 2014). In the context of doctor-patient communication, this convergence requires openness from both parties to adopt shared rules within the interaction.

Combining Grice's cooperative principle with Wittgenstein's insights underscores that effective communication relies on mutual trust. Trust facilitates the creation of a shared linguistic framework and fosters a therapeutic alliance. To achieve this, speakers should adhere to Grice's four maxims: (i) the maxim of quality - be truthful; (ii) the maxim of quantity - provide the appropriate

amount of information; (iii) the maxim of relation - remain relevant; and (iv) the maxim of manner - be clear and orderly while avoiding ambiguity and obscurity. While violations of these maxims (e.g., through metaphors, irony, or hyperbole) may occur without breaching the cooperative principle, maintaining awareness of these guidelines ensures effective communication and helps interlocutors stay aligned. In this linguistic framework, both explicit content and implicatures contribute meaningfully to communication. Trust, as the foundation of this interaction, is essential for interpreting both the speakers' active contributions (maxims) and their passive contributions (implicatures). The implicit dimension of communication, though unspoken, plays a critical role in shaping discourse. However, when this dimension is marked by mistrust rather than trust, communication becomes significantly impaired. For instance, consider a scenario where you ask a housemate whether they have locked the apartment door. If you trust their response, a simple "Yes" suffices to conclude the conversation. However, if mistrust exists, you may ask additional clarifying questions, such as "Did you double-lock it? Did you use the latch? Is the alarm activated?". A similar dynamic applies in doctor-patient interactions. If a physician mistrusts a patient's adherence to prescribed medication, they may interrogate the patient with questions such as "Have you taken your pills? All of them? Every day? After meals, as instructed?". This lack of trust transforms the interaction into an interrogation, disrupting the clinical relationship.

These challenges become particularly pronounced during moments of risk communication. Trust, mutual understanding, and the ability to navigate implicatures are crucial for effective discussions of risks, where the stakes are inherently high (Chalmers 2002). Mismanagement of these implicit dimensions can lead to significant misunderstandings, undermining the therapeutic alliance and the quality of care (Consolandi et al. 2024).

3. Communicating risks

Effective communication is anchored in reciprocal trust, a principle critical not only for successful doctor-patient interactions but also for the broader doctor-patient relationship (Dugdale 2019; Jackson 2002). Trust, as previously emphasized, is foundational to building a robust therapeutic alliance. By applying the Gricean maxims, this alliance can be envisioned as a virtuous circle where trust and truth mutually reinforce one another. The concept of trust becomes clearer when dissecting its definitions. According to the Cambridge Dictionary, the noun "trust" is defined as «the belief that you can trust someone or something». Its verb form, "to trust," means «to believe that someone is good and honest and will not harm you, or that something is safe and reliable; to hope and expect that something is true». Trust, therefore, is inherently a

triadic relationship: "X trusts Y to do Z". In clinical settings, this is translated to "P (the patient) trusts D (the doctor) to do X (a specific action)", and conversely, "D trusts P in doing X".

Similarly, the concept of truth, defined as "the quality of being true", gains depth when considered alongside its adjective form, "true", which implies being «sincere or loyal, and likely to continue to be so in difficult situations» and «having all the characteristics necessary to be accurately described as something». Two recurring themes emerge from these definitions: (i) sincerity, loyalty, and reliability, which relate to the trustworthiness of the relationship, and (ii) accuracy, emphasizing the necessity of precise descriptions or evidence supporting the interaction (e.g., a specific treatment in the clinical context).

This dual aspect of trust and truth highlights a central challenge in risk communication: providing precise descriptions is often impractical or unhelpful when addressing open-ended subjects like risk. At the same time, withholding certain information can undermine trust, even though, as Klitzman (2007, 514) notes «the desire to establish trust can conflict with the imperative to disclose the whole truth». This tension complicates the virtuous circle of trust and truth, especially when truth must navigate patient expectations ("...will not harm you...") and the inherent uncertainties of communication. As Wittgenstein observed, every individual enters communication with their own language game, a unique system of rules shaped by personal experiences and roles. In doctor-patient interactions, these language games are distinct: the doctor, clad in professional attire and documenting the interaction in a medical record, represents the medical profession. The patient, by contrast, typically appears unwell and lacks visible markers of their symptoms. This juxtaposition underscores two primary systems of rules: those of the professional healthcare provider and those of the unwell individual. These roles carry profound implications. The doctor, bound by legal and ethical standards, assumes professional responsibility for the patient's life, while the patient, relying on the doctor's expertise, retains personal responsibility for their own well-being.

Within this framework, two levels of communication coexist: (i) the professional level and (ii) the personal level. The professional level pertains to the doctor's role as a representative of their medical profession. In this capacity, the doctor bears the responsibility to convey information about the patient's condition, treatment options, potential outcomes, and associated risks. Here, the patient plays a more passive role, primarily as a recipient of information. The personal level encompasses the emotional dimensions and personal histories of both parties. This level is where vulnerability, fear, and doubt come to the forefront. Unlike the professional level, both the doctor and the patient are active participants, bringing their personal narratives and emotional experiences into the interaction.

These two levels are intertwined, yet they reveal a double asymmetry in the doctor-patient relationship: i) the first asymmetry arises from the knowledge gap: the doctor, as the medical expert, possesses a depth of understanding about the patient's condition and body that far exceeds the patient's own knowledge. This expertise places the patient in a passive role on the professional level, where they depend on the doctor's guidance; ii) the second asymmetry highlights the dual role of the doctor. On one hand, they must maintain their professional identity and adhere to the boundaries of their role, ensuring clear and accurate communication within the context of medical care. On the other hand, the doctor is a human being with personal emotions and vulnerabilities, which can surface during interactions. Balancing these professional and personal dimensions is a challenge unique to the medical field. This dynamic is reflected in legal frameworks, such as Italy's Gelli-Bianco Law (n. 24/2017), which addresses criminal liability arising from medical errors, underscoring the gravity of professional responsibilities. These asymmetries reveal the complexity of doctor-patient communication, particularly in the realm of risk communication. Risk communication is not conducted in a neutral environment; it is shaped by the interplay of professional expertise, personal emotions, and the delicate balance of trust and truth. Successfully navigating this intricate dynamic requires sensitivity to the implicit dimensions of communication and a commitment to fostering mutual trust within the therapeutic alliance.

4. From theory to practice

This study approached the challenge of misunderstanding classification as a text classification task, which involves assigning a specific label or class to a given segment of text. As there are no directly comparable classification tasks currently available, we began developing the system from scratch. To address the task effectively, we selected *XLM-RoBERTa* (Conneau et al. 2020), a robust general-purpose language model, as the foundation for our system. *XLM-ROBERTA* is renowned for delivering state-of-the-art performance across multiple languages and a variety of tasks (Wang and Banko 2021). The pre-trained *xlm-roberta-base* model, freely available on the Hugging Face Model Hub (Wolf et al. 2020), provided an excellent starting point for adaptation to our specific task requirements.

Our dataset[1] consists of 32 doctor-patient interviews. In total, these interviews include 7,172 conversational turns, with an average of 230 turns per interview. These interactions were meticulously analyzed, and misunderstan-

[1] Dataset available at (Zenodo 2024). See Consolandi et al. 2020 and Consolandi 2024 for an exhaustive overview of the data collection and the methodology.

dings were manually annotated, yielding 320 instances of misunderstandings, representing 4.4% of the total conversational turns. We carried out an in-depth analysis of the misunderstandings and problematic understandings that surfaced during the doctor-patient interviews, and distinct patterns emerged based on the participants' roles. For doctors, the majority of these instances appeared during the anamnesis phase, where gathering a patient's medical history often introduced ambiguities or misinterpretations[2]. On the other hand, for patients and caregivers, misunderstandings were most prominent in discussions about future treatment plans, with a smaller but notable number occurring during the diagnosis phase. These findings underline how different phases of communication carry unique challenges depending on the role and perspective of each participant.

To fine-tune the model and assess its performance, we split the dataset into three parts following a typical 80/10/10 proportion. This resulted in a training set of 254 instances, while both the development set and test set contained 32 instances each. Importantly, for consistency and practicality during training, conversational turns containing more than 50 interactions were truncated at this limit. This step helped streamline the data without losing its core interpretability. However, as often happens with real-world datasets, we encountered a significant class imbalance, particularly in the training data, which raised concerns about the model's ability to generalize effectively across all types of misunderstandings. However, the test set was more balanced, better representing how misunderstandings naturally occur, which gave us confidence that the evaluation results would closely reflect real-world situations. That said, one limitation that surfaced during evaluation was the absence of certain classes in the test set, which could hinder the model's ability to fully adapt to all misunderstanding types.

To mitigate the imbalance issue in the training set, we opted for a simple yet effective strategy: random oversampling of the underrepresented classes. By increasing the representation of minority classes, this method ensures that the model has a fair opportunity to learn from all misunderstanding types, rather than being disproportionately biased toward the majority classes. This balanced approach of working with real-world distributions while addressing data limitations through targeted techniques strengthens the foundation of our classification system. By refining the training process and accounting for the inherent complexities in the data, we aim to improve the model's ability to identify and classify misunderstandings with greater nuance and accuracy. Oversampling helps address class imbalance and allows the model to classify smaller classes more effectively. However, we noticed that increasing the number of training epochs with this approach leads to a drop in accuracy for the

[2] In-depth analysis of the results of the study protocol can be found in Consolandi et al. 2024.

oversampled classes. This is likely due to overfitting, where the model becomes too focused on the repeated examples in the training data, reducing its ability to generalize to new instances. The system is now able to understand the content of the different instances and to better classify this type of misunderstanding.

5. Evaluation phase

To evaluate our assumptions about the ability of LLMs to identify and analyze misunderstandings in dialogues, we conducted an experiment using two conversations involving patients 16 and 18. These cases were chosen because they exhibited clear instances of miscommunication. In the first dialogue, the patient initially states that they smoke but later clarifies that they are referring to e-cigarettes (vaping) when asked about the quantity of cigarettes. In the second dialogue, the patient mistakenly assumes the doctor is referring to a surgical procedure when the doctor begins discussing therapy.

The experiment aimed to achieve three primary objectives:

- Assess the system's ability to recognize misunderstandings.
- Investigate whether the length of the dialogue fragment influences the results.
- Determine if the system can detect potential misunderstandings, even when the participants do not acknowledge them.

The third objective is particularly complex, yet crucial to our study. In the context of doctor-patient communication, misunderstandings that go unnoticed or unaddressed can pose significant challenges to effective interaction.

We set up the experiment using the OpenAI Python library (n.d.) and the latest GPT-4 model. To ensure consistent results, we used a fixed temperature (0) setting. To maintain consistency with the dialogues, we wrote the prompts in Italian. Given the qualitative focus of the experiment, which prioritized understanding over performance metrics, we did not invest heavily in extensive prompt engineering. We thus simply ask to highlight misunderstandings between participants if found.

Based on our criteria, only one instance from the result qualifies as a true misunderstanding, while the rest are speculative, relying on the reasonable assumption that doctor-patient communication is often prone to difficulties. As anticipated, when we reduced the size of the dialogue surrounding the identified misunderstanding, GPT-4 demonstrated a noticeable improvement in its ability to focus on the specific misunderstanding. Despite GPT-4 presenting the misunderstanding across several points, all of them ultimately relate to a single issue - the patient's mistaken belief that she is undergoing surgery, despite the doctor's clarification that this is not possible.

Our final evaluation aimed to determine whether GPT-4 truly identifies misunderstandings or simply reacts to participants' responses. To explore this, slight modifications to the dialogues were required. Fortunately, the dialogues were authentic doctor-patient exchanges where misunderstandings are naturally resolved, allowing us to make adjustments. Specifically, we moved the explanation about the surgery's impossibility earlier in the conversation and altered the patient's response to be more assertive. These minor changes led to a significant shift in the analysis.

The analysis revealed a key flaw: GPT-4's recognition of misunderstandings seemed to depend more on the participants' reactions than the substance of the dialogue.

6. Conclusions

This study addresses the challenge of detecting misunderstandings within risk communication dialogues in healthcare. We begin by examining this issue from a philosophical perspective and subsequently present a focused use case that demonstrates the significant shortcomings of state-of-the-art approaches in achieving satisfactory performance. The findings support the hypothesis that the complexity of this task necessitates further exploration. Notably, framing the problem as a text classification task yields promising initial results. However, the data distribution is highly imbalanced, with a significant bias toward a single class. To mitigate this imbalance, two complementary strategies are proposed: increasing the volume of annotated data for the underrepresented classes and employing more sophisticated data augmentation techniques beyond simple oversampling.

While the initial results are encouraging, our evaluation highlights persistent challenges, even for advanced large language models such as GPT-4. Specifically, these models struggle to accurately identify misunderstandings, often conflating them with general communication uncertainty. In particular, misunderstandings that remain unaddressed by dialogue participants are frequently overlooked entirely by GPT-4.

Future research will prioritize several areas, including the acquisition of additional annotated data and the development of innovative approaches to better detect and interpret moments when dialogues deviate from their intended purpose.

References

Ardissono, Liliana, Guido Boella, and Rossana Damiano. 1997. "A computational model of misunderstandings in Agent Communication." *AI*IA 97: Advances in Artificial Intelligence - 5th Congress of the Italian Association for Artificial Intelligence, Rome, Italy, September 17-19, 1997, Proceedings*, 48-59.

Austin, John Langshaw. 1962. *How to do things with words.* Oxford: Oxford University Press.

Blečić, Martina. 2017. "The Place for Conversational Implicature in Doctor-Patient Communication." *Prolegomena* 16 (2): 117-29.

Chakrabarty, Mrinmoy, Vijay Thawani, and Pinki Devi. 2012. "Transparent Communication in High-Risk Infections: A Bioethical Perspective." *Asian Bioethics Review* 4 (2): 143-49.

Chalmers, Clark C. 2002. "Trust in Medicine." *Journal of Medicine and Philosophy* 27 (1): 11–29.

Conneau, Alex, Kartikay Khandelwal, Naman Goyal, et al. 2020. "Unsupervised Cross-lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*: 8440-51.

Consolandi, Monica, Carlo Martini, Michele Reni, et al. 2020. "COMMUNI.CARE (COMMUNIcation and patient engagement at diagnosis of PAncreatic CAncer): study protocol." *Frontiers in Medicine* 7 (34).

Consolandi, Monica. 2023. "Implicit understandings and trust in the doctor-patient relationship: a philosophy of language analysis in pre-operative evaluations." *Theor Med Bioeth*.

Consolandi, Monica. 2024. "Philosophy Leading the Way: An Interdisciplinary Approach to Study Communication of Severe Diagnoses." *Plos One* 19 (7).

Consolandi, Monica, Mara Floris, Nicolò Pecorelli, et al. 2024. "Communication, understanding and engagement of patients with pancreatic cancer at time of diagnosis." *Pancreatology* 24 (3): 437-44.

Curi, Umberto. 2017. *Le parole della cura. Medicina e filosofia.* Milano: Cortina Raffaello Editore.

Diamond-Brown, Lauren. 2016. "The doctor-patient relationship as a toolkit for uncertain clinical decisions." *Social Science & Medicine* 159: 108-15.

Duffy, Daniel F., Geoffrey H. Gordon, Gerald Whelan, et al. 2004. "Assessing Competence in Communication and Interpersonal Skills: The Kalamazoo II Report." *Acad Med* 79: 495-507.

Dugdale, Lydia. 2019. "Patient as Gift." *The Hastings Center Report* 49 (4): 4-5.

Engdahl, Emma, and Rolf Lidskog. 2014. "Risk, communication and trust: Towards an emotional understanding of trust." *Public Understanding of Science* 23 (6): 703-17.

General Medical Council. 1998. *Seeking patients' consent: the ethical considerations*. London: General Medical Council.

Gillett, Grant. 2006. "Medical science, culture, and truth." *Philosophy, Ethics, and Humanities in Medicine* 1 (13).

Gigerenzer, Gerd, and Adrian Edwards. 2003. "Simple tools for understanding risks: from innumeracy to insight." *BMJ* 327 (7414): 741-44.

Grice, Herbert Paul. 1975. "Logic and conversation." In *Syntax and Semantics 3: Speech Acts*, edited by Peter Cole and Jerry L. Morgan, 41-58. New York: Academic Press.

Han, Paul K. J., William M. P. Klein, and Neeraj K. Arora. 2011. "Varieties of Uncertainty in Health Care: A Conceptual Taxonomy." *Medical Decision Making:* 31 (6): 828-38.

Heyland, Daren K., Graeme M. Rocker, Peter M. Dodek, et al. 2002. "Family satisfaction with care in the intensive care unit: Results of a multiple center study." *Critical Care Medicine* 30 (7): 1413-18.

Ibrahim, Farzana, Per Sandström, Bergthor Björnsson, Anna Lindhoff Larsson, and Jenny Drott. 2018. "'I want to know why and need to be involved in my own care…': a qualitative interview study with liver, bile duct or pancreatic cancer patients about their experiences with involvement in care." *Supportive Care in Cancer* 27 (7): 2561-67.

Legge 8 marzo 2017, no. 24. "Disposizioni in materia di sicurezza delle cure e della persona assistita, nonché in materia di responsabilità professionale degli esercenti le professioni sanitarie." *Gazzetta Ufficiale*, 17 marzo 2017, no. 64.

Jackson, Jennifer. 1991. "Telling the Truth." *Journal of medical ethics* 17 (1): 5-9.

Jackson, Jennifer. 2002. *Truth, trust and medicine.* New York: Routledge.

Klitzman, Robert. 2007. "Pleasing doctors: when it gets in the way." *BMJ* 335 (7618): 514.

Lipshitz, Raanan, and Orna Strauss. 1997. "Coping with Uncertainty: a Naturalistic Decision-Making Analysis." *Organizational Behavior and Human Decision Processes* 69 (2): 149-63.

Malherbe, Jean-Françoise. 2014. *Elementi per un'etica clinica. Condizioni dell'alleanza terapeutica.* Trento: FBK Press.

Miller, Franklin G., and Steven Joffe. 2008. "Benefit in phase 1 oncology trials: therapeutic misconception or reasonable treatment option?" *Clinical Trials* 5 (6): 617-23.

OpenAI Python. n.d. Accessed May 27, 2025. https://github.com/openai/openai-python.

Riva, Silvia, Marco Monti, Paola Iannello, et al. 2012. "The Representation of Risk Medical Experience: What Actions for Contemporary Health Policy?" *PLoS ONE* 7 (11).

Rossi, Maria Grazia, and Fabrizio Macagno. 2020. "Coding Problematic Understanding in Patient–provider Interactions." *Health Communication* 35 (12): 1487-96.

Sadegh-Zadeh, Kazem. 2012. *Handbook of analytic philosophy of medicine*. Springer.

Sbisà, Marina. 2007. *Detto non detto. Le forme della comunicazione implicita.* Roma-Bari: Laterza.

Tanenbaum, Sandra. 1999. "Evidence and expertise: the challenge of the outcomes movement to medical professionalism." *Academic Medicine* 74 (7): 757-63.

Timmermans, Stefan, and Alison Angeli. 2001. "Evidence-based medicine, clinical uncertainty, and learning to doctor." *Journal of Health and Social Behavior* 42 (4): 342-59.

Visschers, Vivianne H.M., Ree M. Meertens, Wim W.F. Passchier, and Nanne N.K. De Vries. 2009. "Probability Information in Risk Communication: A Review of the Research Literature." *Risk Analysis* 29 (2): 267-87.

Wang, Cindy, and Michele Banko. 2021. "Practical Transformer-based Multilingual Text Classification." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, 121-29.

Wei, Dong, Xu Anqi, and Xue Wu. 2020. "The mediating effect of trust on the relationship between doctor-patient communication and patients' risk perception during treatment." *PsyCh Journal* 9 (3): 383-91.

Weinfurt, Kevin P., Daniel P. Sulmasy, Kevin A. Schulman, and Neal J. Meropol. 2003. "Patient expectations of benefit from phase I clinical trials: linguistic considerations in diagnosing a therapeutic misconception." *Theoretical Medicine and Bioethics* 24 (4): 329-44.

Weinfurt, Kevin P. 2008. "Varieties of uncertainty and the validity of informed consent." *Clinical Trials* 5 (6): 624-25.

Weinfurt, Kevin P. 2018. "Propositions and Pragmatics." *The American Journal of Bioethics* 18 (9): 18-20.

Wells, Rebecca Erwin, and Ted J. Kaptchuk. 2012. "To Tell the Truth, the Whole Truth, May Do Patients Harm: The Problem of the Nocebo Effect for Informed Consent." *American Journal of Bioethics* 12 (3): 22-29.

Whitby, Lionel. 1951. "The science and art of medicine." *The Lancet* 258 (6674): 131-33.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Oxford: Basil Blackwell. Original work published 1953 as *Philosophische Untersuchungen*.

Wolf, Thomas, Lysandre Debut, Victor Sanh, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstration*, 38-45.

Zenodo. 2024. "COMMUNI.CARE (Communication and Patient Engagement at Diagnosis of Pancreatic Cancer): Study Protocol." Last modified June 4, 2024. https://zenodo.org/records/11473315.

Zipkin, Daniella A., Craig A. Umscheid, Nancy L. Keating, et al. 2014. "Evidence-Based Risk Communication. A Systematic Review." *Annals of Internal Medicine* 161 (4): 270-80.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-549-3
DOI 10.57574/59655493XX
pag. 23-44 (gennaio-giugno 2025)
Open access article licensed under CC-BY-NC-SA

Enhancing technoscientific heritage and introducing concepts into culture in the age of data, semantic web, and Artificial Intelligence

The ITinHeritage project

Caroline Djambian*

Abstract: In Semantic Web age, enhancing heritage means bringing collections beyond the confines of the museum. But this ethical mission of transmitting culture and knowledge to the general public is made complex by the massive and heterogeneous data that museums now have to manage. As data are now the new collections of museums, they are the starting point for the ITinHeritage project. It focuses on Information Technology (IT) heritage, the preservation of which is of general interest, and raises the question of "How to pass on complex scientific and technical concepts embodied in artefacts not mediators in themselves?" Backed by European science and IT museums, we develop innovative Digital Humanities approaches, combining terminology, Linked Open Data (LOD), and Artificial Intelligence (AI) technologies. In this article we set out the context of our questioning and present the basis of the answer we wish to provide, i.e. the onto-terminological work.

Keywords: Digital Humanities, Scientific and technical heritage, Ontology, Knowledge graph, Linked Open Data (LOD).

1. Introduction

The ITinHeritage project was born of a meeting, as a terminologist and researcher in organization of scientific and technical knowledge in its digital mediation, with members of the Association pour un Conservatoire de l'Informatique et de la Télématique (ACONIT) in Grenoble (France). Grenoble is the birthplace of French computer science, and the Conservatory owns one of the largest collections in Europe. If Information Technology (IT) is the emblem and the witness of contemporary human technoscientific evolution, and the heart of our current society, the more I got to know this museum space, the more I realised how, paradoxically, little IT heritage is known and understood by the public and authorities, and therefor how endangered it was. Further investigations revealed that the difficulties encountered by the

* GRESEC Laboratory (Groupe de recherche sur les enjeux de la communication), Grenoble Alpes University, Grenoble, France. caroline.djambian@univ-grenoble-alpes.fr. ORCID: 0009-0007-3625-305X.

Conservatory were general to IT museums. That's why I decided to create a network of several European museums around a common desire to promote and pass on this heritage. This gave me a fantastic research area with collections of thousands of artefacts. But how, then, preserve this heritage and pass on complex scientific and technical concepts embodied in artefacts that are not mediators in themselves, both to scientists and to the general public?

To answer this question, my initial idea was to use IT technologies to serve their own heritage. So, I began by basing my study on current digital developments embodied in major semantic web projects, which offer an unprecedented opportunity to enhance and disseminate heritage and the knowledge it conveys to the public. These Digital Humanities projects from the Galleries, Libraries, Archives, Museums (GLAM) movement, are based above all on the metadata's standardization work (here, inventories of museum collections), in order to open them up. For example, we can mention Europeana (Doerr et al. 2010), the Société des Musées du Québec, the Royal Rijksmuseum of the Netherlands around the works of the great masters of Dutch painting (Dijkshoorn et al. 2018), JocondeLab for the museums of France (Juanals and Minel 2016), the Smithsonian American Art Museum (Szekely et al. 2013), the Amsterdam Museum (De Boer et al. 2013) and the Getty in Los Angeles, the Bibliothèque Nationale de France with data.Bnf (Simon et al. 2014), Biblissima (Gehrke et al. 2015).

But while this represent a major task, it cannot be an end in itself, because making data easier to consult is not enough to guarantee the museum's mission of transferring culture and knowledge. If data is the starting point, the appropriation of knowledge is our objective, and it is through language that it can be captured. That's why we will focus here on the terminological work that founds the ITinHeritage project, carried out with the help of a multidisciplinary team from the fields of information and communication sciences, computer sciences, linguistics and history of science.

After having presented the particularities of IT's technoscientific heritage, we will look in this paper at: how this heritage, like others, is tending towards the immateriality of data; and how the latter is becoming the raw material for a revisited corpus linguistics approach, placing language at the heart of the transfer and appropriation of knowledge that we are aiming for. On this basis, we will explain why, given the complexity and variability of technoscientific language, the semasiological approach cannot be sufficient, but rather the first stage of a terminological work. Finally, we will explain our method for building an onto-terminology of IT heritage, and how we complement it by Artificial Intelligence (AI) technologies, in order to semantically link museums' open data, while representing technology in its dynamic evolution.

2. An evolving technoscientific subject for study: the Information Technology (IT) heritage

Technosciences, these complex systems that combine technology and science, are the foundation of our contemporary world. They are developing alongside our humanity, transforming it on an unprecedented scale and at an unprecedented speed. Information Technology (IT) is the very embodiment of technosciences and their relationship to our society. They are defined by UNESCO (2009, 121) as «a diverse set of technological tools and resources used to transmit, store, create, share or exchange information. These technological tools and resources include computers, the Internet (websites, blogs and emails), live broadcasting technologies (radio, television and webcasting), recorded broadcasting technologies (podcasting, audio and video players, and storage devices) and telephony (fixed or mobile, satellite, visio/video-conferencing, etc.)». But even though Information Technologies are pervading our humanity, their rapid and massive evolution has not left room for their in-depth and distanced study, for their definition and for their patrimonalization.

Yet studying this heritage can give us a better understanding of our contemporary world and how it has changed since the Second World War. That's what we are aiming for in the ITinHeritage research project, through an heritage-based and interdisciplinary approach funded in Digital Humanities (Djambian et al. 2024a). Our project starts from European scientific and IT museums that long ago began the painstaking work of inheriting, conserving and promoting this heritage. The London Science Museum (London, UK) and the Heinz Nixdorf MuseumsForum HNF (Paderborn, Germany), the Association pour un Conservatoire de l'Informatique et de la Télématique (ACONIT) (Grenoble, France), the NAM-IP computer museum (Namur, Belgium), the Museo degli strumenti per il calcolo (Pisa, Italy) and the HomeComputerMuseum (Helmond, Netherlands) are our partners.

The aim of our study is therefore to define and circumscribe this labile field, through the various forms that its heritage can take. The first form of this heritage is the one that is most obvious at first glance: its physical and digital objects that are its explicit expressions. This includes objects and their documentation, their digital duplicate, the software that represents 1/3 of certain collections, and above all, the metadata of all these artefacts. The growing proportion of data makes it the new collections of science. But it is the tacit expression of this heritage, and the knowledge it encompasses, that we would like to focus on. Much more difficult to grasp, because it is created in the act of experiencing the world, it has been little studied and even less considered as heritage. It echoes the empeiria of Kant (1997) or the metis of Aristotle (1986), whereas the exoteric (Jacob 2001) or pure knowledge of Kant is concretized in writing or in object. Gathering and transmitting this tacit, esoteric

knowledge, is a real challenge because, to be captured, empeiria needs to be shaped into gestures and discourse. We are therefore placing the specialized language at the heart of the ITinHeritage project by combining complementary and innovative approaches aimed at the interaction between terminology, in its conceptual nature, and Artificial Intelligence (AI).

3. New museums' collections and new bodies of work: the data

Data are not only the new artefacts of heritage, they are also the new corpora for researchers (Djambian et al. 2024b). As with many research projects, our terminology work begins with the compilation of a corpus. To build it, we have opted for an approach based more on the information and communication sciences than on linguistics, focusing not on the volume of the corpus but on its high density of specialized terms and concepts, as manifestations of the tacit knowledge that we want to enhance. To this end, we have built a Knowledge Graph that gathers the metadata from our partners museums' inventories of collections. Knowledge Graphs are part of the Semantic Web paradigm, which enable to open and link data with other data from around the world and to easily integrate new data from the web, in a reciprocal enrichment process. To build the ITinHeritage Knowledge Graph we used the Jena environment. We first collected and harmonized the metadata of the various museums, as each one uses its own formats (XML, PDF, ...) and categorises knowledge according to its own references. This harmonization followed the European open data FAIR principles (Findable, Accessible, Interoperable, Reusable). Then we converted metadata to CSV (Comma-Separated Values) and RDF (Resource Description Framework) standards so that it could be opened and linked on the web (Linked Open Data, LOD). But, in order to move from the web of data to the semantic web, it is essential to structure the Knowledge Graph using an ontology. So, we built on the Protégé environment a first-level ontology based on the CIDOC-CRM (International Committee for Documentation of the International Council of Museums, Conceptual Reference Model) (Bekiari 2021) and the EDM (Europeana Data Model) (Doerr et al. 2010) (Fig. 1), two meta-models that are benchmarks in the heritage field.

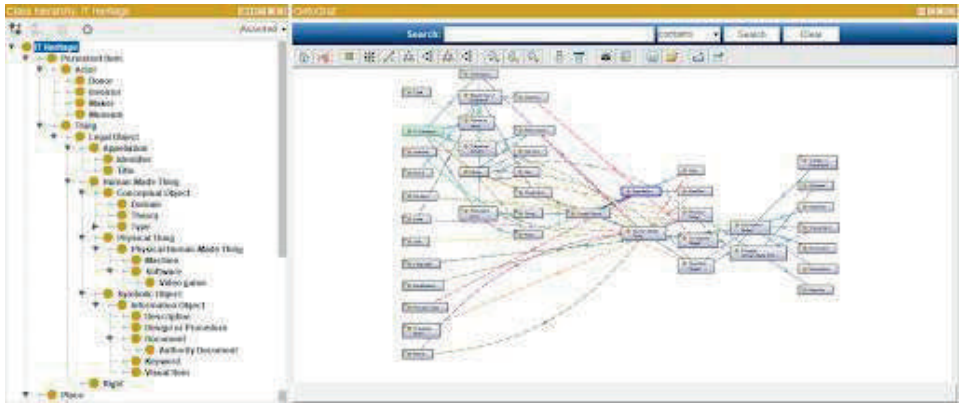


Figure 1: View of he first-level ITinHeritage ontology.

The metadata of 25,500 artifacts are now contained in our Knowledge Graph. The primary aim is to open up museum collections beyond their walls, so that this data can be made available to the public at large, in a spirit of open science. But for us, this Knowledge Graph represent above all our corpus, packed with terms and concepts in the field, written by experts. It represents the new types of corpora resulting from Big Data, which linguists will now have to encounter. In order to define the tacit knowledge of the field, this multilingual corpus (French, English, Italian, German) allows us to study the discourses that express it. This is the starting point for our onto-terminological work based on ISO standards (5078, 704 and 1087).

4. How can technoscientific concepts be made appropriate in the face of the variety and technicality of language?

If we focus on language, we must take it in all its complexity. The first barrier to dealing with technoscientific language is its complexity. Bertrand Russell (1956) asserted that in an ideal language, a word would designate an object, and each object would have its own meaning. However, the complexity of language was already noted in antiquity by Democritus, who observed that a word can designate several objects, that the same object can be designated by several words, that words designating objects vary over time, and finally, that the links between a word and an object have various motivations (Diels 1903). This is why, in languages as complex as technosciences, it is necessary to come to a conceptualization of a world whose elements are designated by unambiguous names (Roche 2007). This step is not an option when we want to transmit specialized knowledge to a wide audience.

To be able to pass on technoscientific knowledge to the layman, the first step is to explain the meaning of the names of the concepts, i.e. the specific terms of the field. It is the notional dimension that makes it possible to cover the various dimensions of a domain and achieve the objective of popularization of its knowledge. To this end, terminology work cannot be carried out without the involvement of the holders of these concepts, as in our case, experts and historians of IT. That's why our work is carried out with the support of Pr. Marie Gevers and Christophe Ponsard (Namur University) from the NAM-IP museum, Xavier Hiron from the ACONIT conservatory, and other experts from the HomeComputerMuseum and the Museo degli Strumenti per il Calcolo).

As we shall see in detail below, we need to integrate a meta-linguistic discourse, and textual and semantic strategies, that will compensate for the accumulation of specialized terms in specialized discourse. In order to make specialized terms, i.e. the names of concepts, appropriable, the descriptive strategy to integrate technoscientific concepts into the culture of the individual receiver must be based on the differentiation of notions (Gaudin 1996).

This is because technoscientific language in itself constitutes an obstacle to access to specialist knowledge, just as much as the lack of mediation of the scientific and technical objects it names. It is the concept that can be appropriated rather than the terms that designate it, even if Putnam (1984) demonstrates that we can manipulate specialized languages without mastering their notions. This is because acquiring a lexical system is not enough, and remains a barrier to the uninitiated because of its opacity. Efforts should therefore be focused on the appropriation of a conceptual system, i.e. on terminology work in addition to linguistic work.

As we said earlier, understanding a new term is based on comparing it with, or differentiating it from a pre-existing notional network. However, the difficulty of integrating new knowledge persists when the existing notional network has no link, no common references, to the new object and/or term. As presented in previous works (Djambian et al. 2024a) and based on those of Gaudin (1996), we can then turn towards two possibilities: the exploitation of semantic-syntactic relations and lexical relations.

Firstly, semantic-syntactic relations can be analyzed and exploited to express complex knowledge. We refer to relationships such as "typical object", "typical action" and "typical agent" (Lerat 1987 and 1988), and "typical application" (Gaudin 1996, 17). These relations list the typical collocations of a lexical unit, making it possible to provide a detailed description of the use made of the object named, or in some way, to substitute for an experience that we could have of the object in the real world.

This is particularly well suited to our ITinHeritage project, with metadata describing the artefacts of partners museums. Here is an example of the de-

scription of the artifact named "Tabulator BS 120" from the ACONIT Conservatory (Grenoble, France):

> La tabulatrice Bull BS 120 se compose d'un lecteur de cartes (traitant 150 cartes par minutes), d'une imprimante (cadence de 150 lignes par minute sur une largeur de 92 colonnes), d'une perforatrice de cartes (débitant soixante-quinze cartes par minute). Elle dispose d'un calculateur mécanique qui lui permet d'exécuter les quatre opérations arithmétiques, des opérations logiques et de mémoriser des informations. Elle offre un système de programmation amovible : le "tableau de connexion", qui était spécialement câblé pour chaque traitement. Sa technologie est entièrement électromécanique.

- Typical action: lire, additionner et imprimer
- Typical object: caractères
- Typical agent: calculateur mécanique, lecteur de cartes, imprimante, perforatrice de cartes (ateliers mécanographiques à cartes perforées)
- Typical application: calcul électromécanique

Semantic-syntactic relations are proving to be a valuable tool for translating technoscientific discourses and gradually make the work of textual analysis moving towards that of knowledge representation. Their interest lies in the contextualization of the object in its uses. This helps to gain understanding, even if it is still insufficient to acquire knowledge. The path towards the concept to be understood is thus supported by a description of the object's properties, echoing some characteristics of the concept. We will present below, another method of working with terminology along these lines.

However, we clearly see in this example that relying only on semantic-syntactic relationships is insufficient to access meaning in the face of the complexity and density of technoscientific language. A translation work related to the recipient's references remains necessary, but excessively burdensome if carried out manually, for instance, for the construction of definitions. We can take the example of the typical application "electromechanical calculation" which can be presented as a "precursor of computer programming", and provides the non-expert with the most commonly accepted reference of "computer programming". It is thus possible to use the inclusions of technoscientific languages in everyday language, even if not all fields are represented in the same proportions, because a specialized language is not to be understood as a whole with defined boundaries, and because it abounds in obstacles to the appropriation of knowledge. In our field of Information Technologies, which can serve as an example, we can observe complex terms (mechanographic workshops with punched cards), eponyms (Turing machine, Moore's law), acronyms complicated by version numbers (IBM 1130), etc.

However, contemporary technosciences are characterized by both a great technicality and a high degree of socialization. Their objects shape our daily li-

ves, especially in the realm of Information Technologies. Consequently, many of their terms sketch out certain notions within us, such as a "keyboard" or a "computer mouse". As demonstrated by Gaudin and Lerat, as well as based on empiricist philosophical theories (Aristotle, Condillac), it is through the application made around these objects, that is to say, the experience that humans have of these objects, that meaning can be constructed towards more complex notions and that names can be attached to these objects. The common language and the specialized language overlap and enrich each other in the wake of the adoption of technologies by the general public.

However, another significant obstacle lies in the linguistic diversities within the same field. These diversities can be geographical (diatopic), temporal according to technological developments (diachronic), or community-based (diastratic). Continuing with Information Technologies, a "first-generation calculator" for the amateur is referred to as a "mechanical calculator" by a historian of computing, or as an "antique calculator" by a general historian. One of the objects that this concept covers is the "abacus", as the "first digital tablet". It is obvious that, for the public of the 21st century, a "digital tablet" refers to a completely different object, i.e. a modern digital pad (laptop). But the object designated in the IT heritage dates back to Antiquity and consists of clay balls and tokens, from the romans "calculi". These abacuses were used for arithmetic until 7000 BC, and evolved into the instrument with rows of movable pieces, better known from 500 BC as the "abacus". It is also interesting to note that English does not distinguish the evolution from the abacus to the "boulier", which for French are two distinct objects (abaque/boulier). These linguistic differences within the same field reflect distinct representations of the world, which are themselves derived from different experiences of it: «... we only manipulate reality through the representations we have of it» (Roche 2005, 50).

In the field of IT, as in all technosciences, especially those that are socially established and characterized by rapid or abundant technological evolution, we can observe these linguistic divergences. Another example that we cite in our previous works (Djambian et al. 2024a) which can illustrate these divergent representations of the world within the same domain, due to different experiences of it, is that of the "Baby" as amateurs name it. The museums designate it as the "Manchester Baby" in reference to its place of construction, by the more comprehensive formulation of "Manchester Small-Scale Experimental Machine" (London Science Museum), because the "Small-Scale Experimental Machine" (SSEM). It was the world's first von Neumann architecture machine built at Manchester's Victoria University in 1948, known by the general public as the first "computer". But, according to historian of computing, mainstream "computers" are "stored-program computers", in the strict sense of the term, i.e. von Neumann-type computers, which have a very precise me-

aning: they are electronic calculating machines with a central memory large enough to hold the program being executed, as well as the data. This definition makes the EDVAC (Electronic Discrete Variable Automatic Computer) the first electronic stored program computer, developed by the Pennsylvania University for the United States Army. So, we can see the extent to which the variety of languages in use reflects different representations of the world and its objects, and even impacts the vision of the history of a field, and can reflect underlying political and economic hegemonies and socio-cultures.

These linguistic differences within the same field can only lead to a conceptual confusion. However, in the relationship between everyday language and specialized sub-languages (diastraties), words remain the sole foundation that allows us to access the concepts they designate, especially nowadays where access to knowledge primarily occurs through digital or, at least, formal mediation. Regarding IT heritage, we find designations of concepts in the texts composed of the metadata that describe the collections of partner museums and serve as the basis for our terminological work. Given the variety of languages used in the field, this choice is based on the fact that these texts are written by experts in the field, in a language that is already sufficiently controlled to achieve consensus and understanding by different audiences.

Through a linguistic analysis, we are able to extract phrases that can maintain lexical relationships. Generic or partitive lexical relationships play an essential role in the appropriation of knowledge and prove to be more effective for terminological work. According to ISO 704 standard, they form its foundation and allow for a more immediate positioning of the term within its conceptual context. They place it within a whole. Lexical relationships are the second opportunity provided for the integration of new terms and knowledge, and they are strongly represented in our corpus, as evidenced by the description presented above of the "Tabulator BS 120".

5. The semasiological approach underlying terminological work

Our previous works (Djambian et al. 2024a) have shown us the usefulness of an analysis of linguistic uses as a first step to identify the names of concepts and their meanings. It is therefore on the basis of a semasiological approach that we support our terminological work, starting from our Knowledge Graph which gathers the metadata of the partners museums' inventory collections. The construction of our corpus has followed the text selection criteria of ISO 5078 (ISO 2025). This standard has also guided our semi-automatic lexical extraction work, where human intervention has always complemented automatic extraction. To this end, we have combined statistical and linguistic techniques. Our work initially focused on the largest corpus, which is the French corpus of the ACONIT Conservatory, selecting the fields that are richest in

terms of texts: Description, Name, Model, and Use. To carry out the automatic lexical extraction, we turned to the TermoStat (Drouin 2003) and Sketch Engine (Kilgarriff et al. 2014) extraction tools. We relied on Python's Natural Language Toolkit (NLTK) too, composed of libraries and programs for Natural Language Processing (NLP), supporting classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities (Bird et al. 2009). Indeed, programming languages are now considered relevant alternatives to traditional extraction tools when facing corpora derived from big data (Anthony 2021).

Our analysis was first based on the TermoStat extraction tool. It has made it possible to highlight simple and complex terms among lexical categories such as nouns, adjectives and verbs. It should be noted that TermoStat bases its analysis on a method of contrast between specialized and non-specialized corpora. The terms we are looking for are the most specific to the field and therefore those which have a lower frequency in the general reference corpus. However, TermoStat has shown real limitations regarding the terms specific to technosciences such as IT: for example, it is impossible to extract acronyms or terms starting with a capital letter followed by a series of numbers. The extraction tools prove to be poorly suited for these languages. These cases, related in the field of IT, to machines' names and their models or manufacturers, are nonetheless of critical importance in the domain (e.g., Gamma 30, IBM 1130, ...). Another problem is that TermoStat understands the Gamma 3 electronic calculator, marketed by Bull in 1952, as a verb. So, this first phase of extractions done by Termostat, which we had hoped would be more successful with the help of Python's NLKT library, proved unsatisfactory. The level of noise required an excessively demanding manual intervention.

Finally, Sketch Engine is the tool that has demonstrated the best results. For the lexical extraction, we considered with particular attention the multi-words terms results (complex terms) such as "mémoire à tores magnétiques", which are the most frequent lexical pattern in IT and the most meaningful. The first single-word terms obtained, such as "micro-ordinateur", "disquette", "macintosh", "microprocesseur", "azerty", "hewlett-packard", "powerbook", "olivetti", and "alphanumérique" are mainly generic terms, often the ones most commonly used in everyday language, and therefore the most familiar to the public, but less significant in the field. They are not part of the specialized expert language, unlike the complex terms, acronyms or terms beginning with an uppercase letter followed by a series of numbers. For these, the Corpus Query Language (CQL) of the Concordance section of Sketch Engine was used because it allows the analysis in various possible contexts, extracts, and analyzes more complex grammatical patterns.

Some observations that we detailed in precedents works (Rossi et al. to be published), have been made based on this first linguistic analysis. The verbs,

which are very present in our corpus, often serve as markers of partitive relationships (Lefeuvre and Condamines 2017), such as "to possess", "to contain", "to include", and "to involve". They indicate components of machines or their technical functioning. The adjectives, for their part, mainly refer to the shape or function of the machines. If the Description field pertains to a museum mediation strategy centered on a detailed description of the object, the Usage field engages in a narrative mediation with a complex structure that, beyond the artifact, describes the practices or actors related to the history of the object. The verbs here are more often focused on the relationship between actions or processes: "to enable", "to be able", "to serve", "to act", and "to function".

We have worked on the same fields within the English language corpus formed by the metadata from the London Science Museum and the HNF for the purposes of unbiased comparative study. The lexical extraction work on the "Description" field shows, contrary to the French corpus, a focus on artefacts but also on their contextualization, a mediation strategy that we had observed in the "Use" field in French. This significant difference in mediations reflects a difference in the representations of the world, which is also observable in the Italian corpus from the Museo degli Strumenti per il Calcolo in Pisa. The language practices among these different cultures also reflect a distinct relationship to objects in their patrimonialization. Here, the Adj+Noun model is the most productive, with examples such as "personal computer", "video game", or "electronic calculator". The most frequent verbs focus on concrete actions: "to manufacture", "to build", "to work", although one can also find a significant number of occurrences for verbs such as "to include", which more marginally, indicate partitive relationships. At this point, the semasiological analysis allows us to highlight some interesting trends within our corpora: the prevalence of a morphological pattern for the creation of terms in the domain, as well as the presence of markers of conceptual relationships.

6. About the need to use socially and historically established consensual representations: the necessary completion by the onomasiological approach

However, the conclusion reached at the end of this semasiological work is that it highlights the extreme dependence of the lexicons extracted with the initial corpora, and it does not resolve, quite the contrary, the linguistic variety and complexity of the domain. It can, in no way, in its current state, lead to a consensual conceptualization. In addition, what should we do with these countless lexicons? Our previous experiences have shown us the difficulty to build up a semantic network based on this semasiological work (Djambian 2011). It poses the problem of finding only the variety of names of the do-

main's concepts in the texts and not the concepts. This study, important foundation for the maturation of terminological work, is limited to the analysis of words in their uses. The previous lexical analysis already shows the disparities that appear from one language to another, but more simply, from one museum collection manager (the metadata writer) to another.

As we said above, one cannot limit oneself to studying how a field is named without being interested in the conceptualizations that underlie these linguistic usages. To compensate for linguistic variations and bring out a common meaning, it is advisable to take an interest in the extralinguistic part of terminology work, centered on the relationship of the concept to the object, and not just in the relationship between a term and its signified, as in the linguistic study we presented in the semasiological phase. The meaning of a word must be independent of usage and consensually standardized within a community that refers to a conceptualization of the world. Thus, it must be remembered that all terminological work should be based on concepts and not on terms (Felber 1987). Terms serve as a point of entry for the terminologist, as a denomination of a concept (Roche 2005). They represent an expression of a conceptual system that, in turn, reflects the way we perceive the objects in our world. The purpose of terminological work is then to establish a consensual and standardized conceptual system.

The onomasiological approach allows for a more comprehensive inclusion of linguistic specificities and variations, as well as the representation of the conceptual system. This formalization stage, conducted with the assistance of field experts, is based on the convention in the Latin sense of *foedus*, which captures the full dimension of the societal and historical context in which a concept is born and evolves. As the concept is an eminently social construction and, to exist, must be validated within a given collective receptor: «The validity of solitary thought is principally dependent on the justification of linguistic statements within the effective community of argumentation. It is not possible for a single being to follow a rule and validate his thought within the framework of a "private language". Rather, thought is public in principle» (Apel 1987, 90).

But how the conventions of intercomprehension and the terminologization of words are founded? For Kripke names are socially elaborated too and are linked to the concept only by «our interaction with other speakers in the community, an interaction under which we are linked to the referent itself» (Kripke 1982, 82). For him, the definition of real-world objects involves substituting for «the idea of necessary and sufficient properties, the idea of a bundle of properties, only some of which must be satisfied in each particular case» (Kripke 1982, 116), in other words, a prototype or what Wittgenstein also calls a family resemblance (1961). The reference fixed by the act of naming is perpetuated through evolutions in time, by successive interactions in commu-

nication, bequeathing these properties which, through abstraction, become conceptual essential characteristics. Hilary Putnam (1990, 47) discusses reference as a social consensus extended to «the contribution of the environment».

For objects of the real world, the reference, which may be fixed arbitrarily and consensually by standards, is ephemeral and refers us to the role of the expert, since in the socially situated exchange of language, there is a distribution of roles and a "linguistic division of labour". Hilary Putnam argues that meaning does not emerge simply by being related to our pre-existing thoughts, but depends on our context and our interaction with it. Identification and validation within this socially established context are integral to the construction of an individual conceptual universe. Putting a concept into words is therefore fundamental because it links the concept to a set of variable linguistic usages, corresponding to practices and organizing a representation of the world, a logosphere, as «the mesh of meaning enveloping reality» (Lafont 1993, 292). It therefore establishes a direct link between language and the objects of the real world and encompasses individual linguistic varieties and consensual community norms.

While ontological construction follows onomasiological paths, the consideration and study of discourse (the semasiological approach) remains essential for understanding the processes by which concepts are constructed, constantly reformulated in discourse, in a historical dimension that texts make observable. «The history of concepts plays a part in understanding them» (Gaudin 1996, 609). The concept-discourse dynamic is, therefore, bilateral and iterative and is built in a given socio-cultural context. By studying discourse, it is essential to take account of this social and historical construction to retrace the genesis of the concept and carry out what Alain Rey (1990, 778) calls «the archaeology of conceptual constructs, ideological and scientific systems, or in an oversimplified term, the history of ideas».

Through a logical formalization, the ontological construction will reproduce this "logosphere" within which experts, who have a socio-linguistic responsibility in the transfer of knowledge, will be able to express prototypes bringing together bundles of essential characteristics that are sufficiently consensual to encompass the varieties socially established in the various communities and over time. The organization of concepts in the form of a notional system requires an inquiry into the essential characteristics that form the concepts. This formalization must designate the concepts beyond their usages to explicitly position them within the overall organization of the notional network. The ordering of concepts according to their hierarchical relationships and differences allows for the construction of a standardized vocabulary based on the generic or partitive lexical relations previously observed during the lexical extraction process.

For example: a “supercomputer”, whose synonym is “supercalculator”, from 1961 to now, has the essential characteristics of being designed to achieve the highest possible performance, particularly in terms of computing speed, and of being a mainframe computer. It is a kind of “second generation stored program computer”, characterized by the use of “transistors and batch systems”. The “supercomputer” has evolved to a “massively parallel stored program computer” from 1980, which characteristic is to use a large number of “computer processors” or separate “computers” to perform a set of coordinated calculations in parallel.

The dimension of hyperonymy (is a kind of) or meronymy (is composed of) is managed in the ontology by the subsumption relationship inherent in the Porphyry tree organization of classes (Porphyre 1984). The synonymy relationship is managed in class (concept) annotations by labels, using models such as SKOS (Simple Knowledge Organization System) (e.g. skos:altLabel). The semantic-syntactic relationships mentioned above can also be reproduced in a logical formalization as an ontology by integrating certain concepts (classes) and using description logic based on the subject-predicate-object triptych. The subject (concept or class) is linked to the object (instance) by a predicate (property or relation), which enables real sentences to be modeled (Fig. 2). For example, the object (instance) “punched card tabulator” is a “mechanographic machine”, which is a “calculation device” (subsumption relation). It was invented in 1887 (Date class) by Herman Hollerith (Inventor class), who founded the International Business Machines Corporation (IBM) (Constructor class), for the 1890 census (Use class) of the United States (Place class). It is composed of “60 counter dials”, “punched cards”, “a reading head” (Description class).

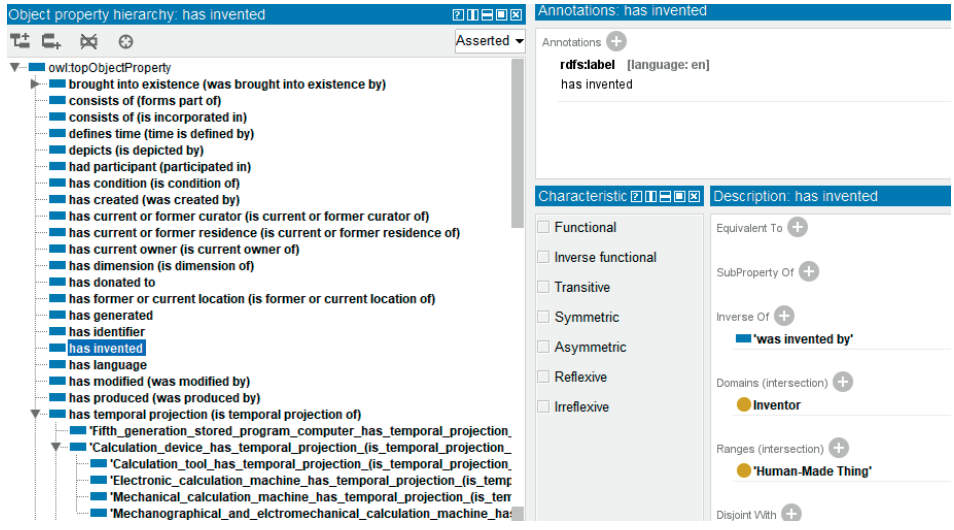


Figure 2: An example of the “has invited” property linking inventors to their inventions.

The construction of the IT domain's onto-terminology is facilitated by the use of, in a first time, the Protégé environment for the Knowledge Graph structuration, and in a second time, on the TEDI tool (Termino-ontological EDItor) (Roche and Papadopoulou 2019), specially designed to carry out terminological work according to the ISO 704 and 1087 standards. In particular, it allows to describe precisely the essential and distinctive characteristics of each concept, thus to define the ontology and analyze the conceptualization in the field, while generating at the same time, a dictionary, always built in accordance with the ISO standards mentioned above. In response to the study of the semantico-syntactic relations that we have described above, which are intended to contextualize a concept by highlighting the properties of an object that emerge from its uses, terminology work on TEDI enables us to take a direct interest in the essential and distinctive characteristics of the said concept by approaching them from various angles of analysis. This makes it possible to define an ontology and its concepts through different facets that cannot be represented in the Protégé environment. We can mention, for artefacts in the field of IT, the technical axis (mechanical, electromechanical, electronic), the theoretical axis (analog, logic, or digital), the axis by program (programmed, partially programmed, non-programmed), by architecture (Von Neaumann, Harvard, parallel, quantum, ...), etc. Thus, a Von Neumann-type architecture computer is an electronic calculating machine, both arithmetic and logical, with a program, in main memory during execution. This modeling of IT heritage knowledge complements the first-level ontology by integration via some CIDOC-CRM upper-ontology classes (Fig. 3).

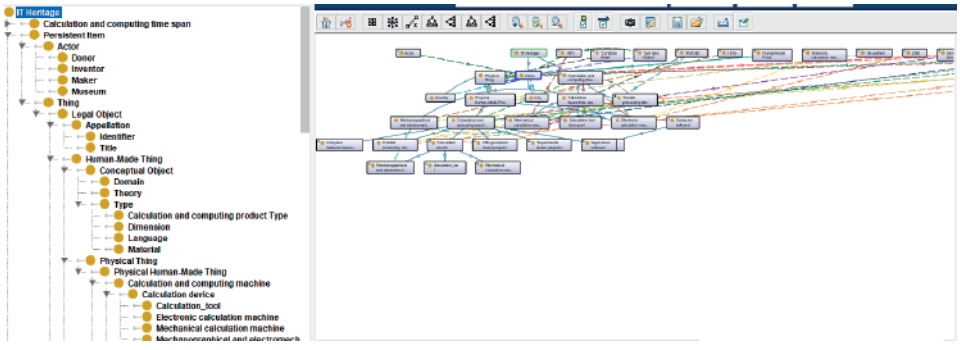


Figure 3: an overview of the ITinHeritage domain ontology.

We aligned on Protégé, the CIDOC-CRM upper-ontology to the OWLTime model (Pan and Hobbs 2005) to obtain a diachronic modelization representing the technological evolutions in the IT field. This way, the concept is located both in the global notional system and in his historical contextualiza-

tion, crucial to its understanding and appropriation as explained supra. As the social context is mandatory too, our future works will focus on practices across history with the help of the "Use" class, to situate these artefacts as socially founded technologies that shape and impact our society in reciprocal ways.

Our future developments will aim to improve the diachronic representation by using Allen intervals instead of OWLTime in the TEDI environment. We are therefore overcoming the fixed-in-time aspect, which is a major bias in ontologies, by making the ITinHeritage ontology diachronic. But domain ontologies also often have the drawback of being scientific works considered as an end in themselves. However, our aim here is to demonstrate their usefulness in overcoming the difficulties of managing large masses of museum data and even beyond. Our ontology must, therefore, be linked to the data to be managed, that is to say, to the more than 25500 artefacts that make up our knowledge graph. So, we need to populate our ontology, which means assigning instances (real-world objects) to our classes (concepts). Obviously, this is impossible to be done manually in such proportions.

As a first intention, we used statistics with TF-IDF (term frequency-inverse document frequency), which measures the importance of a word to a text adjusted for the fact that some words appear more frequently in general, to automatically populate our ontology, i.e. to integrate the museum artefacts as instances of the classes (concepts) in the ontology. Here, again, lexical extraction techniques using the Sketch engine tool were necessary. It aimed to produce a lexicon for each concept in each language (FR, EN, IT, DE), using, this time, a wider corpus composed of web sources as Wikipedia and open access specialized websites or revues. This work still requires extensive human intervention with uncertain results, as TF-IDF is still a relatively old technique. And our representation of domain knowledge remains static.

We need to automate our ontology populating so that, via the synchronization our knowledge graph with the museum data (see above), we can pass on any new entries in the collections. Moreover, our ontology has to be evolutionary because IT cannot be represented in a fixed way when it is continually being created. This is why we are using Artificial Intelligence (AI) via Large Language Models (LLMs), in this case Llamma3, to automatically extend the manually constructed onto-terminology. The lexicons produced previously (we have only used the English language here) were used to guide the AI's results. To train it, we had to provide it with descriptions representative of the domain's terminology, corresponding to each concept. The AI's contextualization and learning capabilities enable it to offer a classification that is refined over time. But the upstream human work required to obtain an operational AI and ontology is still very heavy in terms of lexicon construction, annotation of data corpora, and learning, and it requires the intervention of experts in the field. However, while these problems are not trivial, they are generic

to the use of AI and are in the process of being improved. Our contributions here, therefore, concern the application of automatic methods for extending manually constructed ontologies in order to keep pace with technical, terminological and conceptual developments in the field. Our terminology work is being supported by the LLM to analyze textual fields in the Knowledge Graph corpus, enriched with web data sets. The aim is to be able to detect new terms relating to new objects and integrate them into our ontology, albeit still under human validation.

Finally, since the aim of all this work is to ensure that the general public has access to knowledge in the IT field, we will be making access to the Knowledge Graph possible via a web platform (www.itinheritage.com). Other portals federate museum collections (Szekely et al. 2013; De Boer et al. 2013; Doerr et al. 2010) or safeguard software (Di Cosmo 2018), but the aim of these projects is to enhance the value and heritage of artefacts, whereas our aim is to transfer knowledge. The web platform has therefore been designed with an end-user approach and offers navigation via a time map to represent the technological evolution across history (time and space representation); a visual graph based upon the domain ontology to provide input to the knowledge graph via concepts representing the domain knowledge; and a natural language querying. This man-machine interface is essential because querying the Knowledge Graph (which is in RDF) in natural language requires SPARQL (SPARQL Protocol and RDF Query Language) queries, a computer language that users in the general public cannot master. SparNatural (Clavaud and Francard 2022) is a well-known open-source tool available for querying by non-experts, which can be used to navigate our knowledge graph based on our domain ontology. This approach significantly lowers the barrier to entry for users unfamiliar with technical query languages, making the data within our Knowledge Graph accessible to a broader audience. In this way, we try to increase the possibilities for transferring knowledge of the IT field to the widest possible audience, from scientists to the general public. Firstly, by opening up and then linking data on the semantic web, and secondly by making this data intelligible to introduce specialized concepts into culture. Our future work aims to improve these ontology extension and knowledge access technologies, using a Knowledge RAG (Retrieval Augmented Generation) (Lewis et al. 2020), another technology that also combines AI and Natural Language Processing (NLP), which will enable natural language querying in a much more intuitive and efficient way, and multilingual issues to be taken into account automatically, freeing up a significant amount of human intervention in lexicon generation.

Conclusion

For Smith (2011) heritage is defined in the tension between the objects, languages and representations that form it. In this article, we have shown that this is indeed the case for Information Technologies (IT), which are gradually being elevated to the rank of heritage beyond the everyday uses they represent for us. The uses of its languages fluctuate in step with technological developments and the conceptualizations that are created by the practices around these new technologies. The aim of our work is to highlight and analyze these parallel developments. In this continuity, our future works will continue, through the linguistic approach, analyzing the diachronies and diastracies in IT heritage languages, in a greater depth, for example, by using R language tools on larger comparative corpora. We will combine and enhance too, the terminological work with Artificial Intelligence (AI), to determine its contribution to the modelling of the extra-linguistic layer, always following these developments. By digitally perpetuating IT heritage in both its tangible and intangible forms, and opening it up, we hope that our work will help the general public to gain a better understanding of this prolific and highly technical field. From a scientific point of view, our contributions aim to offer a methodology for the cultural world and beyond for: managing massive and heterogeneous data, opening and segmentizing it (Linked Open Data (LOD)); creating onto-terminologies of scientific and technical domains; and using ultimate AI tools to complement human terminology work. Finally, from an epistemological perspective, the ITinHeritage research project is developing this work to bring to light the new knowledge and representations of the world brought by IT, which are, far beyond a technoscience, the reflection of our societal mutation.

Acknowledgments

This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003).

References

Anthony, Laurence. 2021. “Programming for corpus linguistics.” In A practical handbook of corpus linguistics, edited by M. Paquot, and S. T. Gries. Springer. https://doi.org/10.1007/978-3-030-46216-1_9.

Apel, Karl Otto. 1987. *L’éthique à l’âge de la science*. Presses Universitaires de Lille.

Aristotle. 1986. *Métaphysique*. Librairie Philosophique J. Vrin.

Bekiari, Chryssoula, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, and Athanasios Velios. 2021. "CIDOC CRM." International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). Version 7.1.1.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly Media.

Clavaud, Florence, et Thomas Francart. 2022. "Sparnatural, un éditeur graphique souple et intuitif pour explorer des graphes de connaissances." *Colloque Humanistica 2022*.

De Boer, Victor, Jan Wielemaker, Judith van Gent, et al. 2013. "Amsterdam museum linked open data." *Semantic Web* 4: 237-43. https://doi.org/10.3233/SW-2012-0074.

Di Cosmo, Roberto. 2018. "Software heritage: why and how we collect, preserve and share all the software source code." *IEEE/ACM 40th International Conference on Software Engineering in Society (ICSE-SEIS), Gothenburg, Sweden.*

Diels, Hermann. 1903. *Die fragmente der Vorsokratiker griechisch und deutsch.* Weidmann.

Dijkshoorn, Chris, Lizzy Jongma, Lora Aroyo, et al. 2018. "The Rijksmuseum collection as Linked Data." *Semantic web* 9 (2): 221-30. https://doi.org/10.3233/SW-170257.

Djambian, Caroline, Giada D'Ippolito, et Micaela Rossi. 2024a. "Des représentations du monde des Technologies de l'Information : un travail terminologique en patrimoines scientifiques et techniques." In *Proceedings of TOTh 2024-Terminologie & Ontologie: Théories et applications, June 2024, Le Bourget du Lac, France*, sous la direction de by Christophe Roche. Presses Universitaires Savoie Mont Blanc. https://hal.science/hal-04840125v1.

Djambian, Caroline, Micaela Rossi, Giada D'Ippolito, et al. 2024b. "New terminological approaches for new heritages and corpora: the ITinHeritage project". In *Proceedings of the 3rd international conference on Multilingual digital terminology today. Design, representation formats and management systems - MDTT 2024*, edited by Federica Vezzani, Giorgio Maria Di Nunzio, Beatriz Sánchez Cárdenas, et al. https://ceur-ws.org/Vol-3703/.

Djambian, Caroline. 2011. "Le métier : son savoir, son parler." In *Actes de la Conférence TOTh 2011-Terminologie & Ontologie: Théories et applications*, *May 2011, Annecy*, sous la direction de Christophe Roche. Presses Universitaires Savoie Mont Blanc. https://hal.science/hal-00805596v1.

Doerr, Martin, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, Carlo Meghini, and Herbert Van de Sompel. 2010. "The Europeana Data Model (EDM)." In *Proceedings of IFLA 76.*

Drouin, Patrick. 2003. "Term extraction using non-technical corpora as a point of leverage." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 9: 99-115. https://doi.org/10.1075/term.9.1.06dro.

Felber, Helmut. 1987. *Terminology Manual.* UNESCO.

Gaudin, François. 1996. "Dire les sciences et décrire les sens : entre vulgarisation et lexicographie, le cas des dictionnaires de sciences." *TTR: traduction, terminologie, redaction* 8: 11-27.

Gehrke, Stefanie, Eduard Frunzeanu, Pauline Charbonnier, and Marie Muffat. 2015. "Biblissima's Prototype on Medieval Manuscript Illuminations and their Context." In *Proceedings of International Workshop of Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, Portorož, Slovenia, June 1st, 2015*, edited by Arnaud Zucker, Isabelle Draelants, Catherine Faron-Zucker and Alexandre Monnin, CEUR Workshop Proceedings, 43-48, CEUR-WS.org. http://ceur-ws.org/Vol-1364/paper5.pdf.

ISO. 2025. *ISO 5078:2025 Management of terminology resources — Terminology extraction*.

ISO. 2019. *ISO 1087-1. Travaux terminologique – Vocabulaire – Partie 1 : Théorie et application*.

ISO. 2009. *ISO 704 NF ISO 704. Travail terminologique – Principes et méthodes*.

Jacob, Christian. 2001. "Rassembler la mémoire." *Diogène* 4: 53-76.

Juanals, Brigitte, et Jean-Luc Minel. 2016. "La construction d'un espace patrimonial partagé dans le Web de données ouvert." *Communication* 34 (1). https://doi.org/10.4000/communication.6650.

Kant, Immanuel. 1997. *Critique de la raison pure.* Aubier.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, et al. 2014. "The Sketch Engine: Ten years on." *Lexicography* 1: 7-36.

Kripke, Saul. 1982. *La logique des noms propres*. Les Editions de Minuit.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 3: 9459-74. https://doi.org/10.48550/arXiv.2005.11401.

Lafont, Robert. 1993. *Le dire et le faire*. Praxiling.

Lefeuvre, Luce, et Anne Condamines. 2017. "MAR-REL : une base de marqueurs de relations conceptuelles pour la détection de Contextes Riches en Connaissances (MAR-REL : a conceptual relation markers database for Knowledge-Rich Contexts extraction)." In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 - Articles courts, Orléans, France. ATALA*.

Lerat, Pierre. 1987. "Le traitement des emprunts en terminographie et en néographie". *Cahiers de lexicologie* 50: 137-44.

Lerat, Pierre. 1988. "Terminologie et sémantique descriptive." *La banque des mots* 36.

Pan, Feng, and Jerry R. Hobbs. 2005. "Temporal aggregates in owl-time." In *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2005).*

Porphyre. 1984. *Isagoge*. Librairie Philosophique J. Vrin.

Putnam, Hilary. 1984. *Raison, vérité et histoire*. Editions de Minuit.

Putnam, Hilary. 1990. *Représentation et réalité*. Gallimard.

Rey, Alain. 1990. "Lexico-logiques, discours, lexiques et terminologies 'philosophiques'." *Encyclopédie philosophique universelle. T. II. Les notions.* Presses Universitaires de France.

Roche, Christophe. 2007. "Le terme et le concept : fondements d'une ontoterminologie." In *Actes de la Conférence TOTh 2007-Terminologie & Ontologie: Théories et applications*, *Annecy 1er juin 2007*, sous la direction de Christophe Roche. Presses Universitaires Savoie Mont Blanc.

Roche, Christophe, and Maria Papadopoulou. 2019. "Mind the Gap: Ontology Authoring for Humanists." In *Proceedings of JOWO: The Joint Ontology Workshops*, edited by A. Barton, S. Seppälä and D. Porello. https://ceur-ws.org/Vol-2518/.

Roche, Christophe. 2005. "Terminologie et ontologie." *Langages* 1: 48-62.

Rossi, Micaela, Caroline Djamban, Cécile Frérot, Giada D'Ippolito, Emrick Poncet. En cours de publication. "Rôle et apport des verbes pour la construction et la modélisation de connaissances spécialisées du patrimoine muséal." In *Actes du Colloque Le statut du verbe terminologique entre théories et pratiques*.

Russell, Bertrand. 1956. *Logic and Knowledge*. Allen and Unwin.

Simon, Agnès, Sébastien Peyrard, Vincent Michel, and Adrien Di Mascio. 2014. "We grew up together: data.bnf.fr from the BnF and Logilab perspectives." In *Proceedings of IFLA 2014*.

Smith, Laurajane. 2011. "Heritage and its Intangibility." In *De l'immatérialité du patrimoine culturel*, edited by A. Skounti and Ouidad Tebbaa. Bureau régional de l'Unesco de Rabat.

Szekely, Pedro, Craig A. Knoblock, Fengyu Yang, et al. 2013. "Connecting the Smithsonian American Art Museum to the Linked Data Cloud." In *The Semantic Web: Semantics and Big Data. ESWC 2013,* edited by P. Cimiano, O. Corcho, V. Presutti, L. Hollink, S. Rudolph. *Lecture Notes in Computer Science* 7882. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38288-8_40.

UNESCO. 2009. *Guide to measuring Information and Communication Technologies (ICT) in education*. UNESCO. https://uis.unesco.org/sites/default/files/documents/guide-to-measuring-information-and-communication-technologies-ict-in-education-en_0.pdf.

Wittgenstein, Ludwig. 1961. *Investigations philosophiques*. Gallimard.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-549-3
DOI 10.57574/59655493XX
pag. 45-62 (gennaio-giugno 2025)
Open access article licensed under CC-BY-NC-SA

I Knowledge Organization Systems come strumenti di sorveglianza sanitaria

L'applicazione della ICD in Italia

Erika Pasceri*

Abstract: E-health refers to the use of information and communication technologies (ICT) to enhance the efficiency, quality of care, and access to healthcare services. The adoption of artificial intelligence, telemedicine, and IoT (Internet of Things) devices has enabled a more personalized and preventive approach to health. However, significant challenges remain in terms of privacy, data security, and standardization of information. Knowledge organization systems, such as classifications and controlled vocabularies, are crucial for facilitating interoperability between different healthcare settings. These systems enable effective data collection, management, and sharing, supporting public health surveillance. A key example is the International Classification of Diseases (ICD), which, although fundamental for clinical data analysis, requires continuous updates to meet the evolving needs of the healthcare domain.

Keywords: Health Surveillance, International Classification of Diseases, Knowledge Organization Systems, Knowledge Management, Interoperability.

1. Introduzione

L'Organizzazione Mondiale della Sanità (OMS) definisce l'e-Health come «the cost-effective and secure use of information and communications technologies in support of health and health-related fields, including healthcare services, health surveillance, health literature, and health education, knowledge and research» (World Health Organization. Eastern Mediterranian Region 2025). La sanità digitale rappresenta un insieme di attività e processi caratterizzati da una forte interdisciplinarietà che coinvolge le discipline mediche, archivistiche, documentali ed informatiche. Tali processi, infatti, mirano ad integrare le tecnologie digitali all'interno di strutture sanitarie, ospedali e ambulatori, al fine di migliorare l'erogazione e la qualità delle cure, ottimizzare le risorse e garantire un accesso più rapido ed efficiente ai servizi sanitari. La telemedicina, le cartelle cliniche elettroniche, l'intelligenza artificiale, la gestione

* Dipartimento di Culture, Educazione e Società, Università della Calabria, Rende (CS), Italia. erika.pasceri@unical.it. ORCID: 0000-0003-4440-2266.

elettronica dei dati sanitari e l'Internet of Medical Things sono solo alcune delle innovazioni che stanno contribuendo ad una vera e propria rivoluzione dell'ICT offrendo nuove possibilità diagnostiche e terapeutiche.

La digitalizzazione e ottimizzazione dei processi e dei flussi informativi all'interno delle strutture sanitarie rendono possibili approcci sempre più personalizzati e preventivi, grazie alla raccolta e analisi di grandi quantità di dati strutturati secondo standard di settore. Tuttavia, tale paradigma presenta anche sfide significative, tra cui la protezione della privacy dei pazienti, la gestione della sicurezza informatica e la standardizzazione dei dati per una corretta raccolta e successiva elaborazione. Tra questi fattori, particolare rilevanza assumono gli standard e i sistemi di organizzazione della conoscenza capaci di facilitare la comunicazione tra stakeholder: provider di servizi sanitari, agenzie governative e pazienti utilizzando un linguaggio che possa superare l'utilizzo di sistemi di codifica locali o barriere linguistiche.

Nel corso degli ultimi decenni, le organizzazioni sanitarie hanno subìto profonde trasformazioni attribuibili a diversi fattori: cambiamenti demografici, modifiche normative, nuovi contesti epidemiologici. Tali cambiamenti – anche sostanziali – hanno sottolineato l'urgente necessità di ottimizzare l'efficacia, l'appropriatezza e la sostenibilità dei registri informativi, come strumento alla base di un'organizzazione lavorativa più efficiente, che possa tradursi in una maggiore accuratezza del percorso di cure destinato al paziente. Le tecnologie dell'informazione e della comunicazione hanno contribuito notevolmente all'evoluzione di metodi e tecniche nella gestione documentale, agevolando lo scambio di informazioni e garantendone una più facile interpretazione e reperibilità.

I Knowledge Organization Systems (KOS) in questo contesto risultano essere strumenti in grado di svolgere un ruolo cruciale ai fini di gestione, rielaborazione e condivisione delle informazioni. Maggiore è il grado di strutturazione delle informazioni, maggiore è la possibilità di gestirle e conservarle ai fini di un riuso e recupero puntuale.

La sorveglianza sanitaria riguarda l'osservazione continua dei fattori che influenzano la salute della popolazione, come malattie infettive, malattie croniche, infortuni, e altri eventi sanitari. Il rapporto tra i KOS e la sorveglianza sanitaria può essere declinato in modi differenti, soprattutto in relazione al loro grado di strutturazione e la finalità per cui un KOS è stato creato. Tra i KOS, infatti, si annoverano ontologie, tassonomie, classificazioni, vocabolari controllati e liste di controllo (Zeng 2008) secondo livelli differenti di strutturazione favorendo quindi la gestione, l'interoperabilità e la condivisione delle informazioni in base alle necessità informative e comunicative del contesto di destinazione. Alcuni KOS nascono per essere volutamente generalisti, altri specialistici soprattutto in domini definiti con degli obiettivi e contesti di applicazione ben precisi.

La sorveglianza sanitaria, inoltre, implica la raccolta di dati da diversi sistemi, come ospedali, cliniche, laboratori ed enti governativi. I KOS contribuiscono a favorire l'interoperabilità tra i diversi sistemi, assicurando che i dati possano essere condivisi e confrontati in modo coerente (Blouin Genest 2015). Sebbene da tempo ormai i KOS siano entrati di diritto nella normativa che regola i flussi informativi, la codifica dei dati, la strutturazione dei documenti, in particolar modo nell'ambito sanitario, il loro utilizzo risulta essere molto parziale in confronto alle potenzialità di strumenti che – se compiutamente utilizzati fin dalla creazione del dato – costituiscono un supporto fondamentale di sorveglianza sanitaria, strumento essenziale di prevenzione della salute della popolazione. Lo scarso utilizzo a volte deriva dalla non obbligatorietà dell'elemento "codifica" già a partire dalle linee guida definite per l'implementazione dei documenti nei sistemi informativi[1].

L'analisi qui presentata è partita da un'osservazione del contesto internazionale, focalizzandosi sulle piattaforme per l'interoperabilità dei dati e dei documenti nel settore sanitario. In particolare, è stato esaminato come gli standard internazionali, come HL7 (Health Level Seven), IHE (Integrating the Healthcare Enterprise), OpenEHR, e protocolli come FHIR (Fast Healthcare Interoperability Resources), siano adottati a livello globale per garantire la compatibilità tra sistemi sanitari diversi e per promuovere l'efficace scambio di dati sanitari. Queste tecnologie, infatti, costituiscono la base per le iniziative internazionali di interoperabilità, come quelle promosse dall'OMS e dalla Commissione Europea, che definiscono linee guida per l'adozione di standard condivisi. Tale approccio ha consentito di identificare le best practices applicabili anche in contesti nazionali. Successivamente, l'analisi si è focalizzata sul contesto italiano, prendendo ad esempio un caso d'uso specifico nel Sistema Sanitario Nazionale (SSN) italiano, con particolare attenzione all'impiego della International Classification of Diseases (ICD) nei principali utilizzati quali le schede di dimissione ospedaliera (SDO), i certificati di morte, le cartelle cliniche e il profilo sanitario sintetico. Per un'analisi più attenta e contestualizzata è stato quindi necessario ripercorrere l'iter normativo che ha portato all'utilizzo obbligatorio di ICD e come questo sia stato declinato nei diversi contesti applicativi. In sintesi, l'approccio metodologico ha permesso di comprendere come i KOS – se compiutamente utilizzati – possano essere strumenti fondamentali per ottimizzare la gestione dei dati sanitari, garantire l'efficace scambio di informazioni e migliorare i processi di sorveglianza sanitaria, pur nel rispetto delle normative vigenti in ambito europeo e nazionale.

---

[1] A tale proposito si vedano le linee guida per l'implementazione in formato elettronico dei documenti pubblicati da HL7 Italia (HL7 Italia n.d.), che è parte di HL7 International ed è responsabile della localizzazione dello standard nella realtà italiana e, più in generale, ha l'obiettivo di stimolare e convogliare i contributi regionali e nazionali allo sviluppo dello standard e favorire la modernizzazione dello scenario sanitario italiano.

2. Knowledge Organization, Knowledge Management e le organizzazioni sanitarie

Organizzare la conoscenza è una capacità innata dell'essere umano che gli permette, partendo dalle informazioni già conosciute, di strutturare in modo efficace i dati per generare nuove conoscenze e comprendere il mondo in modo più completo e articolato. In ogni campo del sapere sono nate diverse modalità di organizzazione della conoscenza, che hanno rappresentato un fondamento cruciale per le ricerche e gli sviluppi successivi. Nell'ambito delle scienze della biblioteca e dell'informazione, la cosiddetta *Knowledge Organization* è un campo di ricerca che indaga tutte le tecniche (descrizione, indicizzazione, classificazione, ecc.) con cui la conoscenza può essere ordinata in modi utili per la sua consultazione ed il suo utilizzo (Gnoli et al. 2006).

Il campo di ricerca, apprendimento e pratica prevede l'organizzazione e la rappresentazione dei documenti nei diversi sistemi informativi (banche dati, sistemi di classificazione, cataloghi di biblioteche, Internet, biblioteche, archivi, ecc.), e include attività di descrizione, archiviazione e classificazione di , soggetti e concetti, sia da parte degli esseri umani che da parte delle applicazioni informatiche (Hjørland 2016).

Nel contesto digitale, indubbiamente favoriti dalle nuove tecnologie dell'informazione e della comunicazione, i KOS acquisiscono una forma ancora più ampia in qualità di risorse capaci di intermediare tra l'uomo e la macchina e tra persone che utilizzano codici comunicativi differenti, garantendo così il recupero e lo scambio informativo coerente e non ambiguo grazie ad una precisa strutturazione dell'informazione.

Con lo svilupparsi della teoria del *Knowledge Management* (Nonaka e Takeuchi 1995), nasce l'idea base che gestire la conoscenza mediante un processo sistematico volto non solo a creare nuovi tipi di conoscenze, ma anche a gestirle, rielaborarle, condividerle, sia un requisito fondamentale per mantenere la competitività delle organizzazioni. L'evoluzione tecnologica e la disponibilità di strumenti informatici pervasivi e sempre più intelligenti hanno avuto un impatto dirompente all'interno dei processi delle moderne organizzazioni, le quali necessitano di gestire consapevolmente il proprio patrimonio informativo, che nel tempo ha acquisito forme senz'altro diverse dal passato. È un patrimonio fondamentale fin dal momento in cui ha origine; utile non solo per gli sviluppi futuri dell'organizzazione e per il valore storico che assumerà nel tempo, ma altresì fondamentale per la memoria recente dell'organizzazione e per garantire un'immediata e maggiore efficienza. Considerata l'imponente proliferazione di dati e documenti la gestione documentale è ormai il *core business* delle organizzazioni in epoca moderna, settore tanto complesso quanto strategico, che si avvale dell'interazione tra diverse discipline, quali l'informatica, l'archivistica e la documentazione.

Tuttavia, nonostante l'innovazione tecnologica abbia investito in maniera prepotente ed importante anche il settore sanitario, è necessario – ancor prima di adottare soluzioni informatiche estremamente innovative e promettenti – porre maggiore attenzione alla qualità del dato in sé, al suo valore: come viene prodotto, gestito e trasformato in conoscenza utilizzabile e condivisibile. Di conseguenza le cartelle cliniche, i referti, le lettere di dimissione ospedaliera, i verbali di pronto soccorso, i referti di laboratorio, le prescrizioni mediche, e tanto altro, costituiscono una fonte inestimabile di conoscenza, utilizzabili in molteplici contesti e con diverse finalità. Affinché dati e documenti prodotti diventino "risorse" all'interno delle organizzazioni moderne, specie quelle sanitarie, è necessario tuttavia un processo sistematico di gestione e organizzazione della conoscenza, volto a trasformare la conoscenza "statica" in conoscenza "dinamica", e dunque significativa e rielaborabile con il supporto di strumenti e applicativi interoperabili.

Le piattaforme digitali, strumenti di data mining e intelligenza artificiale (IA) aiutano a identificare pattern e tendenze in grandi volumi di dati. Per esempio, l'analisi predittiva, un ramo dell'IA, permette di prevedere l'insorgenza di epidemie e di monitorare la diffusione di malattie infettive in tempo reale, favorendo la prevenzione e risposta rapida (Sharma et al. 2024). L'uso di tecniche di machine learning e deep learning consente di analizzare enormi quantità di dati da diverse fonti, identificando correlazioni che potrebbero sfuggire all'analisi manuale, come nel caso di nuove malattie o fattori di rischio emergenti. Questo tipo di analisi è essenziale per la sorveglianza delle malattie croniche, la gestione dei dati sui vaccini e per il monitoraggio delle emergenze sanitarie, come nel caso di epidemie o pandemie (Hanna et al. 2025).

3. L'eHealth in Europa: standard e interoperabilità dei dati sanitari

Nel contesto europeo, lo sviluppo di servizi e strumenti di *digital health* o *e-health* è divenuto uno dei principali obiettivi dei vari paesi, e ha raggiunto nel tempo livelli di maturità più o meno alti, sulla base di fattori differenti che hanno influenzato la sua naturale evoluzione. È d'obbligo considerare l'aspetto dei costi che questo processo comporta. Questo impegno si è concretizzato attraverso una serie di iniziative, politiche e sociali, con l'obiettivo di migliorare l'interoperabilità tra i sistemi sanitari dei vari paesi membri, consentendo così una gestione migliore delle informazioni sanitarie a livello transfrontaliero. Una fra tutte è l'introduzione dell'eHealth Digital Service Infrastructure (eHDSI) (Unione Europea 2025)[2], un'iniziativa promossa dall'Unione Euro-

[2] La Commissione Europea mette a disposizione di tutti i cittadini dell'EU una pagina dedicata all'infrastruttura di servizi digitali per l'assistenza sanitaria online al fine di garantire la continuità delle cure mediche ai cittadini europei che viaggiano all'interno dell'UE. Per tutti i cittadini è possibile riconoscere la disponibilità di tali servizi evidenziati dal marchio "MyHealth@EU".

pea nell'ambito del programma *Connecting Europe Facility (CEF) (Unione Europea 2021)*, con l'obiettivo di creare un'infrastruttura digitale comune per la sanità, facilitando la condivisione sicura dei dati sanitari tra gli Stati membri. Alla base del funzionamento si evidenzia l'implementazione di strumenti quali l'*e-precription*[3] e il *Patient Summary*[4].

Gli standard di dati, specie in ambito sanitario, vengono spesso implementati a livello regionale o locale, mettendo potenzialmente a rischio l'interoperabilità, causata dall'utilizzo di versioni obsolete o non allineate a quelle di altri sistemi informativi. Tanto vale sia per le istituzioni pubbliche che per quelle private. L'acquisizione di dati 'di buona qualità' rappresenta la fase più importante e delicata, in quanto pone le basi per le successive analisi o implementazioni. Nel contesto delle iniziative volte a favorire non solo l'interoperabilità, ma anche tutto ciò che ruota intorno alla sanità digitale, il Regolamento (Unione Europea 2017), sui dispositivi medici (Medical Device Regulation, "MDR") ha avuto un impatto indiretto sulla digitalizzazione, considerando che richiede l'uso di tecnologie per il monitoraggio dei dispositivi medici, inclusi software e applicazioni. Nel 2022 si è dato il via al progetto volto a creare l'European Health Data Space (EHDS) (Unione Europea 2025), che mira a costruire un "spazio comune" per i dati sanitari. L'idea di base è quella di garantire la condivisione secondo protocolli di sicurezza tra i Paesi membri al fine di migliorare le iniziative di ricerca, la diagnosi, la cura e il monitoraggio costante dei pazienti. Tuttavia, raggiungere l'interoperabilità per un'assistenza sanitaria continuativa all'interno e tra gli Stati membri dell'UE rimane una sfida. Se da un lato l'EHDS nasce con l'intento di mettere a fattor comune i progressi tecnologici, etici e politici di oltre due decenni e di aggiungere una nuova componente all'erogazione dell'assistenza sanitaria ospedaliera e alla ricerca clinica, per poter sfruttare appieno il potenziale dell'EHDS, oltre a stabilire relazioni tra le diverse organizzazioni, è necessario un certo grado di maturità digitale negli ospedali, in particolare nei settori dove la sicurezza informatica è predominante per la gestione dei dati e per l'interoperabilità (Deimel et al. 2025).

In tale contesto l'uso dei KOS non solo a livello internazionale, ma soprattutto a livello locale per ciascun Paese membro, risulta essere indispensabile poiché rappresenta il modo con cui i dati clinici possono essere compresi e utilizzati in modo coerente tra sistemi sanitari nello stesso Paese e tra Paesi diversi assegnando loro un significato univoco e condiviso. A tale proposito l'EHDSI ha reso disponibile un Master Value Sets Catalogue (MVC), che contiene l'e-

[3] Permette a un cittadino europeo di ottenere farmaci prescritti nel proprio Paese anche in un altro Stato membro, dove la prescrizione viene riconosciuta e gestita digitalmente.

[4] Il Patient Summary o Profilo Sanitario Sintetico è costituito da un insieme identificabile di dati sanitari essenziali e comprensibili che include gli eventi clinici necessari a garantire la continuità assistenziale.

lenco dei concetti clinici ammessi e i relativi codici nei KOS approvati[5]. Senza il catalogo dei Value Sets, la codifica clinica nei documenti eHDSI resterebbe arbitraria e non vincolata a concetti condivisi, vanificando l'obiettivo finale di raggiungere l'interoperabilità sul piano semantico. Il MVC garantisce la coerenza, l'aggiornamento e la tracciabilità, rendendo possibile l'elaborazione sicura e affidabile dei dati sanitari transfrontalieri.

La mancanza di riferimento o mappatura verso un KOS standard impedirebbe qualunque automatismo nella lettura dei dati clinici e comprometterebbe la sicurezza e l'efficacia delle cure: un termine 'libero' o non codificato può essere ignorato, frainteso o tradotto in modo errato. Al contrario, la presenza di concetti formalizzati consente la validazione, il controllo, la tracciabilità e l'aggiornamento coerente delle informazioni. In sintesi, i KOS forniscono alla eHDSI la semantica necessaria affinché gli standard sintattici (HL7 Clinical Document Architecture, FHIR) siano effettivamente utili in un contesto clinico transfrontaliero.

I sistemi di organizzazione della conoscenza in ambito sanitario sono in grado di integrare fonti di dati eterogenee, sia quantitative che qualitative. Questo include dati clinici (ad esempio, referti medici, cartelle cliniche elettroniche), dati epidemiologici (come i tassi di incidenza e prevalenza di malattie), dati demografici (come età, sesso, etnia e fattori sociali), e dati ambientali (come inquinamento o condizioni climatiche) che possono influenzare la salute. La capacità di combinare e analizzare queste informazioni in modo coerente e integrato è essenziale per monitorare la salute della popolazione e rispondere in modo efficace alle emergenze sanitarie. Ad esempio, l'analisi integrata di dati epidemiologici e sociali può rivelare pattern di vulnerabilità in particolari gruppi di popolazione (ad esempio, in base a fattori economici, razziali o geografici) (Buchanan 2003), permettendo così di sviluppare politiche mirate di prevenzione e intervento (Choi et al. 2024)which include data collection, analysis, interpretation and dissemination, but not public health action. Controlling a public health problem of concern requires a public health response that goes beyond information dissemination. It is undesirable to have public health divided into data generation processes (public health surveillance. Allo stesso modo, l'integrazione di dati provenienti da laboratori clinici e da sistemi di sorveglianza ambientale può contribuire a identificare focolai di malattie infettive legati a specifici fattori ambientali (Semenza 2014)present, and future. The key finding from the Working Group II, Fifth Assessment Report (AR5. Per poter essere utilizzate in maniera ottimale le strutture dell'informazione clinica sono passate da modelli narrativi e destrutturati (basati su testo

[5] Il catalogo include oltre 200 value sets, documentati con diversi metadati che includono il nome, l'OID, la lingua ecc. Tra i principali value sets troviamo SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms), UCUM (Unified Code for Units of Measure), ATC (Anatomical Therapeutic Chemical Classification System), LOINC (Logical Observation Identifiers Names and Codes).

libero nei documenti cartacei o digitalizzati) a sistemi sempre più strutturati e codificati, in grado di rappresentare formalmente i concetti medici per l'elaborazione automatica, la condivisione e rielaborazione. Questa evoluzione è stata necessaria per sostenere l'informatizzazione dei processi clinici e l'interoperabilità nei contesti nazionali e internazionali. In questo passaggio, i modelli informativi (es. HL7 CDA, openEHR, FHIR) e i KOS quali SNOMED CT, LOINC, ICD hanno assunto un ruolo centrale.

Tra tutti i KOS ad oggi esistenti in ambito sanitario in questa sede – come anticipato – prenderemo in esame la ICD, che storicamente ha rappresentato uno dei primi e più diffusi strumenti di strutturazione dei dati sanitari originariamente pensato per la codifica statistica delle cause di morte, ma poi ampiamente adottato anche per diagnosi cliniche, gestione amministrativa e sorveglianza sanitaria (Moriyama et al. 2011). L'evoluzione della pratica clinica e l'ampliamento degli usi dei dati hanno portato ad una crescente complessità e specializzazione tradottasi in diverse versioni dell'ICD. Ciò ha portato ad una frammentazione delle implementazioni (ad es. anche all'interno di uno stesso Paese, vengono utilizzate diverse varianti nazionali ICD-10-CM, ICD-10-AM, o come in Italia, ICD9-CM e ICD-10). La presenza simultanea di più versioni nei sistemi sanitari ha creato una disomogeneità semantica che ostacola l'interoperabilità dei dati, specialmente nei contesti che richiedono codifiche coerenti e mappabili, come i flussi transfrontalieri previsti dall'eHDSI o la costruzione dello Spazio Europeo dei Dati Sanitari.

4. L'ICD nel contesto nazionale italiano e il suo utilizzo per la gestione delle informazioni a livello centrale

A partire dal 1893, la Conferenza dell'Istituto internazionale di statistica, che ebbe luogo a Chicago, approvò la Classificazione internazionale delle cause di morte e l'Italia, a partire dal 1924, diede avvio alla classificazione per le statistiche concernenti la mortalità. Dal 1979, è stata poi rilasciata una versione modificata ed ampliata, introducendo gli interventi e le procedure diagnostiche e terapeutiche del sistema di classificazione, l'*International Classification of Diseases* 9th *revision Clinical Modification* (ICD9-CM)[6]. Le modifiche, che nel tempo sono state sempre più sostanziali, hanno avuto lo scopo di permettere sia una classificazione maggiormente precisa ed analitica delle formulazioni diagnostiche, sia di introdurre la classificazione delle procedure diagnostiche e terapeutiche.

[6] Il termine *clinical* è usato per mettere in evidenza le modifiche introdotte: rispetto alla ICD-9, fortemente caratterizzata dall'orientamento a scopo di classificazione delle cause di mortalità, la ICD-9-CM si è orientata soprattutto a classificare le informazioni riguardanti la morbosità.

L'introduzione della classificazione ICD-9 nel sistema sanitario italiano e la sua traduzione ufficiale sono avvenute con il Decreto del Ministero della Sanità del 26 luglio 1993 (Decreto Ministeriale 26 luglio 1993), che ha definito le modalità di codifica delle informazioni cliniche contenute nella SDO. Essa rappresenta lo strumento standard attraverso cui vengono raccolti, su tutto il territorio nazionale, i dati relativi a ogni episodio di ricovero presso strutture pubbliche e private. Sebbene inizialmente ideata per finalità amministrative, la Scheda di Dimissione Ospedaliera (SDO) si è affermata nel tempo come un sistema informativo strategico, grazie al valore delle informazioni raccolte, sia cliniche che organizzative. È oggi utilizzata per la programmazione sanitaria, il monitoraggio dei Livelli Essenziali di Assistenza (LEA), l'analisi epidemiologica, la valutazione delle performance ospedaliere, oltre che per finalità medico-legali[7]. Il già citato Decreto del 1993 ha rappresentato un passaggio chiave, definendo più chiaramente il contenuto del flusso informativo e introducendo il concetto di "debito informativo" nei confronti del Ministero della Salute. Il decreto ha anche stabilito che, a partire dal 1° gennaio 1995, la SDO avrebbe sostituito la precedente rilevazione dell'attività ospedaliera, fino ad allora effettuata tramite il modello ISTAT/D10 (Bracci et al. 2022). Negli anni successivi, sono stati adottati diversi provvedimenti per aggiornare e ampliare il tracciato informativo della SDO. In particolare, con il (Decreto del 27 ottobre 2000, n. 380), è stato introdotto un primo importante aggiornamento, che ha previsto l'adozione della versione 1997 dell'ICD-9-CM, in sostituzione della precedente versione ICD-9, e un ampliamento del tracciato record per raccogliere un maggior numero di informazioni cliniche e amministrative. Il (Decreto 8 luglio 2010, no. 135) ha segnato un ulteriore passo avanti, modificando la frequenza di trasmissione dei dati: da semestrale a trimestrale nel 2010, e successivamente a mensile dal 2011, rafforzando così la tempestività del monitoraggio dell'attività ospedaliera. Nel tempo, il sistema si è arricchito anche di strumenti interpretativi e operativi, come le linee guida e le circolari ministeriali. Tra queste, la Circolare del 23 ottobre 2008, approvata dalla Cabina di Regia del Nuovo Sistema Informativo, ha fornito importanti indicazioni sulla codifica dei dati anagrafici e amministrativi. Nel 2010, l'Accordo Stato-Regioni ha ulteriormente aggiornato le linee guida per la codifica delle informazioni cliniche.

L'ultima revisione significativa è avvenuta con il (Decreto 7 dicembre 2016, no. 261), che ha introdotto importanti novità strutturali nel tracciato informativo della SDO, tra cui:

---

[7] Il percorso normativo che ha portato alla strutturazione del flusso SDO ha avuto inizio con il Decreto del Ministero della Sanità del 28 dicembre 1991, che ha istituito il flusso informativo delle dimissioni ospedaliere come sistema ordinario per la raccolta dei dati. Questo primo passo è stato seguito, nel giugno 1992, dalla pubblicazione delle Linee Guida, che hanno fornito le istruzioni operative per la compilazione, codifica e gestione della SDO, attribuendole lo stesso valore della cartella clinica.

- la registrazione puntuale dei trasferimenti intraospedalieri (con data e ora di ammissione, trasferimento e dimissione);
- la possibilità di rilevare diagnosi pregresse;
- l'identificazione, nel rispetto della privacy, dell'équipe chirurgica responsabile degli interventi;
- e l'ampliamento del contenuto clinico utile al tracciamento completo dell'evento assistenziale.

Il set informativo raccolto tramite la SDO comprende:
- dati anagrafici del paziente (età, sesso, residenza, titolo di studio);
- informazioni sul ricovero (durata, regime, modalità di dimissione);
- e dati clinici, come diagnosi principale, diagnosi secondarie e procedure diagnostiche e terapeutiche eseguite.

Il flusso delle SDO è oggi utilizzato per la gestione economico-amministrativa delle strutture sanitarie, per la programmazione sanitaria e valutazione dei LEA, analisi clinico-epidemiologiche, monitoraggio della qualità, appropriatezza ed efficienza delle prestazioni, valutazione del rischio clinico, e definizione dei criteri di riparto del Fondo Sanitario Nazionale. Nonostante la sua capillare diffusione e l'elevata copertura a livello nazionale, persistono alcune criticità legate alla variabilità nella qualità della compilazione, soprattutto nei primi anni di attuazione, e alla mancata adozione delle versioni più aggiornate dell'ICD (come ICD-10 o ICD-11), che consentirebbero di avere una terminologia aggiornata e sicuramente più estensiva. Tuttavia, la normativa continua a fare riferimento alla versione 2007 dell'ICD-9-CM.

4.1. Certificato di morte

A partire dal 2016, l'ISTAT ha adottato la l'ICD per la codifica delle cause di morte, conformandosi agli standard internazionali. La codifica avviene attraverso strumenti come il sistema Iris, che supportano l'applicazione coerente delle regole ICD e permettono di individuare in modo automatizzato e interattivo la causa iniziale di morte, secondo i criteri stabiliti dall'OMS. Con l'introduzione della versione ICD-11, nella sua edizione aggiornata al maggio 2019 ed entrata in vigore a livello internazionale il 1° gennaio 2022, è stato necessario un rilevante adeguamento organizzativo e formativo del personale addetto alla codifica. L'ICD-11 ha infatti introdotto modifiche strutturali sostanziali rispetto alle versioni precedentemente in uso, in particolare alla ICD-9-CM, che – come detto precedentemente – costituisce il sistema ufficialmente adottato in Italia per la SDO.

Le differenze tra ICD-9 e ICD-11 sono sostanziali e non riguardano soltanto l'ampliamento del numero di entità nosologiche o la maggiore granularità della codifica, ma coinvolgono anche nuovi criteri semantici e logici per la selezione della causa iniziale di morte, nonché la riorganizzazione delle categorie diagnostiche. In questo scenario, il disallineamento tra le versioni utilizzate nei diversi sistemi informativi sanitari può costituire una criticità rilevante per la produzione di dati comparabili e per lo svolgimento di analisi epidemiologiche e statistiche longitudinali. Una criticità di non poco conto considerando che tutti i documenti sono potenzialmente collegati semanticamente e il loro utilizzo integrato può fornire un supporto concreto per gli studi epidemiologici di settore.

4.2. La Cartella clinica elettronica

Obiettivo primario delle cartelle cliniche è il supporto alla cura del paziente attraverso la registrazione precisa e puntuale delle informazioni, ma esse rivestono una più ampia funzione anche a livello giuridico. All'interno delle cartelle cliniche si possono individuare varie sezioni: anagrafica, anamnesi, risultati dell'esame obiettivo, sintomi attuali (stato), valutazione e trattamento, l'esito del trattamento, la lettera di dimissione e chi ha redatto il verbale (Dalianis 2018).

La Cartella Clinica Elettronica (CCE) rappresenta uno strumento essenziale per la gestione centralizzata e digitalizzata delle informazioni sanitarie relative a un paziente durante il periodo di degenza presso una struttura ospedaliera. Essa consente la registrazione e l'accesso strutturato a dati clinici, visite specialistiche e indagini diagnostiche da parte del personale sanitario autorizzato (Ingenito 2021).

Nel contesto normativo italiano, uno dei primi riferimenti alla CCE si trova nell'articolo 47-bis, comma 1, del (Decreto Legge 9 febbraio 2012, no. 5). Tale disposizione stabilisce che senza ulteriori oneri a carico delle strutture, i piani sanitari nazionali e regionali possono e devono promuovere l'adozione della gestione elettronica della documentazione clinica, ponendola sullo stesso piano dei sistemi di prenotazione online, favorendo dunque una minore pressione sui costi e una migliore accessibilità da parte dei pazienti.

Successivamente, l'articolo 13, comma 5, del (Decreto Legge 18 ottobre 2012, no. 179) ha introdotto il comma 1-bis all'interno del già citato articolo 47-bis, stabilendo che, a partire dal 1° gennaio 2013, la conservazione delle cartelle cliniche può avvenire anche esclusivamente in formato digitale. Pur non penalizzando con sanzioni pecuniari le amministrazioni inadempienti, questo ulteriore aggiornamento incentiva l'utilizzo del documento elettronico dandogli piena validità probatoria anche ai fini della sua esclusiva conservazione digitale. Queste norme hanno quindi sancito il passaggio normativo e

operativo dal supporto cartaceo alla documentazione elettronica, sottolineandone i benefici sia in termini economici che di efficienza e accessibilità. Oltre alla completezza informativa, un aspetto cruciale della CCE riguarda le sue caratteristiche strutturali e funzionali. Il documento informatizzato deve essere progettato per costituire una base solida che possa consentire future elaborazioni e analisi, assicurando l'adozione di un insieme minimo di dati, l'utilizzo di codifiche, formati standard. Inoltre, devono essere sempre riportate le informazioni relative all'efficacia delle terapie e allo stato clinico del paziente.

Infine, nell'ambito della trasformazione digitale del sistema sanitario, è fondamentale che le piattaforme utilizzate per la gestione delle CCE siano interoperabili, ovvero capaci di integrarsi con altri sistemi informativi, registri istituzionali e, se necessario, con le cartelle cliniche di membri della stessa famiglia, sempre nel rispetto delle normative vigenti in materia di protezione dei dati personali.

4.3. Patient Summary

Il Patient Summary (PS) o Profilo Sanitario Sintetico (PSS) istituito ufficialmente in Italia nel 2015, come già accennato, è un documento riassuntivo che raccoglie le informazioni cliniche essenziali di un paziente, con l'obiettivo di fornire un quadro completo e immediatamente accessibile dello stato di salute, delle condizioni mediche e delle necessità di trattamento del paziente, anche in riferimento a prestazioni erogate al di fuori del Servizio Sanitario Nazionale (Decreto 26 settembre 2023, no. 165). Tale documento, obbligatorio in molti contesti sanitari, in particolare nell'ambito del Fascicolo Sanitario Elettronico (FSE) (Ministero della Salute et al. n.d.) è pensato per essere condiviso tra diversi livelli di assistenza, consentendo una visione complessiva dello stato di salute del paziente e facilitando il trasferimento delle informazioni tra professionisti sanitari. Si tratta di un documento cruciale per garantire la continuità e la sicurezza delle cure, poiché consente al personale sanitario di avere un'immediata panoramica sulla storia medica del paziente, su diagnosi, trattamenti precedenti, allergie, terapie in corso e altri aspetti clinici vitali soprattutto in situazioni emergenziali. L'utilizzo di ICD-9-CM per la sua compilazione è stato normato dal Disciplinare Tecnico allegato al Regolamento citato sul FSE per ciò che concerne la lista dei problemi in atto sul paziente al momento della visita da parte del Medico di Medicina Generale (MMG), che è il responsabile della creazione, compilazione e aggiornamento del Piano di Sorveglianza Sanitaria (PSS).

Conclusioni

La gestione della conoscenza ha rappresentato, sin dall'antichità, un pilastro fondativo dell'agire umano, analogamente a quanto accade oggi con la trasformazione digitale, che si configura come un catalizzatore imprescindibile per l'ampliamento, la condivisione e la valorizzazione della conoscenza collettiva. Tuttavia, affinché tale conoscenza possa generare reale valore aggiunto, risulta imprescindibile che essa sia correttamente standardizzata, codificata e condivisa, in maniera strutturata e interoperabile.

Questi principi trovano applicazione concreta in molteplici ambiti e in particolare nel quadro della e-Health, intesa come l'insieme delle soluzioni tecnologiche applicate alla promozione, tutela e gestione della salute pubblica. Negli ultimi anni, anche a seguito dell'emergenza pandemica da SARS-CoV-2, si è assistito a una significativa accelerazione nella diffusione di soluzioni digitali in sanità, tra cui la telemedicina, che si è imposta come risposta necessaria alla limitazione dell'accesso in presenza ai servizi di assistenza non urgenti.

La sanità digitale rappresenta un'opportunità strategica per la ridefinizione dei modelli di cura, abilitando nuove modalità di assistenza, miglioramento dei percorsi esistenti, e ampliando il potenziale della ricerca clinica e traslazionale su scala internazionale. L'analisi condotta sul livello di maturità digitale dei sistemi sanitari (Boscolo et al. 2024) evidenzia tuttavia che il progresso in questo ambito non dipende esclusivamente dalla disponibilità di risorse economiche, dalla normativa di riferimento o dalle infrastrutture tecnologiche, ma richiede una progettazione integrata, sistemica e armonizzata tra tutti i livelli istituzionali e operativi. In tale contesto, l'utilizzo strutturato di sistemi di classificazione internazionali, come l'ICD, nato con finalità statistiche ed epidemiologiche, si dimostra sempre più uno strumento efficace anche per la programmazione sanitaria, la sorveglianza epidemiologica e la gestione dell'igiene pubblica. In Italia – come discusso – l'ICD trova applicazione in numerosi documenti clinico-assistenziali – dalle cartelle cliniche alle lettere di dimissione, dai certificati di morte al Profilo Sanitario Sintetico (PSS) – ma si rileva un utilizzo spesso non sistematico, parziale o semanticamente errato, con conseguente perdita di efficacia nei processi di interoperabilità semantica e tecnica. L'adozione coerente ed estensiva di standard terminologici internazionali è una condizione necessaria per abilitare ecosistemi informativi interoperabili, in grado di supportare decisioni cliniche, gestionali e politiche basate su dati di qualità. L'analisi del panorama italiano evidenzia che, anche se disponiamo di un quadro legislativo e programmatico avanzato, si continua a riscontrare una disomogeneità applicativa tra i vari livelli di governo (nazionale, regionale, locale), determinando una frammentazione operativa che ostacola la piena integrazione dei sistemi. La discontinuità tra iniziative nazionali e regionali, unite alla scarsa sensibilità del personale sanitario nei confronti delle tema-

tiche digitali – in particolare quelle relative all'interoperabilità e alla gestione corretta dei dati – rappresentano ulteriori criticità strutturali. Un esempio emblematico è il limitato utilizzo del Profilo Sanitario Sintetico da parte dei medici di medicina generale, il cui corretto impiego sarebbe invece strategico per migliorare la continuità assistenziale, in particolare nei casi di assistenza sanitaria transfrontaliera, previsti dalla normativa europea. A livello europeo l'EHDS si configura, infatti, come uno strumento strategico per affrontare le sfide evidenziate, ponendosi l'obiettivo di garantire accesso, condivisione e riutilizzo standardizzato dei dati sanitari in modo interoperabile, sicuro ed eticamente fondato. L'EHDS, promuovendo l'adozione di standard comuni, può contribuire a superare la frammentazione normativa e infrastrutturale nazionale, favorendo un approccio sinergico tra infrastrutture digitali, istituzioni regolatorie e attori dell'implementazione tecnologica.

In tale scenario emerge con forza la necessità di investire sistematicamente nella formazione continua e nella qualificazione professionale degli operatori sanitari coinvolti nei processi di digitalizzazione. L'effettiva adozione di strumenti complessi quali ICD, SNOMED CT o LOINC richiede competenze specifiche, non solo tecniche ma anche concettuali, per evitare errori di codifica, ambiguità semantiche e problemi di interoperabilità.

Nel contesto nazionale italiano, è opportuno sottolineare l'esistenza e il ruolo del gruppo LOINC Italia (LOINC Italia 2025)[8], che opera come referente ufficiale del Regenstrief Institute, l'Ente statunitense responsabile della creazione del sistema di codifica LOINC. Tale gruppo costituisce il riferimento italiano per la diffusione, coordinamento e supporto all'adozione del sistema LOINC in Italia, con l'obiettivo di promuovere il suo utilizzo, ma anche fornire formazione e supporto continuo. Il gruppo è inoltre attivamente impegnato nell'organizzazione di attività formative rivolte a professionisti sanitari, responsabili di sistemi informativi con l'intento di diffondere la conoscenza del sistema e favorire l'adozione consapevole e tecnicamente corretta dei codici. Tale supporto riguarda principalmente le operazioni di mappatura svolte a partire dai codici "locali" in uso presso i singoli laboratori verso lo standard LOINC, l'integrazione nei sistemi informativi sanitari e l'allineamento con i requisiti normativi nazionali ed europei in materia di interoperabilità semantica e standardizzazione delle informazioni cliniche (Chiaravalloti et al. 2025; Liscio et al. 2025).

Tale esempio rappresenta un importante esempio di come la competenza specialistica sia fondamentale nell'adozione e utilizzo consapevole di un sistema che nasce – come nel caso citato – per un contesto applicativo ben limitato.

---

[8] L'Istituto di Informatica e Telematica (IIT) del Consiglio Nazionale delle Ricerche (CNR) è partner italiano ufficiale del Regenstrief Institute (LOINC n.d.), con il quale ha siglato nel 2014 un Memorandum of Understanding che lo riconosce autore della versione ufficiale in lingua italiana dello standard internazionale di codifica LOINC.

A tale proposito, si potrebbe dunque pensare che nel vasto panorama dei KOS in ambito sanitario esistono anche dei sistemi che potrebbero meglio supportare le analisi e la strutturazione delle informazioni in modo da poterle gestire a livello centrale, come ad esempio SNOMED-CT nello studio di (Cazzaniga et al. 2023) o altre terminologie più specialistiche e settoriali come nello studio di (Togni et al. 2025).

Una governance multilivello e centralizzata dotata di competenze tecnico-normative, in grado di supervisionare l'adozione coerente dei sistemi di classificazione, garantire l'aggiornamento delle versioni in uso, e coordinare i processi di armonizzazione semantica tra le diverse componenti del sistema sanitario (clinica, amministrativa, epidemiologica). Solo attraverso una visione strategica unitaria, affiancata da una capacità operativa distribuita e sostenuta da adeguate competenze professionali, sarà possibile costruire un ecosistema informativo sanitario nazionale realmente interoperabile, integrato e orientato al valore, capace di generare benefici concreti per il cittadino, i professionisti e il sistema salute nel suo complesso.

Ringraziamenti

Questo studio prende spunto dalla tesi della Dott.ssa Eleonora Palumbo (Università della Calabria, Corso di Laurea Magistrale in Gestione e Conservazione dei Documenti Digitali), seguita dalla sottoscritta in qualità di relatrice. Un ringraziamento particolare per aver fornito alcuni materiali e talune analisi.

Riferimenti bibliografici

Blouin Genest, Gabriel. 2015. "World Health Organization and Disease Surveillance: Jeopardizing Global Public Health?" *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine* 19 (6): 595-614. https://doi.org/10.1177/1363459314561771.

Boscolo, Paola Roberta, Gianmario Cinelli, Francesca Guerra, e Francesco Petracca. 2024. "PNRR sanità: la lenta metamorfosi digitale del SSN." *Agenda Digitale*, 4 dicembre, 2024. https://www.agendadigitale.eu/sanita/pnrr-sanita-la-lenta-metamorfosi-digitale-del-ssn/.

Bracci, Tania, Simona Cinque, Francesco Grippo, e Chiara Orsi, a cura di. 2022. *Codifica delle cause di morte con l'ICD-10*. Istituto Nazionale di Statistica.

Buchanan, David. 2003. "Social Epidemiology: Berkman, LF, Kawachi I (Eds) Oxford University Press, New York, 2000, pp. 391." *Health Education Research* 18 (3): 404-07. https://doi.org/10.1093/her/cyg020.

Cazzaniga, Giorgio, Albino Eccher, Enrico Munari, et al. 2023. "Natural Language Processing to extract SNOMED-CT codes from pathological reports." *Pathologica* 115 (6): 318-24. https://doi.org/10.32074/1591-951x-952.

Chiaravalloti, Maria Teresa, Grazia Serratore, Fabio Del Ben, and Agostino Steffan. 2025. "LOINC Mapping Experiences in Italy: The Case of Friuli-Venezia Giulia Region." In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies.* HEALTHINF, vol. 2. https://doi.org/10.5220/0013380800003911.

Choi, Bernard C.K., Noël C. Barengo, and Paula A. Diaz. 2024. "Public Health Surveillance and the Data, Information, Knowledge, Intelligence and Wisdom Paradigm." *Revista Panamericana de Salud Pública* 48 (marzo): 1. https://doi.org/10.26633/RPSP.2024.9.

Dalianis, Hercules. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer International Publishing. https://doi.org/10.1007/978-3-319-78503-5.

Decreto 26 settembre 2023, n. 165. "Regolamento recante modificazioni e integrazioni al regolamento recante norme concernenti l'aggiornamento della disciplina del flusso informativo sui dimessi dagli istituti di ricovero pubblici e privati." *Gazzetta Ufficiale*, 21 novembre 2023, no. 272.

Decreto 7 dicembre 2016, n. 261. "Regolamento recante modifiche ed integrazioni del decreto 27 ottobre 2000, n. 380 e successive modificazioni, concernente la scheda di dimissione ospedaliera." *Gazzetta Ufficiale*, 7 febbraio 2017, no. 31.

Decreto 8 luglio 2010, n. 135. "Regolamento recante integrazione delle informazioni relative alla scheda di dimissione ospedaliera, regolata dal decreto ministeriale 27 ottobre 2000, n. 380." *Gazzetta Ufficiale*, 20 agosto 2010, no. 194.

Decreto Legge 27 ottobre 2000, n. 380. "Regolamento recante norme concernenti l'aggiornamento della disciplina del flusso informativo sui dimessi dagli istituti di ricovero pubblici e privati." *Gazzetta Ufficiale*, 19 dicembre 2000, no. 295.

Decreto Legge 9 febbraio 2012, n. 5. "Disposizioni urgenti in materia di semplificazione e di sviluppo." *Gazzetta Ufficiale*, 6 aprile 2012, no. 82.

Decreto Ministeriale 26 Luglio 1993. "Disciplina del flusso informativo sui dimessi dagli Istituti di ricovero pubblici e privati." *Gazzetta Ufficiale*, 3 agosto l993, no. 180.

Decreto Legge 18 ottobre 2012, n. 179. "Ulteriori misure urgenti per la crescita del Paese." Gazzetta Ufficiale, 19 ottobre 2012, no. 245.

Deimel, Lucas, Maike Hentges, and Rainer Thiel. 2025. "The European Health Data Space and the Future of Cross-Border Healthcare." In *Digital Maturity in Hospitals*, edited by Armin Scheuer and Jörg Studzinski. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-80704-6_11.

Gnoli, Claudio, Vittorio Marino, e Luca Rosati. 2006. *Organizzare la conoscenza: Dalle biblioteche all'architettura dell'informazione per il web*. Tecniche nuove.

Hanna, Matthew G., Liron Pantanowitz, Rajesh Dash, et al. 2025. "Future of Artificial Intelligence: Machine Learning Trends in Pathology and Medicine". *Modern Pathology* 38 (4): 100705. https://doi.org/10.1016/j.modpat.2025.100705.

Hjørland, Birger. 2016. "Knowledge Organization." *Knowledge Organization* 43 (7): 475-84.

HL7 Italia. n.d. "Realm Italiano. Specifiche, Guide e White Paper di HL7 Italia". Consultato il 19 maggio 2025. https://www.hl7.it/realm-italiano/.

Ingenito, Chiara. 2021. "La rete di assistenza sanitaria on-line: la cartella clinica elettronica." *Federalismi.it. Rivista di diritto pubblico italiano, comparato, europeo* 5: 71-95.

Liscio, Alessandro, Maria Teresa Chiaravalloti, Vincenzo Giannattasio Dell'Isola, and Monica Moz. 2025. "Advancing LOINC Mapping with AI: Insights from LARA's RAG Implementation and Expert Evaluation." In *International Conference on Artificial Intelligence in Medicine (AIME)*, edited by R. Bellazzi, L. Sacchi, J. Juarez Herrero, and B. Zupan. *Lecture Notes in Computer Science* 15735. 230-34, Springer, Cham. https://doi.org/10.1007/978-3-031-95841-0_43.

LOINC. n.d. "Italian Translations." Consultato il 19 maggio 2025. https://loinc.org/international/italian/.

Ministero della Salute, Dipartimento per la trasformazione digitale, Ministero dell'Economia e delle Finanze. 2023. "Cosa contiene il Fascicolo Sanitario Elettronico." https://www.fascicolosanitario.gov.it/cosa-contiene.htm.

Moriyama, Iwao M., Ruth M. Loy, and Alastair H.T. Robb-Smith. 2011. *History of the Statistical Classification of Diseases and Causes of Death. National Center for Health Statistics*, edited by Harry M. Rosenberg and Donna L. Hoyert. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Hyattsville, Md: U.S. https://www.cdc.gov/nchs/data/misc/classification_diseases2011.pdf.

Nonaka, Ikujiro, and Hirotaka Takeuchi. 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press: New York, NY. https://doi.org/10.1093/oso/9780195092691.001.0001.

Semenza, Jan. 2014. "Climate Change and Human Health." *International Journal of Environmental Research and Public Health* 11 (7): 7347-53. https://doi.org/10.3390/ijerph110707347.

Sharma, Alok, Artem Lysenko, Shangru Jia, Keith A. Boroevich, and Tatsuhiko Tsunoda. 2024. "Advances in AI and Machine Learning for Predictive Medicine." *Journal of Human Genetics* 69 (10): 487-97. https://doi.org/10.1038/s10038-024-01231-y.

Togni Serena, Saracino Lucia, Cieri Mariangela, et al. 2025 "Implementing Oncologic Nursing Care Plans in Electronic Health Records with Two Taxonomies: A Pilot Study." *Western Journal of Nursing Research* 47 (3): 159-68. doi:10.1177/01939459241310402.

Unione Europea. 2017. "Regolamento (UE) 2017/745 del Parlamento europeo e del Consiglio del 5 aprile 2017 relativo ai dispositivi medici, che modifica la direttiva 2001/83/CE, il regolamento (CE) n. 178/2002 e il regolamento (CE) n. 1223/2009 e che abroga le direttive 90/385/CEE e 93/42/CEE del Consiglio". Gazzetta ufficiale dell'Unione europea, 5 giugno 2017. https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32017R0745.

Unione Europea. 2021. "Regolamento (UE) 2021/1153 del Parlamento europeo e del Consiglio del 7 luglio 2021 che istituisce il meccanismo per collegare l'Europa e abroga i regolamenti (UE) n. 1316/2013 e (UE) n. 283/2014." *Gazzetta ufficiale dell'Unione europea*, 14 luglio 2021. https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32021R1153.

Unione Europea. 2025. "Regolamento (UE) 2025/327 del Parlamento Europeo e del Consiglio dell'11 febbraio 2025 sullo spazio europeo dei dati sanitari e che modifica la direttiva 2011/24/UE e il regolamento (UE) 2024/2847." *Gazzetta Ufficiale dell'Unione Europea*, 5 marzo 2025. https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=OJ:L_202500327.

World Health Organization. Eastern Mediterranean Region. 2025. "eHealth." https://www.emro.who.int/health-topics/ehealth/.

Zeng, Marcia Lei. 2008. "Knowledge Organization Systems (KOS)." *Knowledge Organization* 35 (2-3): 160-82. https://doi.org/10.5771/0943-7444-2008-2-3-160.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-549-3
DOI 10.57574/59655493XX
pag. 63-84 (gennaio-giugno 2025)
Open access article licensed under CC-BY-NC-SA

La cartella clinica tra analogico e digitale: gestione e conservazione a norma

Grazia Serratore*

Abstract: This paper analyzes the status of the implementation of the Electronic Medical Record (EMR) in Italy, focusing on the transition from analog to digital systems. Through a definition of medical record and outlining its structure and primary functions, the distinction between the Electronic Health Record (EHR) and the EMR is then clarified, examining their Italian equivalents: Fascicolo Sanitario Elettronico (FSE) and Cartella Clinica Elettronica (CCE). Moreover, an overview of key European initiatives that aim to enhance health data exchange and interoperability in support of EMR development and adoption is presented. The discussion then turns to the Italian context, examining: i) the transition phase from analog to digital in medical record workflows; ii) the differences between the management and preservation of analog and digital medical records; iii) the probative value of analog versus digital records; and, iv) the legal requirements for long-term preservation of medical records and their contents.

Keywords: Electronic Health Record, Cartella Clinica Elettronica, Medical Records, Interoperability, Italian health information systems.

1. Introduzione

La cartella clinica rappresenta la verbalizzazione delle attività di cura svolte con riferimento ad un degente in una struttura sanitaria nel corso di un singolo episodio di cura. Il suo scopo principale è quello di documentare in maniera esaustiva e strutturata i dati clinici, i sintomi riferiti dal paziente, i segni rilevati dal personale sanitario, le diagnosi, le terapie prescritte e gli esiti delle cure. Il suo obiettivo è, dunque, quello di tracciare lo stato di salute del paziente e gli eventi avvenuti nel corso di un ricovero, in modo da favorire uno scambio e un utilizzo efficienti delle informazioni sanitarie tra i diversi attori coinvolti nel processo di cura.

* Dipartimento di Culture, Educazione e Società, Università della Calabria, Rende (CS), Italia; Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Rende (CS), Italia. srrgrz97r52m208c@studenti.unical.it; grazia.serratore@iit.cnr.it. ORCID: 0009-0009-9481-2213.

Considerando l'importanza che la cartella clinica, sia nel suo formato analogico e sia in quello digitale, riveste per finalità di cura, ricerca, sanità pubblica e di natura giuridica, questo contributo si propone di analizzare e descrivere questo strumento di cura, nel contesto europeo ed italiano, e di offrirne una panoramica complessiva sullo stato di attuazione. Secondo la definizione fornita dal Ministero della Salute nel 1992:

> la cartella clinica costituisce lo strumento informativo individuale finalizzato a rilevare tutte le informazioni anagrafiche e cliniche significative relative a un paziente e a un singolo episodio di ricovero. Ciascuna cartella clinica ospedaliera deve rappresentare l'intero episodio di ricovero del paziente nell'istituto di cura: essa, conseguentemente, coincide con la storia della degenza del paziente all'interno dell'ospedale. La cartella clinica ospedaliera ha così inizio al momento dell'accettazione del paziente in ospedale, ha termine al momento della dimissione del paziente dall'ospedale e segue il paziente nel suo percorso all'interno della struttura ospedaliera.

I contenuti della cartella clinica, invece, sono definiti dal D.M. Ministero della Sanità 5 agosto 1977 all'art. 24, che specifica: «è prescritta, per ogni ricoverato la compilazione della cartella clinica da cui risultino le generalità complete, la diagnosi di entrata, l'anamnesi familiare e personale, l'esame obiettivo, gli esami di laboratorio e specialistici, la diagnosi, la terapia, gli esiti e i postumi». Come sostiene Guarasci (2012), la cartella clinica possiede tutte le caratteristiche proprie di un fascicolo: ha come oggetto di riferimento univoco il paziente; è prodotta da un unico soggetto produttore, ossia la struttura sanitaria che la genera; fa riferimento ad uno specifico procedimento, vale a dire l'episodio di cura. Anche se i contenuti di una cartella clinica sono stabiliti a livello normativo, in alcune strutture ospedaliere si riscontra la tendenza ad alimentare questo strumento con tutte le informazioni sanitarie a disposizione, senza limitarle al singolo episodio di cura.

In Italia la cartella clinica ospedaliera viene ricondotta, per la prima volta, ad un formato digitale dall'art. 13, comma 1 bis del decreto-legge n. 179 del 18 ottobre 2012; successivamente il Regolamento in materia di fascicolo sanitario elettronico (DPCM del 29 settembre 2015 n. 178) e il Decreto del Ministero della salute del 7 settembre 2023 art. 3 ribadiscono che una cartella clinica, anche nel suo formato elettronico, rappresenta una delle tipologie documentali che possono alimentare il FSE.

Il citato decreto-legge n. 179 istituisce inoltre il FSE, dandone all'art. 12 la seguente definizione: «Il fascicolo sanitario elettronico è l'insieme dei dati e documenti digitali di tipo sanitario o sociosanitario generati da eventi clinici presenti e trascorsi, riguardanti l'assistito, riferiti anche alle prestazioni erogate al di fuori del Servizio Sanitario Nazionale». In questa visione il FSE ha quindi lo scopo di coprire l'intero percorso assistenziale del paziente, in quanto

alimentato continuativamente dai professionisti sanitari che lo prendono in cura, sia nell'ambito del Servizio Sanitario Nazionale che di quello offerti da operatori privati accreditati. A livello gerarchico e concettuale, il FSE è quindi un macro-insieme di dati e documenti sanitari che include, come sua parte integrante, anche la cartella clinica. Nonostante la denominazione, il FSE non può essere inteso come un fascicolo, per come indicato nella definizione in uso all'archivistica classica, poiché i dati relativi alla storia clinica di un assistito che confluiscono al suo interno, anche se gestiti e conservati in un unico contenitore, afferiscono di fatto a diversi soggetti produttori. Pertanto, il FSE sembrerebbe rispondere maggiormente alla definizione di dossier, soprattutto considerando che «il dossier non vive autonomamente – come il fascicolo archivistico – ma solo in rapporto a scopi specifici e a determinati domini di conoscenza» (Guarasci 2012). Tuttavia, il termine FSE viene ormai utilizzato nel settore per identificare un preciso contenitore documentale e complessivamente anche l'architettura strutturale sottostante. Mentre il FSE viene dunque impropriamente associato al fascicolo, pur configurandosi come un dossier, la cartella clinica invece è un fascicolo archivistico a tutti gli effetti.

Parallelamente a livello internazionale appare evidente la necessità di fare una distinzione terminologica tra i termini EHR e EMR, in modo da disambiguarli semanticamente e consentirne un uso consapevole, senza confonderli su un piano concettuale (Garrett e Seidman 2011). Questa necessità è stata già sottolineata nel 2008 all'interno del report della National Alliance for Health Information Technology all'Office of the National Coordinator for Health Information Technology (ONC)[1] sul tema *Defining Key Health Information Technology Terms* (Department of Health & Human Services 2008), nel quale la differenza di uso e scopo tra EHR e EMR è stabilita proprio dal diverso orientamento semantico dato dai termini *health* e *medical*. Un EHR viene definito come «An electronic record of health-related information on an individual that conforms to nationally recognized interoperability standards and that can be created, managed, and consulted by authorized clinicians and staff across more than one health care organization». Poiché questo strumento ambisce a garantire una visione globale dello stato di salute del paziente, i contenuti destinati ad alimentare un EHR non dipendono esclusivamente da una singola organizzazione sanitaria. Conseguentemente i sistemi di EHR (EHRs) possono essere creati, gestiti e consultati da stakeholders appartenenti ad organizzazioni sanitarie differenti, racchiudendo così l'intera storia clinica di un paziente. I primi prototipi di EHR sono stati sviluppati negli Stati Uniti a partire dagli anni Settanta e prevedevano solo la possibilità di effettuare la conservazione dei

[1] L'Office of the National Coordinator for Health Information Technology è una divisione del Department of Health and Human Services (HHS) degli Stati Uniti. È la principale entità federale incaricata di coordinare gli sforzi a livello nazionale per implementare soluzioni informatiche per lo scambio elettronico di informazioni e dati sanitari.

dati dei pazienti su computer autonomi, generando così di fatto delle cartelle cliniche computerizzate, anche note come Computer Store Records (CSR). A partire dagli anni Novanta, invece, lo sviluppo di soluzioni tecnologiche più performanti, ha permesso l'implementazione di sistemi informativi per la registrazione dei pazienti e del loro stato di salute (McCullough et al. 2010; Akindele e Soyemi 2021), che hanno successivamente e gradualmente incluso maggiori funzionalità (Chao et al. 2013; Adler-Milstein et al. 2015; Windle e Windle 2015; Campanella et al. 2016)"issue":"18","note":"publisher: American College of Cardiology Foundation","page":"1973-1975","source":"jacc.org (Atypon. Nel dominio sanitario italiano il concetto di EHR è considerato equivalente a quello di FSE. All'interno del medesimo report si offre, inoltre, una definizione di EMR come «An electronic record of health-related information on an individual that can be created, gathered, managed, and consulted by authorized clinicians and staff within one health care organization», correlandone così la provenienza dei contenuti ad una singola struttura sanitaria di erogazione. Tale definizione consente di considerare un EMR come equivalente concettuale della cartella clinica, in virtù del carattere circoscritto nell'ambito operativo di una singola organizzazione sanitaria.

La questione terminologica appena descritta rappresenta non solo un problema di definizioni, quanto piuttosto un vero e proprio problema concettuale, che rischia di creare confusione sia tra gli addetti ai lavori sia tra i non addetti, soprattutto per quanto riguarda la corretta gestione e conservazione dei dati e dei documenti sanitari all'interno di questi contenitori informativi, in modo da assicurarne le caratteristiche di immodificabilità, integrità, leggibilità, reperibilità e autenticità, secondo quanto stabilito dal Decreto Legislativo 7 marzo 2005, n. 82, agli articoli 20, 21, e 44.

2. La trasformazione digitale della cartella clinica, l'interoperabilità ed il ruolo degli operatori sanitari

Dopo l'avvento dei Sistemi di Cartella Clinica Elettronica (SCCE), la cartella clinica cartacea (CCC) ha continuato ad essere utilizzata all'interno dei sistemi informativi sanitari, a volte coesistendo con quella elettronica, a volte come unico strumento. Essa determina un'accessibilità limitata ai dati, poiché la disponibilità e la consultazione del supporto fisico dipendono dalla possibilità dei professionisti sanitati di accedervi nel reparto di ricovero, nel corso di un consulto o nell'archivio della struttura. Questa modalità analogica di gestione della cartella clinica presenta significative criticità operative: non consente, ad esempio, di coordinare in modo efficiente le prescrizioni mediche e le somministrazioni farmacologiche, né di tracciare in tempo reale gli spostamenti dei pazienti all'interno della struttura. Tali funzioni richiederebbero,

infatti, una visione integrata e aggiornata dei dati clinici, nonché la possibilità di recuperare le informazioni mediante interrogazioni sui dati, ovvero delle capacità che una CCC non possiede. Inoltre, la sua usabilità è spesso ottimizzata per specifici reparti o unità operative, rendendo meno intuitiva la ricerca delle informazioni di interesse per il personale sanitario di altri reparti. Questo causa limitazioni in termini di interoperabilità semantica, che ne riducono l'efficacia nell'agevolare il complesso e dinamico scambio di informazioni sanitarie all'interno di una struttura ospedaliera. L'implementazione e l'utilizzo di CCE permettono di superare queste limitazioni, consentendo di tracciare ogni azione degli attori coinvolti nel processo di cura, assicurando una visualizzazione più efficace dei dati e la possibilità di eseguire previsioni e analisi più esaustive sulla base dei dati sanitari disponibili (Holroyd-Leduc et al. 2011).

Un SCCE è lo strumento informatico deputato a svolgere sia una funzione documentale e sia di supporto alle attività cliniche. Le funzioni documentali comprendono la gestione, la fruizione e la condivisione dei dati anagrafici, amministrativi e sanitari degli assistiti, mentre le funzioni di supporto all'attività clinica si riferiscono alla possibilità di gestire, con l'ausilio di questo strumento, la terapia farmacologica, la redazione della lettera di dimissione ospedaliera e la visualizzazione dei referti. Inizialmente l'implementazione di un SCCE in una azienda ospedaliera richiede un notevole investimento di risorse economiche, in modo da soddisfare alcuni prerequisiti infrastrutturali, come l'adeguata copertura di rete delle aree cliniche, una dotazione di dispositivi commisurata ai bisogni del personale e l'implementazione di un sistema personalizzato per i bisogni della struttura; in seguito, devono essere previsti dei costi aggiuntivi per il mantenimento e l'aggiornamento (Cresswell et al. 2023). Le aziende ospedaliere sono inoltre tenute a formulare dei piani di continuità operativa, così che, in caso di un mal funzionamento, non ci siano ricadute sulla salute e sui processi di assistenza ai pazienti. In parallelo, risulta necessario dotarsi di repository sicuri per la conservazione a norma delle CCE, secondo quanto indicato dall'art. 51 del Decreto Legislativo 7 marzo 2005, n. 82 e dal Regolamento UE 2016/679 all'art. 25.

Gestire un ecosistema di dati sanitari in una struttura clinica significa anche affrontare alcune sfide cruciali, come assicurare l'interoperabilità tra unità operative diverse, garantire la sicurezza e la privacy dei dati, sviluppare interfacce intuitive per agevolare il lavoro degli operatori, monitorare la qualità e la completezza delle informazioni registrate, offrire programmi di formazione e supporto continuo al personale e definire una governance chiara dei dati, stabilendo ruoli, responsabilità e regole di accesso e utilizzo (Orhan e Kafes 2021). Solo attraverso una gestione attenta e coordinata di tutti questi aspetti è possibile sfruttare appieno le potenzialità di un SCCE e migliorare la qualità e l'efficienza dell'assistenza sanitaria. Tuttavia, affinché un SCCE possa garantire l'interoperabilità delle informazioni, è necessario intervenire sia a livello

tecnico sia a livello semantico. A livello tecnico e di processo occorre scegliere le architetture e i protocolli con cui i dati devono essere trasmessi da un sistema informativo all'altro, ovvero ad esempio da un reparto all'altro, e adottare adeguate modalità di gestione dei dati per garantire il mantenimento della loro valenza giuridica e, soprattutto, la loro riservatezza. A livello semantico, invece, è necessario utilizzare codifiche e standard comuni in modo che sistemi diversi interpretino le informazioni scambiate in modo chiaro e senza ambiguità.

Gli operatori sanitari, in particolare, sono stati riconosciuti come un fattore chiave nella trasformazione digitale del settore sanitario e, per questa ragione, è necessario investire sulla loro formazione digitale per vincere eventuali resistenze legate all'abitudine dell'utilizzo dell'analogico nella pratica clinica (Gewald et al. 2017; Brown et al. 2020). Considerando la costante evoluzione delle tecnologie dell'informazione è necessario non solo assicurare un aggiornamento delle competenze digitali ma, anche, prevedere l'erogazione di corsi di formazione periodici per il personale. L'età è sicuramente un fattore discriminante nell'utilizzo di strumenti digitali e i dati ISTAT, relativi all'anno 2021, hanno rivelato che l'Italia ha il personale medico più anziano d'Europa: il 55% dei medici aveva almeno 55 anni, e nel 2022, sia nell'ambito del SSN sia nel privato, l'età media dei medici specialisti era pari a 53,7 anni (SID 2024). Oggi un numero crescente di professionisti sanitari, soprattutto giovani e specializzandi, è nativamente abituato all'uso degli strumenti digitali e riconosce i benefici derivanti dalla digitalizzazione della sanità, mostrando aspettative elevate sulla qualità e l'efficienza dei sistemi. Tuttavia, se da un lato chi ha meno competenze digitali trova maggiori difficoltà nell'adottare nuove tecnologie perché teme di non utilizzarle correttamente, dall'altro chi è più esperto esprime preoccupazioni legate all'affidabilità dei sistemi, alla loro aderenza ai reali flussi informativi e al supporto alla cooperazione. Nello specifico, le principali esigenze evidenziate riguardano la facilità di recupero delle informazioni e di compilazione, la possibilità di ricostruire cronologicamente il percorso clinico-assistenziale e l'uso di un linguaggio condiviso tra operatori (Vehko et al. 2019). È probabile che un futuro ricambio generazionale contribuisca ad un cambio di paradigma e favorisca l'integrazione del digitale nella pratica clinica.

3. Il contesto europeo e i progetti per l'interoperabilità transfrontaliera

In Europa, i primi sistemi computerizzati per gestire le informazioni sui pazienti risalgono alla fine degli anni Sessanta. Uno dei primi prototipi risale nello specifico al 1968, quando a seguito della decisione di separare l'Università di Lovanio (Belgio) in due sedi – quella fiamminga, che sarebbe rimasta a Leuven, e quella francofona, da trasferire a Bruxelles per ospitare la facoltà di medicina – le autorità accademiche pianificarono lo sviluppo di un nuovo

ospedale a Bruxelles: le Cliniques Universitaires Saint-Luc. In questo contesto di riorganizzazione, si decise di informatizzare il sistema di cartella clinica, i flussi di informazioni relativi ad ammissioni, fatturazione, laboratori e terapia intensiva della nascente struttura ospedaliera. Il modello sviluppato divenne rapidamente un punto di riferimento, attirando numerose delegazioni ospedaliere dal Belgio e da altri Paesi europei interessati ad adottare sistemi simili (Roger France 2014). A partire dagli anni Dieci del Duemila, in Europa, i sistemi informativi ospedalieri, o Hospital Information Systems (HIS), hanno invece conosciuto un rapido sviluppo, supportati dalla sempre più diffusa consapevolezza dei benefici derivanti dal loro impiego, come si può leggere nella (Raccomandazione (UE) 2019/243):

> la capacità dei cittadini e dei prestatori di assistenza sanitaria di accedere alle Cartelle Cliniche Elettroniche, in formato digitale, e di condividerle in sicurezza a livello nazionale o transfrontaliero comporta numerosi vantaggi: il miglioramento della qualità dell'assistenza ai cittadini, la riduzione del costo dell'assistenza sanitaria per le famiglie e il sostegno alla modernizzazione dei sistemi sanitari nell'Unione, che si trovano sotto pressione per via dei cambiamenti demografici, dell'aumento delle aspettative e dei costi delle cure.

Tuttavia, nonostante i diversi Paesi membri si trovino ad affrontare delle sfide socioeconomiche molto simili nel campo della demografia, della salute e dell'assistenza sanitaria, questi non presentano un'adozione uniforme delle CCE. In assenza di una normativa vincolante, ogni Paese membro ha adottato in autonomia decisioni in materia di assistenza sanitaria, ha gestito i tempi e le modalità di digitalizzazione delle infrastrutture e dei sistemi sanitari secondo le proprie risorse. Conseguentemente, il rallentamento registrato in alcuni Stati membri ha avuto ripercussioni anche sullo scenario internazionale di interoperabilità.

L'UE svolge, dunque, all'interno di un panorama diversificato, un'azione complementare rispetto alle politiche sanitarie nazionali, ponendosi, attraverso progetti e finanziamenti, a sostegno dell'azione dei singoli Stati membri e svolgendo un ruolo di coordinamento. L'UE si è pronuncia per la prima volta in materia di sanità digitale in ottica transfrontaliera con la Direttiva 2011/24/UE del Parlamento europeo e del Consiglio concernente l'applicazione dei diritti dei pazienti relativi all'assistenza sanitaria transfrontaliera, incoraggiando l'istituzione di norme volte ad agevolare l'accesso ad un'assistenza transfrontaliera sicura e di qualità e a garantire la mobilità dei pazienti nell'UE[2]. Sono

[2] Prima di questo momento, si faceva riferimento alla Direttiva 2002/22/CE, del Parlamento Europeo e del Consiglio del 7 marzo 2002, relativa al servizio universale e ai diritti degli utenti in materia di reti e di servizi di comunicazione elettronica. Questa auspicava la diffusione delle moderne tecnologie informatiche in tutti i settori, ma non faceva esplicito riferimento al dominio sanitario.

di seguito menzionati alcuni progetti rappresentativi delle azioni compiute a livello europeo in materia di interoperabilità transfrontaliera.

Il programma Connecting Europe Facility (CEF), istituito con il Regolamento UE n° 1316/2013, ha previsto lo sviluppo dell'eHealth services-National Contact Point eHealth (NCPeH) e l'istituzione della Digital Service Infrastructure (eHDSI). Per garantire una migliore coordinazione e collaborazione tra gli Stati membri, sono stati istituiti i punti di contatto nazionali per i servizi di eHealth che fungono da intermediari tra le autorità nazionali e le istituzioni europee nell'ambito degli scambi transfrontalieri di documenti clinici. L'eHDSI, invece, è un'infrastruttura digitale nata per assicurare lo scambio sicuro e interoperabile di informazioni sanitarie elettroniche tra gli Stati membri, collegando tra loro i diversi punti di contatto nazionali di eHealth e consentendo di scambiare *Patient Summary* e *ePrescription*. I primi scambi test si sono svolti tra Estonia e Finlandia nel gennaio 2019.

Nel 2021, in linea con quanto già realizzato, è stato avviato il progetto NCPeH+ che punta ad offrire servizi a favore degli assistiti italiani che necessitano di cure all'estero e di cittadini europei che necessitano di assistenza nel nostro Paese. I servizi ad oggi sono stati implementati solo per alcune regioni, tra cui Lombardia, Veneto ed Emilia-Romagna. Questo progetto si colloca nello sviluppo dell'*EHealth Data Space* (EHDS) con l'intento di assicurare la copertura dei servizi transfrontalieri, rendendoli disponibili in modo crescente alla popolazione europea.

Il programma EU4Health 2021-2027, invece, è stato istituito dal Regolamento UE 2021/522, per rinforzare i sistemi sanitari e sanare le vulnerabilità emerse con la pandemia da Covid-19. EU4Health persegue gli obiettivi specifici di: rafforzare l'uso e il riutilizzo dei dati sanitari per la prestazione di assistenza sanitaria e per la ricerca e l'innovazione; promuovere la diffusione di strumenti e servizi digitali, nonché la trasformazione digitale dei sistemi sanitari; sostenere la creazione di uno spazio europeo dei dati sanitari (European Commission 2024).

Infine, si cita XpanDH, progetto avviato nel 2023 nell'ambito del programma Horizon Europe, che mira a sviluppare le capacità di persone e organizzazioni nel creare e adoperare soluzioni digitali sanitarie interoperabili. Il progetto ambisce anche allo sviluppo di specifiche tecniche per la costruzione di un formato europeo per lo scambio di CCE (EEHRxF) e a creare un modello funzionale per favorirne l'adozione, migliorare la cooperazione sanitaria e promuovere i Personal and European Health Data Spaces (European Union 2023).

I progetti citati sono solo alcune delle attività che rivelano l'impegno profuso per il raggiungimento di una sanità più digitale, inclusiva e interoperabile. Tuttavia, mentre da un lato le azioni e i progetti, a livello comunitario sono sempre più ambiziosi e vi è una maggiore consapevolezza dei benefici derivanti

dall'adozione diffusa di SCCE, dall'altro si riscontra ancora in molti Paesi una loro applicazione non sistematica. Tra le criticità maggiormente riscontrate si registrano: la presenza di dati sanitari relativi ad un medesimo paziente che provengono da fonti eterogenee e non interoperabili, nonché la presenza di dati duplicati; l'incompatibilità dei sistemi informativi elettronici adottati a livello locale; la raccolta di dati in modalità cartacea e la conseguente difficoltà di integrazione nei SCCE; il personale sanitario con scarse competenze digitali; la mancanza di risorse finanziarie, risorse umane e infrastrutture (Bogaert et al. 2021). L'accesso alle informazioni e la continuità delle cure sono dunque ostacolati da una limitata comunicazione tra i diversi attori, che impedisce di avere un quadro completo sullo stato di salute di un paziente e induce sia alla duplicazione di alcuni dati e sia alla creazione di alcune lacune nella storia clinica dell'assistito (OECD 2020; Sipido et al. 2020).

Un ulteriore motivo del rallentamento nell'adozione degli SCCE è riconducibile ad una loro impostazione per la gestione documentale, mentre oggi si assiste a un cambiamento di paradigma che pone sempre più attenzione sulla gestione, il riutilizzo e l'aggregazione dei dati sanitari. Nel dominio sanitario i dati hanno assunto una maggiore centralità in tutti i processi assistenziali, per cui è necessario adeguare le logiche che sottendono allo scambio comunicativo e alla gestione delle informazioni cliniche. Un errore comune, infatti, è stato quello di provare a traslare le procedure analogiche in ambiente digitale, continuando a ragionare in termini di documenti e non di dati (McDonald 1997; Orhan e Kafes 2021; Lenz e Reichert 2007). Ripensare alle procedure, agli scambi, alle modalità di gestione delle informazioni, tuttavia, può non essere sufficiente per risolvere il problema. È necessario che i dati e i documenti siano costruiti *ab origine* secondo standard condivisi, i quali svolgono il ruolo di lingua franca, garantendo una comunicazione priva di incomprensioni e ambiguità.

4. La Cartella Clinica Elettronica in Italia

In Italia, i SCCE possono essere considerati a tutti gli effetti dispositivi medici, secondo quanto stabilito dal Decreto legislativo n. 37 del 25 gennaio 2010, all'art. 1: «dispositivo medico: qualunque strumento, apparecchio, impianto, software, sostanza o altro prodotto, utilizzato da solo o in combinazione, compresi gli accessori tra cui il software destinato dal fabbricante ad essere impiegato specificamente con finalità diagnostiche e/o terapeutiche e necessario al corretto funzionamento del dispositivo stesso». Anche se nel decreto non viene fatto esplicito riferimento alle cartelle cliniche o ai SCCE, è possibile affermare che, a livello concettuale, questi rientrino pienamente nella definizione di dispositivo medico fornita dal legislatore, poiché i dispositivi medici elencati includono, non solo strumenti fisici, ma anche software destinati a

essere impiegati con finalità diagnostiche o terapeutiche. I SCCE perseguono queste medesime finalità, in quanto la loro funzione incide direttamente sulla qualità e sull'efficacia dei processi di cura.

In questi ultimi anni, lo sviluppo della CCE ha rappresentato una priorità per le strutture sanitarie italiane, come dimostrano i dati riportati dall'osservatorio Sanità Digitale del Politecnico di Milano relativi agli anni 2022 e 2025. Se nel 2022, infatti, il 42% delle strutture ospedaliere disponeva di una CCE attiva in tutti i reparti e il 23% aveva delle CCE attive solo parzialmente (Politecnico di Milano 2022), nel 2025 la CCE risulta presente nell'85% delle strutture sanitarie italiane, anche se solo nella metà dei casi si tratta di una soluzione diffusa in tutti i reparti. Il dato positivo è che il 10% delle strutture attualmente sprovviste prevede la sua introduzione entro la fine 2025, con un incremento che porterà dunque alla quasi copertura totale (Sironi 2025). In tutti gli altri casi, invece, risulta più difficile determinare le decisioni intraprese dalle singole aziende ospedaliere. Il panorama italiano in materia di sanità digitale appare dunque molto frammentato, soprattutto al livello di adozione e di sviluppo di differenti SCCE, sia in termini di contenuti che di modalità di gestione. Pur avendo sul mercato software sempre più sofisticati, le funzionalità offerte dai SCCE dei principali produttori presenti sul territorio nazionale e adottate negli ospedali sono spesso basilari, soprattutto a causa di vincoli economici, della scarsa formazione informatica del personale medico, della resistenza al cambiamento e della complessità dei processi di integrazione con i sistemi informativi già preesistenti. A ciò si aggiunge una generale mancanza di coordinamento a livello centrale che favorisca l'adozione uniforme di soluzioni più avanzate e standardizzate.

Le funzionalità implementate nei SCCE attualmente in uso sono: consultazione dei dati anagrafici e amministrativi del paziente; annotazione dell'anamnesi; registrazione dei parametri vitali; visualizzazione di esami di laboratorio e diagnostica per immagini; registrazione del trattamento farmacologico; annotazione del diario clinico; gestione e visualizzazione delle informazioni di riepilogo sulle condizioni di salute del paziente; e redazione della lettera di dimissione. Tuttavia, non sempre queste funzioni possono essere svolte in regime di interoperabilità tra reparti o tra strutture diverse, né supportano l'analisi clinico-epidemiologica dei dati, limitando il potenziale strategico della digitalizzazione in ambito sanitario. Ciò è dovuto primariamente alle differenze sostanziali nelle scelte adottate in strutture sanitarie diverse o, talvolta, nei reparti di una medesima struttura, come ad esempio l'utilizzo di soluzioni software differenti che prevedono la gestione di tipologie documentali dissimili (Brizzi 2021). Ne conseguono disomogeneità di contenuti e problemi negli scambi di informazioni che hanno delle ricadute sulla qualità e sui tempi di erogazione delle cure. Inoltre, ad oggi, in molte strutture sanitarie italiane si registra una coesistenza della dimensione digitale e di quella analogica, che spesso consiste

nella trasposizione digitale dei moduli cartacei utilizzati per documentare le attività svolte nei reparti e negli ambulatori o la stampa di documenti nativi digitali, causando inevitabilmente un rallentamento nel processo assistenziale.

Ulteriore ostacolo all'interoperabilità è dato dalla mancanza di un set minimo di dati all'interno della cartella clinica, che costituisca un nucleo comune di riferimento da arricchire poi con personalizzazioni cliniche necessarie per ogni specialità. Questo problema era già emerso nell'analogico, portando il Ministero della Salute a varare nel 2012 le Linee Guida per lo sviluppo di un modello di Cartella Paziente Integrata (CPI) (Ministero della Salute 2012), che suggerivano un modello di dati da seguire per uniformare i contenuti delle cartelle cliniche e, per ragioni di completezza ed opponibilità a terzi, proponevano di includere anche le registrazioni e le annotazioni degli infermieri e del personale paramedico. L'obiettivo era la creazione di un modello standard di CPI, progettato in analogico per facilitare le successive fasi di digitalizzazione, che avrebbe dovuto essere trasferibile e fruibile da parte delle aziende sanitarie di tutte le Regioni, offrendo una schematizzazione comune dei processi sanitari e un dizionario di riferimento per la terminologia generale. Tuttavia, queste linee guida sono state perlopiù disattese o non applicate in una maniera così estensiva da permettere di godere dei benefici previsti negli intenti progettuali.

La confusione provocata da una lunga fase di transizione e da una costituzione ibrida della cartella clinica è evidente da un utilizzo spesso ambivalente delle espressioni "cartella clinica elettronica" e "cartella clinica informatizzata", senza riuscire a cogliere le differenze che invece sussistono a livello concettuale. Le due denominazioni rappresentano, infatti, due diverse modalità di gestione delle informazioni cliniche dei pazienti. La cartella clinica informatizzata presenta una gestione ibrida tra analogico e digitale: i documenti che confluiscono nella cartella clinica informatizzata sono nativamente digitali e hanno una loro gestione elettronica nelle fasi di produzione, redazione e modifica, ma completano il proprio ciclo di vita documentale in formato analogico (Orizio e Gotti 2016). Questa prassi, ancora diffusa, risulta concettualmente scorretta, poiché la gestione ibrida della cartelle clinica informatizzata introduce criticità dal punto di vista giuridico e archivistico, poiché il documento originale resta il documento nativo digitale e firmato digitalmente. Inoltre, seppur radicate nella consuetudine clinica, questa pratica compromette la coerenza del sistema documentale e rischia di vanificare i vantaggi della digitalizzazione dei processi nel contesto sanitario. Una CCE, invece, è un documento nativo digitale e svolge il suo ciclo di vita documentale interamente in un ambiente digitale, in quanto redatta, gestita e conservata esclusivamente in modalità digitale. Nel 2012, l'Associazione Italiana Sistemi Informativi in Sanità definiva la CCE ospedaliera come l'«insieme delle informazioni gestite da processi del percorso diagnostico-terapeutico-assistenziale ospedaliero supportati dalle tecnologie informatiche». Tuttavia, non si tratta esclusivamente di un supporto delle

tecnologie dell'informazione alle consuete pratiche mediche, quanto dell'adozione e del mantenimento di un servizio. Infatti, la CCE non deve essere considerata una copia della CCC, ma deve corrispondere ad un mutamento a livello concettuale sulle modalità di intendere e gestire i processi amministrativi, organizzativi e manageriali della struttura sanitaria, in modo che questa possa effettivamente essere uno strumento di ausilio nella pianificazione e nella valutazione delle cure e si ponga come mezzo di comunicazione tra i professionisti sanitari.

Nel contesto nazionale italiano negli ultimi anni si sono sviluppati alcuni prototipi ed applicazioni particolarmente significativi di CCE. Nel 2016 la Regione Emilia-Romagna ha redatto il documento "Cartella Clinica Elettronica. Linee Guida Tecniche per l'acquisizione, l'adeguamento e l'implementazione clinica", contenente delle indicazioni molto dettagliate per l'implementazione di un SCCE all'interno di ogni struttura ospedaliera regionale (Regione Emilia-Romagna 2016). Grazie a queste linee guida è stato avviato un processo di informatizzazione delle tre aziende ospedaliere della città di Bologna, che ha portato alla costruzione di un modello comune di CCE Pediatrica (CCEP) (Sist et al., 2022). Nel 2021, la Fondazione Policlinico Universitario Agostino Gemelli IRCCS, grazie al lavoro di una equipe multidisciplinare, ha definito le specifiche tecniche e funzionali per adottare un SCCE nelle aree di terapia intensiva, con l'obiettivo di registrare direttamente nel sistema le misurazioni ottenute da diversi dispositivi di monitoraggio, alcuni dei quali posizionati ai posti letto dei pazienti. Sempre nello stesso anno, negli Spedali civili di Brescia è stata messa a regime una piattaforma che ha coinvolto e collegato oltre cento centri di onco-ematologia italiani, con lo scopo di valutare da remoto l'idoneità dei pazienti ad essere candidati allo svolgimento di terapie CAR-T, una delle cure più innovative per diversi tipi di leucemie, linfomi e tumori. L'idoneità dei pazienti viene valutata mediante teleconsulto e i dati clinici vengono registrati in una CCE specialistica. Infine, si riscontrano dei miglioramenti anche nelle Regioni italiane che, fino ad ora, sono sempre rimaste più indietro nei processi di transizione digitale. Ad esempio, nel 2023, in Calabria il reparto di Cardiologia dell'Ospedale Civile Nicola Giannettasio di Corigliano-Rossano (CS) è stato scelto come sito pilota per l'utilizzo delle CCE nella pratica clinica e, sempre in Calabria, si trova un altro centro di eccellenza nel campo della Cardiologia e della telecardiologia; nell'Ospedale Civile Ferrari di Castrovillari (CS), infatti, la telecardiologia e le televisite permettono, ormai da circa quattro anni, la compilazione telematica dei piani terapeutici e uno snellimento nelle procedure di gestione delle visite di controllo, soprattutto nel caso di malati cardiopatici cronici.

5. Gestione e conservazione della cartella clinica elettronica

La gestione e la conservazione della cartella clinica presentano differenze significative a seconda che si tratti di documenti nativi digitali o analogici, con impatti rilevanti sia sui processi organizzativi sia sulle responsabilità degli attori coinvolti. Nell'analogico si possono distinguere due fasi nella gestione delle cartelle cliniche. Nella prima, che coincide con il periodo di degenza del paziente nell'ospedale, la cartella clinica rimane nel reparto per essere consultata e arricchita con nuove informazioni. Nella seconda fase, a seguito della dimissione del paziente, la CCC deve invece pervenire, entro il termine massimo di trenta giorni dalla dimissione, alla Direzione del Presidio Ospedaliero per essere inserita e custodita nell'archivio centrale della struttura sanitaria. Per ogni fase di gestione delle cartelle cliniche, la responsabilità della conservazione è attribuita a figure diverse: inizialmente, al primario responsabile dell'unità operativa e, poi, al direttore sanitario (Decreto del Presidente della Repubblica 27 marzo 1969, n. 128, art. 7).

Secondo il decreto-legge del 18 ottobre 2012 n. 179, art. 13 comma 1-bis: «la conservazione delle cartelle cliniche può essere effettuata, senza nuovi o maggiori oneri a carico della finanza pubblica, anche solo in forma digitale». L'attuazione di questa norma rappresenta un passaggio cruciale per l'ammodernamento dei sistemi informativi del dominio sanitario, poiché sancisce la piena legittimità della conservazione digitale delle cartelle cliniche, legittimando la messa in essere di un ciclo di vita interamente digitale per questa tipologia documentale. Le CCE sono gestite all'interno di SCCE, assicurando che siano mantenute inalterate nel tempo le caratteristiche di affidabilità, autenticità, integrità, leggibilità e reperibilità dei documenti (Decreto Legislativo 7 marzo 2005, n. 82, art. 44, comma 1 ter).

Le singole strutture sanitarie, in qualità di soggetti produttori, hanno l'obbligo di garantire la corretta gestione del ciclo di vita delle CCE creando i propri documenti, gestendoli e, infine, conservandoli *in house* o *in outsourcing* (Sorrentino et al. 2020)authenticity, integrity and readability over time are essential principles on which the efficiency of healthcare facilities could be measured. Realizing digital preservation in full compliance with regulations is extremely important, especially in a sensitive domain such as healthcare, and above all considering this process in the frame of the Italian Federated Electronic Health Record, namely Fascicolo Sanitario Elettronico (FSE. Il sistema di gestione informatica delle CCE è organizzato per assicurare l'indicizzazione e la ricerca dei documenti, mentre il sistema di conservazione dei documenti informatici garantisce l'integrità, la leggibilità e l'agevole reperibilità dei documenti e delle informazioni identificative. Il sistema di gestione e quello di conservazione sono distinti ma interoperanti; questi devono essere in grado di assicurare correttamente il trattamento dei contenuti e richiedono, sin dalla

loro progettazione, una definizione di tutti i flussi di dati e di documenti che avvengono all'interno della struttura. Deve essere definito a monte un modello di processi molto dettagliato, con indicazione dei requisiti di accesso ai documenti in base alle funzioni svolte dai professionisti sanitari coinvolti nei processi di cura. Ogni utente si autentica mediante la verifica di opportune credenziali e, a seconda del ruolo ricoperto, il sistema procede con l'individuazione dei livelli di accesso. Pertanto, i SCCE devono essere appositamente progettati per monitorare gli accessi, garantendo una maggiore e veloce reperibilità delle informazioni del paziente nel rispetto dei principi di *privacy by design* e di *privacy by default* (Regolamento UE 2016/179, art. 25) e adottando misure tecniche e organizzative funzionali a garantire un livello di sicurezza dei dati adeguato al rischio (Regolamento UE 2016/179, art. 32).

Per quanto concerne i tempi di conservazione delle cartelle cliniche, la Circolare n. 61 del Ministero della Sanità del dicembre 1986 sancisce che queste devono essere conservate illimitatamente, poiché costituiscono un atto ufficiale, indispensabile a garantire la certezza del diritto, oltre a costituire una fonte documentaria per le ricerche di carattere storico sanitario. La medesima Circolare indica un arco temporale di vent'anni come periodo minimo e sufficiente per la tenuta delle radiografie e, in analogia con quanto appena riportato, si ritiene che anche altra documentazione diagnostica possa essere assoggettata allo stesso periodo di conservazione di venti anni. La necessità di conservare illimitatamente le cartelle cliniche viene indirettamente ribadita anche dal Decreto 7 settembre 2023 "Fascicolo Sanitario Elettronico 2.0", dove all'art. 10, comma 2 si legge «I dati e i documenti del FSE, inclusi il profilo sanitario sintetico […] e il taccuino personale dell'assistito […], e fatta eccezione per la cartella clinica e i documenti afferenti alla stessa, vengono cancellati dal titolare del trattamento decorsi trent'anni dalla data del decesso dell'assistito stesso, con periodicità annuale». Questo sottolinea il ruolo unico e insostituibile rivestito dalla cartella clinica che costituisce una fonte essenziale per tutelare i diritti del paziente e degli eredi, supportare eventuali accertamenti giudiziari e ricostruire percorsi assistenziali anche a distanza di molti anni. La sua conservazione illimitata riflette, quindi, l'esigenza di garantire continuità documentale e memoria clinica e storica, oltre a rispondere a finalità di interesse pubblico e scientifico.

Con l'affermarsi di una sanità digitale sempre più orientata ai dati, è importante prendere in considerazione anche la normativa concernente la gestione e la conservazione dei dati personali e particolari. In materia interviene il (Regolamento UE 2016/179 art. 5) che, introducendo il principio di minimizzazione, stabilisce che non possono essere raccolti più dati di quanti ne siano effettivamente necessari e questi devono essere mantenuti solo per il tempo necessario per raggiungere le finalità del trattamento. La sfida che si prospetta è, dunque, quella di gestire l'intero ciclo di vita delle cartelle cliniche nel pieno

rispetto della normativa in materia di privacy, assicurando che il trattamento di questi dati particolari avvenga sempre nel rispetto delle finalità specifiche per cui sono raccolti e vengano utilizzati nella misura strettamente necessaria (Ferraro 2021).

Inoltre, le Linee guida di attuazione del Fascicolo Sanitario Elettronico 2.0 (Agenzia per l'Italia Digitale 2022) stabiliscono che i dati clinici destinati a confluire nel FSE, in futuro, dovranno essere acquisiti direttamente dai sistemi aziendali produttori e presentarsi già nativamente strutturati secondo lo standard HL7 FHIR[3] oppure, se necessario, esservi mappati. In questo caso sarà fondamentale verificare che i sistemi utilizzati dalle strutture sanitarie rispettino le regole semantiche e sintattiche previste dallo standard e che i dati vengano convertiti nel formato HL7 FHIR quando non siano già prodotti nativamente secondo tale modello. Invece, i documenti prodotti dovranno essere conformi al formato Clinical Document Architecture 2 HL7 (CDA 2 HL7) e essere iniettati nei corrispondenti PDF in formato PaDES. All'interno delle cartelle cliniche native digitali, i documenti e i dati devono essere corredati da opportuni metadati così da poter garantire la reperibilità e la valorizzazione in vista della conservazione a lungo termine. Una corretta compilazione dei metadati assicura la coerenza e la correttezza formale dei documenti e dei dati, che possono essere utilizzati anche per effettuare ricerche statistiche, scientifiche e di altro tipo (Guaglianone et al. 2022)

6. Il valore probatorio della cartella clinica elettronica e il rispetto della privacy

La giurisprudenza riconosce la cartella clinica come atto pubblico di fede privilegiata che, indipendentemente dalla tipologia di supporto utilizzato, possiede una duplice finalità: sanitaria e giuridica. La finalità sanitaria è legata alla sua natura di attestazione della diagnosi e di contenitore di informazioni utili per il trattamento sanitario del paziente. La finalità giuridica, invece, è legata alla sua valenza probatoria poiché documento redatto da un pubblico ufficiale nell'esercizio delle sue funzioni[4]. In quanto atto di fede privilegiata, la cartella clinica ha un valore probatorio contestabile solo con querela di falso e questo aspetto è particolarmente significativo in situazioni di contenzioso, in cui si attesta come mezzo di tutela di diritti e di obblighi per lo Stato, per il paziente

[3] HL7 FHIR, acronimo di *Health Level Seven - Fast Healthcare Interoperability Resources*, è uno standard ideato e sviluppato da HL7 International per rispondere alle nuove sfide della sanità digitale, tra cui il supporto automatizzato alle decisioni cliniche e l'elaborazione delle informazioni attraverso sistemi informativi.

[4] Codice Civile, art. 2699 "Atto pubblico": «L'atto pubblico è il documento redatto, con le richieste formalità, da un notaio o da altro pubblico ufficiale autorizzato ad attribuirgli pubblica fede nel luogo dove l'atto è formato».

e per la struttura sanitaria (Zagra et al. 2011). La sua adeguata compilazione riveste, quindi, una grande importanza nella formulazione di un giudizio di responsabilità medica.

Ogni singola annotazione ha autonoma efficacia probatoria dal momento stesso in cui viene inserita nella cartella clinica dai professionisti sanitari che intervengono nel processo di cura. In fase di redazione vi è, quindi, l'obbligo di inserire tutti gli atti diagnostici e terapeutici, senza possibilità di modifiche successive. Secondo quanto stabilito dal Codice di deontologia medica del 2014 all'art. 26 (Federazione Nazionale Ordini Medici Chirurghi e Odontoiatri 2014):

> il medico redige la cartella clinica, quale documento essenziale dell'evento ricovero, con completezza, chiarezza e diligenza e ne tutela la riservatezza; le eventuali correzioni vanno motivate e sottoscritte. Il medico riporta nella cartella clinica i dati anamnestici e quelli obiettivi relativi alla condizione clinica e alle attività diagnostico-terapeutiche a tal fine praticate; registra il decorso clinico assistenziale nel suo contestuale manifestarsi o nell'eventuale pianificazione anticipata delle cure nel caso di paziente con malattia progressiva, garantendo la tracciabilità della sua redazione.

La cartella clinica svolge quindi il ruolo di attestazione puntuale e contestuale delle informazioni relative allo stato di salute di un paziente e deve possedere degli specifici requisiti formali.

Nell'analogico e nel digitale è ugualmente necessario che ogni annotazione e ogni azione clinica siano riconducili al suo autore grazie all'apposizione della firma. Tuttavia, in quanto documenti informatici, le CCE devono essere firmate con firma digitale, firma elettronica qualificata o firma elettronica avanzata per soddisfare il requisito della forma scritta e mantenere l'efficacia prevista dall'articolo 2702 del Codice Civile[5] (Decreto Legislativo 7 marzo 2005, n. 82, art. 20 comma 1 bis). Un documento informatico cui è apposta una firma elettronica soddisfa il requisito della forma scritta e sul piano probatorio è liberamente valutabile in giudizio, sulla base delle sue caratteristiche oggettive di qualità, sicurezza, integrità e immodificabilità (Decreto Legislativo 7 marzo 2005, n. 82, art. 21 comma 1). I medici devono necessariamente sottoscrivere ogni operazione effettuata nel sistema, provvedendo a completare la propria identificazione informatica attraverso appositi strumenti di firma.

Le CCE vengono alimentate dal personale sanitario nell'erogazione delle cure, senza che il paziente sia tenuto a fornire il proprio consenso all'alimentazione. Analogamente, quando un medico deve consultarne i contenuti, a prescindere dalla sua qualifica di libero professionista o di operatore di una

[5] Codice Civile, art. 2702 "Efficacia della scrittura privata": «La scrittura privata fa piena prova, fino a querela di falso, della provenienza delle dichiarazioni da chi l'ha sottoscritta, se colui contro il quale la scrittura è prodotta ne riconosce la sottoscrizione, ovvero se questa è legalmente considerata come riconosciuta».

struttura sanitaria pubblica o privata, in virtù del segreto professionale che è tenuto a rispettare non deve sempre richiedere il consenso del paziente, poiché il segreto professionale viene riconosciuto come una misura adeguata a tutelare il diritto alla riservatezza dei pazienti (Regolamento UE 2016/179 art. 9, par. 3). Tale possibilità di fare a meno del consenso è valida solo per i trattamenti necessari per finalità di cura e inerenti alla salute, «per tutelare un interesse vitale dell'interessato o di un'altra persona fisica che si trovi nell'impossibilità di prestare il consenso», per motivi di interesse pubblico nel settore della sanità pubblica, come la protezione da gravi minacce per la salute a carattere transfrontaliero o la garanzia di parametri elevati di qualità e sicurezza dell'assistenza sanitaria; a fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici (Regolamento UE 2016/179 art. 9, par. 2).

Il diritto del paziente di accedere ai propri dati personali, inclusi quelli relativi alla propria salute, è un principio fondamentale sancito dal GDPR all'articolo 15, che specifica il diritto dell'interessato a ottenere conferma del trattamento dei suoi dati personali e a ricevere una copia di tali dati. L'assistito può richiedere e ottenere la riproduzione fedele ed integrale della propria cartella clinica, dei dati e dei documenti in essa contenuti, in qualsiasi momento ne abbia comprovata necessità; tuttavia, secondo l'articolo 12, paragrafo 5, del GDPR, solo la prima copia della cartella clinica deve essere fornita gratuitamente al paziente. La struttura sanitaria può addebitare un costo solo per eventuali copie aggiuntive o nel caso in cui la richiesta sia considerata manifestamente infondata o eccessiva.

7. Conclusioni

Il presente contributo mira ad analizzare e descrivere lo strumento sanitario della CCE nel panorama italiano, individuando alcune delle principali cause della comprovata entropia che caratterizza i sistemi informativi sanitari. La principale limitazione di queste analisi, tuttavia, risiede nella difficoltà di reperire dati aggiornati e completi sullo stato di adozione delle CCE e di implementazione dei SCCE sia a livello nazionale che europeo. Si rileva, infatti, una notevole scarsità di fonti sul tema e una certa difficoltà nell'accedere a dati raccolti sistematicamente e affidabili.

Nonostante queste limitazioni, il lavoro offre un quadro analitico delle principali iniziative europee finalizzate a promuovere lo scambio di dati sanitari e l'interoperabilità transfrontaliera, in relazione allo sviluppo e all'adozione delle CCE. Maggiore spazio è riservato alla descrizione del contesto italiano, approfondendo: la lunga fase di transizione dal cartaceo al digitale nei flussi documentali clinici; le differenze tra le fasi di gestione e conservazione delle cartelle cliniche analogiche e digitali; la funzione giuridica della cartella clinica e il valore probatorio dei documenti in essa contenuta; e, infine, le disposizioni

normative sulla conservazione a lungo termine e in materia di privacy relative alle cartelle cliniche, nonché ai dati e ai documenti in esse contenute.

In un panorama estremamente frammentato, che vede tuttavia alcuni esempi virtuosi all'interno del contesto nazionale italiano, l'ipotesi di istituire modelli di CCE a livello regionale o specialistico può rappresentare un primo passo significativo per il raggiungimento di un adeguato livello di uniformità locale, da cui partire per potenziare successivamente l'interoperabilità a livello nazionale. La definizione di un set di dati condiviso, volto a semplificare le modalità di compilazione e a sviluppare interfacce comuni, potrebbe costituire una tappa fondamentale verso la standardizzazione dei contenuti delle CCE. In questa direzione, le diverse Implementation Guide HL7 FHIR sviluppate per la codifica e la strutturazione di alcune tipologie documentali, come ad esempio i referti di laboratorio, possono rappresentare un prezioso punto di partenza per garantire l'uniformità dei contenuti delle CCE e la loro standardizzazione secondo criteri condivisi a livello nazionale ed internazionale. L'applicazione di criteri uniformi nella fase di implementazione e concettualizzazione delle CCE e dei SCCE consentirebbe, inoltre, di migliorare l'erogazione delle cure e agevolare le fasi di compilazione, disponendo di interfacce simili.

Ciò che emerge con chiarezza è l'esigenza di proseguire con gli sforzi intrapresi fino ad ora per evitare che la reticenza culturale e metodologica al cambiamento, spesso radicata in modalità operative ancora fortemente ancorate all'analogico, porti a vanificare gli investimenti nella sanità digitale, che, seppur onerosi in una fase iniziale, possono produrre risultati significativi e duraturi nel lungo periodo, se adeguatamente valorizzati.

Riferimenti bibliografici

Adler-Milstein, Julia, Jordan Everson, and Shoou-Yih D. Lee. 2015. "EHR Adoption and Hospital Performance: Time-Related Effects." *Health Services Research* 50 (6): 1751-71. https://doi.org/10.1111/1475-6773.12406.

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.

Akindele, Akinade, and Opeyemi Soyemi. 2021. "Use of Electronic Health Records in the USA: A Lesson for Nigeria." *Library Philosophy and Practice* (e-journal). https://digitalcommons.unl.edu/libphilprac/5486.

Bogaert, Petronille, Marieke Verschuuren, Herman Van Oyen, and Hans van Oers. 2021. "Identifying Common Enablers and Barriers in European Health Information Systems." *Health Policy* 125 (12): 1517-26. https://doi.org/10.1016/j.healthpol.2021.09.006.

Brizzi, Ferdinando. 2021. *Dati sanitari, GDPR e COVID-19: il caso della ricerca: tra scienza e diritto*. Key Editore.

Brown, Janie, Nicole Pope, Anna Maria Bosco, Jaci Mason, and Alani Morgan. 2020. "Issues Affecting Nurses' Capability to Use Digital Technology at Work: An Integrative Review." *Journal of Clinical Nursing* 29 (15-16): 2801-19. https://doi.org/10.1111/jocn.15321.

Campanella, Paolo, Emanuela Lovato, Claudio Marone, et al. 2016. "The Impact of Electronic Health Records on Healthcare Quality: A Systematic Review and Meta-Analysis." *European Journal of Public Health* 26 (1): 60-64. https://doi.org/10.1093/eurpub/ckv122.

Chao, Weng Chi, Hao Hu, Carolina Oi Lam Ung, and Yong Cai. 2013. "Benefits and Challenges of Electronic Health Record System on Stakeholders: A Qualitative Study of Outpatient Physicians." *Journal of Medical Systems* 37 (4): 9960. https://doi.org/10.1007/s10916-013-9960-5.

Cresswell, Kathrin, Stuart Anderson, Catherine Montgomery, Christopher J. Weir, Marek Atter, and Robin Williams. 2023. "Evaluation of Digitalisation in Healthcare and the Quantification of the 'Unmeasurable'." *Journal of General Internal Medicine* 38 (16): 3610-15. https://doi.org/10.1007/s11606-023-08405-y.

Decreto del Presidente della Repubblica 27 Marzo 1969, n. 128. "Ordinamento Interno Dei Servizi Ospedalieri." https://www.edizionieuropee.it/law/html/50/zn86_11_019.html#_ART7_.

Decreto 7 settembre 2023. "Fascicolo sanitario elettronico 2.0." Gazzetta Ufficiale no. 249, 24 ottobre 2023. https://www.gazzettaufficiale.it/eli/id/2023/10/24/23A05829/sg.

Decreto del Presidente del Consiglio dei Ministri 29 settembre 2015 n. 178. "Regolamento in materia di fascicolo sanitario elettronico." Gazzetta Ufficiale no. 263, 11 novembre 2015. https://www.gazzettaufficiale.it/eli/id/2015/11/11/15G00192/sg.

Decreto legislativo 25 gennaio 2010, n. 37. "Attuazione della direttiva 2007/47/CE che modifica le direttive 90/385/CEE per il ravvicinamento delle legislazioni degli stati membri relative ai dispositivi medici impiantabili attivi, 93/42/CE concernente i dispositivi medici e 98/8/CE relativa all'immissione sul mercato dei biocidi." Gazzetta Ufficiale no. 60, 13 marzo 2010. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2010;037.

Decreto Legislativo 7 Marzo 2005, n. 82. "Codice dell'Amministrazione Digitale." Gazzetta Ufficiale no. 112, 16 maggio 2005, Suppl. Ordinario no. 93. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-03-07;82.

Decreto-legge 18 ottobre 2012, n. 179 recante: "Ulteriori misure urgenti per la crescita del Paese." Gazzetta Ufficiale no. 294, 18 dicembre 2012, Suppl. Ordinario no. 208. https://www.gazzettaufficiale.it/eli/id/2012/12/18/12A13277/sg.

Department of Health & Human Services. 2008. *Report to the Office of the National Coordinator for Health Information Technology on Defining Key Health Information Technology Terms*, 28 Aprile 2008.

Direttiva 2011/24/UE del Parlamento europeo e del Consiglio, del 9 marzo 2011, concernente l'applicazione dei diritti dei pazienti relativi all'assistenza sanitaria transfrontaliera. https://eur-lex.europa.eu/eli/dir/2011/24/oj.

European Commission. 2024. "EU4Health Programme 2021-2027 – a Vision for a Healthier European Union." https://health.ec.europa.eu/funding/eu4health-programme-2021-2027-vision-healthier-european-union_en.

European Union. 2023. "XpanDH Project." https://xpandh-project.iscte-iul.pt/.

Federazione Nazionale Ordini Medici Chirurghi e Odontoiatri. 2014. "Codice di Deontologia Medica." https://www.omceo.bg.it/ordine/deontologia-e-normativa/il-codice-deontologico.html.

Ferraro, Luigi. 2021. "Il Regolamento UE 2016/679 tra Fascicolo Sanitario Elettronico e Cartella Clinica Elettronica: il trattamento dei dati di salute e l'autodeterminazione informativa della persona", *BioLaw Journal – Rivista di BioDiritto* 4.

Garrett, Peter, and Joshua Seidman. 2011. "EMR vs EHR – What Is the Difference?" *Health IT Buzz*, 4 January 2011. https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference.

Gewald, Heiko, Alicia Núñez, Corinna Gewald, and Leah Vriesman. 2017. "An International Comparison of Factors Inhibiting Physicians' Use of Hospital Information Systems." *CONF-IRM 2017 Proceedings*, May. https://aisel.aisnet.org/confirm2017/1.

Guaglianone, Maria Teresa, Giovanna Aracri, Maria Teresa Chiaravalloti, et al. 2022. "Ensuring the Long-Term Preservation of and Access to the Italian Federated Electronic Health Record." *Applied Sciences* 12 (7): 3304. https://doi.org/10.3390/app12073304.

Guarasci, Roberto. 2012. "Terminologia, semantica e oggetti documentali." In *Il Fascicolo Sanitario Elettronico: Infrastruttura tecnologica e codifica dei dati*, a cura di Isabella Florio, e Maria Teresa Guaglianone.

Holroyd-Leduc, Jayna M., Diane Lorenzetti, Sharon E. Straus, Lindsay Sykes, and Hude Quan. 2011. "The Impact of the Electronic Medical Record on Structure, Process, and Outcomes within Primary Care: A Systematic Review of the Evidence." *Journal of the American Medical Informatics Association* 18 (6): 732-37. https://doi.org/10.1136/amiajnl-2010-000019.

Lenz, Richard, and Manfred Reichert. 2007. "IT Support for Healthcare Processes – Premises, Challenges, Perspectives." *Data & Knowledge Engineering, Business Process Management*, 61 (1): 39-58. https://doi.org/10.1016/j.datak.2006.04.007.

McCullough, Jeffrey S., Michelle Casey, Ira Moscovice, and Shailendra Prasad. 2010. "The Effect of Health Information Technology on Quality in U.S. Hospitals." *Health Affairs* (Project Hope) 29 (4): 647-54. https://doi.org/10.1377/hlthaff.2010.0155.

McDonald, C. J. 1997. "The Barriers to Electronic Medical Record Systems and How to Overcome Them." *Journal of the American Medical Informatics Association* 4 (3): 213-21. https://doi.org/10.1136/jamia.1997.0040213.

Ministero della Salute. 1992. "Linee guida del 17 giugno 1992, la compilazione, la codifica e la gestione della scheda di dimissione ospedaliera istituita ex dm 28.12.1991." https://www.salute.gov.it/portale/documentazione/p6_2_6.jsp?area=68&btnCerca=cerca&iPageNo=3&lingua=italiano.

Ministero della Salute. 2012. "*CPI. Sviluppo di un modello cartella paziente integrata*" https://www.quotidianosanita.it/allegati/allegato1178142.pdf.

OECD. 2020. *Health at a Glance: Europe 2020: State of Health in the EU Cycle. Paris: Organisation for Economic Co-operation and Development*. https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2020_82129230-en.

Orhan, Mustafa, and Mustafa Kafes. 2021. "Systematic Review on Reengineering Digital Processes of Healthcare Institutions." *Turkiye Klinikleri Journal of Health Sciences* 6 (Novembre): 973-84. https://doi.org/10.5336/healthsci.2020-80883.

Orizio, Luca, e Giovanna Gotti. 2016. "La percezione del cambiamento durante il processo d'introduzione della cartella clinica informatizzata: indagine in terapia intensiva." *Scenario® - Il Nursing nella sopravvivenza* 33 (1): 37-39. https://doi.org/10.4081/scenario.2016.58.

Politecnico di Milano. 2022. "Sanità Digitale: nel 2022 aumentano gli investimenti." Consultato il 5 Aprile 2024. https://www.osservatori.net/it/ricerche/comunicati-stampa/sanita-digitale-pnrr-italia.

Raccomandazione (UE) 2019/243 relativa a un formato europeo di scambio delle cartelle cliniche elettroniche. 2019. https://eur-lex.europa.eu/eli/reco/2019/243/oj

Regione Emilia-Romagna. 2016. "I file della Cartella clinica integrata. Salute." https://salute.regione.emilia-romagna.it/assistenza-ospedaliera/file-cci/file-cartella-clinica-integrata.

Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali e alla libera circolazione di tali dati (Regolamento generale sulla protezione dei dati – GDPR). Versione aggiornata con considerando e rettifiche pubblicate nella Gazzetta ufficiale dell'Unione europea L 127 del 23 maggio 2018. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/6264597.

Roger France, Francis. 2014. "About the Beginnings of Medical Informatics in Europe." *Acta Informatica Medica* 22 (1): 11-15. https://doi.org/10.5455/aim.2014.22.11-15.

Sipido, Karin R., Fernando Antoñanzas, Julio Celis, et al. 2020. "Overcoming Fragmentation of Health Research in Europe: Lessons from COVID-19." *The Lancet* 395 (10242): 1970–71. https://doi.org/10.1016/S0140-6736(20)31411-2.

Sironi, Chiara. 2025. "Il futuro della sanità è sempre più digitale." *Tendenze*, 5 Giugno 2025. https://tendenzeonline.info/articoli/2025/06/05/il-futuro-della-sanit--sempre-pi-digitale/.

Sist, Luisa, Mariella Dammiano, Chiara Donati, et al. 2022. "La cartella clinica elettronica in ambito pediatrico." https://www.infermiereonline.org/2022/03/06/la-cartella-clinica-elettronica-in-ambito-pediatrico/

Società Italiana di Diabetologia (SID). 2024. "I medici italiani sono i più anziani d'Europa." Policy Brief. no. 2, Febbraio 2024 https://www.siditalia.it/pdf/policy-brief-SID---2-2024.pdf.

Sorrentino, Elisa, Maria Teresa Guaglianone, Elena Cardillo, Maria Teresa Chiaravalloti, Anna Federica Spagnuolo, and Giuseppe Alfredo Cavarretta. 2020. "La conservazione dei documenti che alimentano il Fascicolo Sanitario Elettronico." *Rivista italiana di informatica e diritto* 2 (1): 35-42. https://doi.org/10.32091/RIID0018.

Vehko, Tuulikki, Hannele Hyppönen, Sampsa Puttonen, et al. 2019. "Experienced Time Pressure and Stress: Electronic Health Records Usability and Information Technology Competence Play a Role." *BMC Medical Informatics and Decision Making* 19 (1): 160. https://doi.org/10.1186/s12911-019-0891-z.

Windle, John R., and Thomas A. Windle. 2015. "Electronic Health Records and the Quest to Achieve the 'Triple Aim'." *Journal of the American College of Cardiology* 65 (18): 1973-75. https://doi.org/10.1016/j.jacc.2015.03.038.

Zagra, Michele, Antonina Argo, and Stefania Zerbo. 2011. "Cartella Clinica." In *Medicina legale orientata per problemi*, a cura di Michele Zagra, Antonina Argo, Burkhard Madea e Paolo Procaccianti. Elsevier.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-549-3
DOI 10.57574/59655493XX
pag. 89-124 (gennaio-giugno 2025)
Open access article licensed under CC-BY-NC-SA

Schiavitù e *invented archives*. Intelligenza Artificiale generativa nella costruzione del *Database on the Slave Trade*

Salvatore Spina*

Abstract: This essay examines the construction of the *Database on the Slave Trade between the Mediterranean and the Atlantic (15th–16th centuries)* as a case study to explore the transformative role of generative Artificial Intelligence tools in historical and archival research. Positioned within the framework of *histoire sérielle*, the project addresses methodological challenges related to the datafication of sources on Mediterranean and Atlantic slavery, showing how *invented archives* function as epistemic spaces where memory, computation, and interpretation intersect. The use of LLMs (Large Language Models) for data extraction, encoding, and modeling raises critical issues concerning scientific validity, algorithmic opacity, and the indispensable role of human oversight. The essay offers a critical reflection on the integration of Digitality and Historiography, highlighting both the limitations and potentials of emerging technologies in transforming historical documents into queryable *cybertexts* and in fostering the collaborative construction of historical memory.
La ricerca si è avvalsa di un finanziamento dell'Unione europea – Next Generation EU – missione 4, componente 2, investimento 1.1, nell'ambito del programma PRIN-PNRR. Il titolo del progetto è: A Database on the Slave Trade between the Mediterranean and the Atlantic (15th-16th centuries).

Keywords: Histoire sérielle, ChatGPT, AI Models, Ollama, Chatbox.

1. Introduzione

La nascita del World Wide Web, nel 1991[1], ha segnato la "storia" della Storia e dell'Archivistica, non solo per la creazione di una nuova *heimat* comunicativa (Ciofalo e Leonzi 2013), della "Nicchia Ecologica Digitale" e della "Digitalità" (Spina 2024), ma per aver aperto un percorso in cui si faceva reale la possibilità di concretizzare progetti di valorizzazione archivistica, fino a quel momento inimmaginabili. Certamente, non tutto si è mosso con linearità; ma

* Dipartimento di Scienze Umanistiche, Università degli Studi di Catania, Catania, Italia. salvatore.spina@unict.it. ORCID: 0000-0001-6367-8183.

[1] Il 6 agosto 1991 veniva lanciato il primo sito web della storia, dal fisico britannico Tim Berners-Lee – proprio da colui che ha pensato, progettato e realizzato il World Wide Web.

fu da subito evidente che, da quella data – che ha visto il primo sito andare online (World Wide Web, n.d.), col quale si dava libero accesso, agli utenti del Web, alle idee e alle caratteristiche proprie dell'"invenzione" del "WWW" –, tutto quello che veniva "uploadato" avrebbe avuto poco a che vedere con il concetto tradizionale di "archivio". Si era, e lo si è ancora, di fronte a qualcosa di concettualmente nuovo: gli *invented archives.*

Grazie al Web, e alla costante implementazione di siti, è stato – quindi – possibile racchiudere documenti archivistici, con la finalità di dare corpo "digitale" a progetti dalla doppia anima, storica e archivistica; a idee di ricerca, o semplicemente ad espressioni di passioni personali – il "Mosaic Netscape" e il "Netscape Navigator" (Figg. 1, 2) (era già il 1994) diedero, infatti, a tutti la possibilità di creare le proprie collezioni, e di fare del Web una struttura informativa globale, superando le aspettative di Vannevar Bush e del suo *Memex* (Bush 1945).

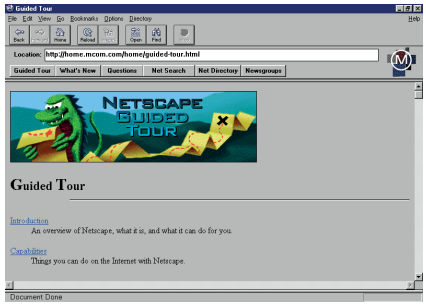


Figura 1. Portale Web "Mosaic Netscape".

Figura 2. Il browser Netscape Navigator.

Questa nuova frontiera ha segnato le prime fasi della storia "informatica" dell'Archivistica, la quale guardava attentamente alla possibilità di mettere online milioni di registri, indici e inventari, che fino a quel momento avevano guidato gli utenti alla ricerca dei documenti. Contestualmente, *a latere*, studiosi e gruppi di ricerca, grazie al Web, davano vita ad una nuova modalità di "far archivio e Storia", che obbligò gli archivisti a fare i conti con una realtà documentaria sempre più interconnessa[2], che affiancò il concetto di "soggetto collettore/creatore" a quello di "soggetto produttore", e il "vincolo tematico" al "vincolo archivistico".

A partire dai primi anni del Ventunesimo secolo, il Web si "dota" di numerosi portali e aggregazioni tematiche, che si costruiscono intorno a svariati filoni e oggetti di ricerca (Trinkle e Merriman 2000); raccolte effettuate da singoli studiosi (professionisti e non), da comunità di ricercatori, da istituzioni accademiche o di ricerca (Rosenzweig 1997; 2001), la cui idea, di certo, ci

[2] Solamente per uno spunto, consiglio la lettura di Xu e Zhang (2022); Gonzalez et al. (2014); Reilly et al. (2000); Clerici e Pra (2012).

porta necessariamente, e per diversi motivi, a guardare all'*histoire sérielle*, ossia a rifocalizzarsi dentro quell'approccio che fondava, negli anni Cinquanta del Novecento, la sua metodologia sull'individuazione di determinati documenti archivistici, e il loro inserimento in una *serie* definita, allo scopo di spiegare un fenomeno storico – che diventava, per Furet (1971; 1981), condizione necessaria per il venire in essere della Storia Quantitativa (Darcy e Rohrs 1995; Dollar e Jensen 1974; Furet 1968; Itzcovich 1996).

2. *History of Slavery*. I dataset e gli archivi inventati come *histoire sérielle*?

L'introduzione dell'espressione *histoire sérielle*, in Storiografia (contemporanea), si deve alla figura di Pierre Chaunu, che ne ha elaborato e teorizzato la portata metodologica principalmente attraverso la sua opera fondamentale *Séville et l'Atlantique, 1504–1650*, pubblicata nel 1959, con la quale proponeva un radicale cambiamento epistemologico, spostando il focus della ricerca storica dall'"evento singolare e irripetibile" all'analisi di "fenomeni ripetitivi", seriali, e collocati all'interno di strutture coerenti e persistenti nel tempo, in cui la ripetizione non è più considerata un limite, ma il punto di partenza per l'interpretazione storica (Furet 1981, 6) – ed è certamente, all'interno di questa cornice teorica, che si colloca la genealogia degli *invented archives*, che hanno consentito di "materializzare" il concetto di "serialità".

Nel concreto della ricerca storica, ogni studioso "lavora nella serialità", operando inevitabilmente selezioni e organizzando le informazioni, isolando dati rilevanti e integrandoli in interpretazioni storiografiche specifiche. Questo processo, che può spaziare dall'analisi di eventi singolari, sino alla *longue durée* braudeliana (Braudel 2003; 2015; Braudel e Salsano 1974), richiede sempre una strutturazione logica delle fonti, la quale, se vero che è sempre stata frutto di una relazione "chiusa" (*close reading*) tra lo studioso e le carte, con il crescente processo di digitalizzazione e "matematizzazione" delle fonti documentarie (Spina 2021), ha richiesto strumenti adeguati, come il database – dove organizzare le sue informazioni –, che diventa indispensabile per lo studioso, il quale può, come evidenziato da Oscar Itzcovich, non solo raccogliere informazioni, ma renderle interrogabili, manipolabili e combinabili in modelli teorici dinamici (Itzcovich 1989; 1993).

"Serialità" e archivi inventati, dunque, si sovrappongono, per dare corpo ad un ambiente digitale appositamente realizzato per raccogliere e rendere accessibili documenti, originariamente disseminati in molteplici archivi fisici, e consentire indagini analitiche profonde e comparative, in grado di rendere visibile l'"invisibile" – come afferma Nowatzki (2021) –, così come nel caso del fenomeno di "lunga durata" della schiavitù, che è stato oggetto di numerose "esperienze

digitali", quali il "North American Slave Narratives" (n.d.), ovvero i progetti "Slave Revolt in Jamaica" (Brown 2021), "Legacy of Slavery in Maryland" (The Maryland State Archives, n.d.), e l'"Early Caribbean Digital Archive" (Northeastern University Library, n. d.), "The Atlantic Slave Trade in Two Minutes" (National African-American Reparations Commission, n.d.), o, ancora, il "BlackDH Schema Project" (Lu e Pollock 2019), che propone un ripensamento radicale del *Text Encoding Initiative* (TEI), troppo neutro e distante dalle necessità di una rappresentazione dei soggetti afrodiscendenti, quindi, in chiave inclusiva.

Di rilevanza internazionale, sicuramente, il "Transatlantic Slave Trade Database" (oggi, Slave Voyages, n.d.) ha rivestito – e riveste – il ruolo di "progetto pilota" per tutte le altre esperienze successive.

Online dal 1999, il portale è frutto di anni di ricerca intrapresa da studiosi diversi, già a partire dagli anni Settanta del Novecento.

Già qualche anno prima, però, nel 1993, documenti relativi al fenomeno della schiavitù erano presenti nella rete, tra i dati raccolti nel sito "The Valley of Shadow", ideato dallo storico Edward L. Ayers nel 1991 – progetto che nasce all'interno del Virginia Center for Digital History[3] (Università della Virginia), allo scopo di condurre un'indagine storica, mai tentata prima, con l'ausilio di strumenti tecnologici (Ayers e Rubin 2000). Qui confluisce il "Free Black Registry (1803-1865)", un apparato documentario che testimonia le dinamiche sociali di una realtà schiavista come quella della Virginia nella prima metà del XIX secolo.

Nell'ottobre del 2007, invece, presso il Nebraska Digital Workshop, tenutosi presso il Center for Digital Research in the Humanities (CDRH) dell'Università del Nebraska, Andrew Torget presenta il "Texas Slavery Project", un portale frutto di un lavoro metodologico statistico che si proponeva di stimare i dati demografici mancanti relativi alla popolazione schiavizzata e ai possessori di schiavi nei territori della Repubblica del Texas nel periodo compreso tra il 1837 e il 1845. Operando su un insieme di 30 contee storiche texane, il progetto ha elaborato un modello predittivo basato sull'analisi comparativa delle tendenze demografiche nei contesti in cui le fonti risultano complete, applicando tali modelli ai casi lacunosi, per arrivare al numero "reale" di schiavi detenuti.

Nel 2011, trova la sua genesi il progetto "Enslaved: Peoples of the Historical Slave Trade", che costituisce, oggi, uno degli snodi fondamentali nel panorama delle Digital Humanities applicate allo studio della schiavitù. Il progetto è una evoluzione di "Slave Biographies: The Atlantic Database Network", finanziato dal National Endowment for the Humanities.

Nel dicembre 2014, nasce "Slavery and Remembrance", da una iniziativa congiunta tra UNESCO Slave Route e la Colonial Williamsburg Foundation.

[3] Il Virginia Center for Digital History (VCDH), uno spazio indipendente all'interno del College of Arts and Sciences dell'Università della Virginia, viene fondato nel 1998 da Edward L. Ayers e William G. Thomas III, allo scopo di sostenere progetti di ricerca in grado di sensibilizzare la comunità scientifica all'uso della tecnologia informatica in ambito umanistico.

Il progetto coinvolge più di cinquanta musei, siti storici e istituzioni in Europa, Africa e America, che hanno dato vita ad un sito web multilingue liberamente accessibile, che si concentra su musei e siti di memoria legati alla schiavitù e al commercio degli schiavi, a beneficio degli utenti in generale, degli studiosi, degli studenti e dei professionisti del patrimonio museale e culturale.

L'anno successivo, esattamente il 6 agosto 2015, appare online un nuovo progetto: "Liberated Africans", un portale scaturito da una collaborazione interdisciplinare che fa capo ad un team di sviluppo, il cui lavoro guarda alla possibilità di ripercorrere le vite di circa 225.000 schiavi africani portati via dalle navi negriere, ricostruire le attività di alcune delle prime corti internazionali del mondo, il tutto con profilo metodologico che può inscriversi nel solco della Public History, ossia nell'apertura del team di sviluppo alla comunità mondiale, garantendo la partecipazione di soggetti studiosi del fenomeno.

Contemporaneamente, ma su altri versanti, nel febbraio 2016, viene avviato il "Georgetown Slavery Archive", che costituisce, oggi, uno degli strumenti più avanzati e metodologicamente significativi per lo studio della schiavitù in relazione alle istituzioni religiose e accademiche statunitensi. Pensato e progettato nell'ambito dell'iniziativa "Slavery, Memory, and Reconciliation", promossa dalla Georgetown University, il progetto si configura come un archivio digitale in continuo aggiornamento, volto a raccogliere, conservare e rendere accessibile la documentazione relativa alla storia della schiavitù connessa ai gesuiti del Maryland e alla stessa Università di Georgetown.

Nell'aprile 2025, Internet Archive accoglie i documenti digitalizzati relativi alla Storia della Schiavitù ad Aruba, conservati presso il National Archives of Aruba (ANA) e la National Library of Aruba (BNA). La documentazione consente di percepire le condizioni di vita degli schiavi e dei loro discendenti, contribuendo a far luce sul loro doloroso passato.

Questi dati, digitalizzati, sono stati ufficialmente aggiunti al Memory of the World Program (MoW-AW) dell'UNESCO, e, da quest'anno, confluiti in Internet Archive, garantendo una ulteriore accessibilità, e una maggiore tutela contro l'obsolescenza.

3. Il "Database on the Slave Trade between the Mediterranean and the Atlantic (15th-16th centuries)

Nel 2022, su bando PNRR, presso l'Università di Teramo e Catania prende avvio il progetto di ricerca volto alla creazione di un database dove raccogliere la documentazione utile a spiegare il fenomeno della tratta degli schiavi, tra Mediterraneo e Atlantico, nei secoli Quindicesimo e Sedicesimo.

Principal Investigators delle due unità sono Carlo Taviani (Teramo) e Lina Scalisi (Catania). Sullo sfondo del progetto, la necessità di spiegare quali di-

namiche caratterizzarono il passaggio tra il tardo Medioevo e l'Età Moderna, quando il "sistema-mondo" – così definito da Fernand Braudel (2003; 2017; Braudel e Salsano 1974) – atlantico si configurò progressivamente come uno spazio economico articolato, in cui circolavano, insieme ai metalli preziosi, alle merci e ai capitali finanziari, anche persone libere e schiavizzate. Questi flussi collegavano gli arcipelaghi atlantici, le coste occidentali dell'Africa e il Mediterraneo, inaugurando precoci percorsi del capitalismo moderno. Mentre gli studi storiografici hanno già ampiamente analizzato le dinamiche economiche[4], politiche e sociali dei secoli XVII e XVIII, le fasi del fenomeno "schiavitù" precedenti all'impresa di Colombo sono state "offuscate" dall'interpretazione storica post-scoperta geografica del Nuovo Mondo, quindi, volta alla descrizione del fenomeno strutturato e intriso di violenza, nel consolidamento del complesso concetto "piantagione-schiavitù", nonché di sfruttamento intensivo delle risorse commerciabili, dai metalli alle fibre tessili, dalle specie animali ai prodotti agricoli. In questo quadro, i popoli africani e i territori del continente – in particolare le coste occidentali e le aree retrostanti – furono non semplicemente spettatori, ma elementi costitutivi dell'ingranaggio mercantile atlantico.

Ciò che viene posto in secondo piano, e che deve rivalutarsi come dimensione essenziale al network, è il ruolo del Mediterraneo; in particolare le sue strutture finanziarie e commerciali. Questo "continente liquido" (Braudel 2017), infatti, si innesta nel sistema, già molto tempo prima delle dinamiche atlantiche, e già prima dell'Età Moderna.

Cosa è, dunque, la figura dell'"esploratore" europeo, in Africa subsahariana, e negli arcipelaghi atlantici tra la metà del XV e l'inizio del XVI secolo?

A partire dagli anni Cinquanta del Quattrocento, si assiste a un mutamento rilevante nelle traiettorie del capitale europeo: da un sistema centrato sull'asse orientale – con il Levante e il Mar Nero come poli strategici – si passa a un orientamento occidentale che privilegia l'Iberia, il Maghreb e i margini atlantici. In questo ridirezionamento dei circuiti economici, un ruolo cruciale fu assunto dai mercanti-banchieri genovesi, i quali, attraverso una rete densa di insediamenti e partecipazioni finanziarie, si posizionarono strategicamente in tutta l'area iberica e negli arcipelaghi atlantici, finanche in Sicilia (Rau 1957; Verlinden 1970; Sicking e Wijffels 2020; Trasselli 1969; Gioffrè 1971; Bonazza 2023; Armenteros-Martinez 2022; Trasselli 1972; Gioffrè 1962; Schwartz 2004). Arcipelaghi come Madeira e le Canarie divennero poli produttivi centrati sulla coltivazione dello zucchero e sull'impiego intensivo della manodopera servile, mentre Capo Verde si affermò sin dai primissimi decenni come nodo fondamentale per la tratta atlantica degli schiavi. Pur essendo stato studiato il fenomeno dell'apertura dell'Atlantico, soltanto recentemente

[4] Si vedano, tra la letteratura più recente: Delpiano (2021); Fiume (2012); Bono (2017); Sgroi (1979); Eltis e Richardson (2008).

la Storiografia ha iniziato a tracciare connessioni sistematiche tra mercati differenti e contesti sovranazionali, adottando approcci metodologici innovativi e sfruttando database digitali. Inoltre, sebbene la tratta transatlantica degli schiavi costituisca, oggi, un campo storiografico consolidato, le sue origini più remote – e in particolare tra Medioevo e prima età moderna – restano ancora poco investigate, soprattutto per quanto riguarda le sue premesse finanziarie, logistiche e infrastrutturali.

È in questa direzione che si inserisce il progetto per la costruzione del "Database on the Slave Trade between the Mediterranean and the Atlantic (15th-16th centuries)"; finalizzato allo studio della tratta atlantica nelle sue fasi iniziali, con particolare attenzione ai nodi finanziari genovesi, che resero possibile la prima apertura commerciale dell'Atlantico.

L'obiettivo è quello di adottare strutture semantiche di modellazione dei dati, già sperimentate con successo nel settore dei Beni Culturali, per garantire la qualità FAIR (rintracciabili, accessibili, interoperabili e riutilizzabili) (Digital Library 2021) delle informazioni archiviate, ed assicurare la fedeltà epistemologica dei dati storici, e rendere possibile una rappresentazione accurata del fenomeno all'interno di ambienti computazionali, favorendo la condivisione delle fonti e dei risultati tra diversi gruppi di ricerca.

Alla base del database, troviamo la piattaforma open-source "Arches", sviluppata dal Getty Conservation Institute e dal World Monuments Fund – che riunisce professionisti del patrimonio e sviluppatori di software provenienti da tutto il mondo – allo scopo di creare un sistema che consenta di creare database in cui riunire dati che possono "spiegare" un patrimonio culturale, inventari digitali, e altri tipologie di prodotti, attraverso diversi sistemi integrati di visualizzazione dei dati, finanche la georeferenziazione delle informazioni.

Il progetto ha una durata biennale, per questo, data l'enorme quantità di dati da acquisire e "tradurre" in records del database, è stata valutata – positivamente – la possibilità di utilizzare i Large Language Models, i quali, oggi, nelle "vesti" delle varie piattaforme di *Generative Artificial Intelligence*, consentono di lavorare con maggiore efficacia ed efficienza su grossi volumi di documenti.

Il progresso tecnologico, come già evidenziato nel grande filone metodologico su cui si ergono le *Digital Humanities* e la *Digital History* (Spina 2022a), ha ragionevolmente modificato la maggior parte delle metodologie di ricerca, portando anche le *Humanities* dentro la Nicchia Ecologica Digitale. Ciò si è tradotto in una rivalutazione degli approcci quantitativi, per il maggior controllo che lo studioso può operare sui dati e sulla possibilità di derivare, da essi, dei risultati epistemologicamente rilevanti.

Per la prima volta, dunque, gli storici hanno avuto la possibilità di applicare tools di Data Mining, come Weka, RapidMiner, KNIME, Orange, spaCy (Python), Stanford CoreNLP, e altri ulteriori strumenti per poter estrapolare quanti più dati possibili dalla documentazione storica, la quale,

contestualmente, veniva sottoposta ad un processo di trascrizione automatica, grazie alle IA non generative, come Transkribus (Massot et al. 2019; Milioni 2020; Muehlberger et al. 2019; Rabus 2019; Schlagdenhauffen 2020; Erwin 2020; Kahle et al. 2017; READ-COOP, s.d.; Spina 2022b; 2023a) ed eScriptorium (Gautier et al. 2022; Kiessling et al. 2019).

La documentazione archivistica esce – seppur con grossi limiti – dagli spazi di conservazione, per entrare nella rete sottoforma di "digital twins" (Banfi et al. 2023; Ahmadi-Assalemi et al. 2020; Correia et al. 2023; El Saddik 2018; Fuller et al. 2020; Gabellone 2022; Grieves 2023; Hutson et al. 2023) che cercano di tradurre in formato *machine-readable* quelli che possono diventare i Big Data della Storia (Bail 2014; Eijnatten et al. 2013; Graham et al. 2015; Franzosi 2017; Kaplan e di Lenardo 2017; Kennedy et al. 2017; Santoro 2015; Schiuma e Carlucci 2018; Spina 2020; Ruis e Shaffer 2017).

Così, tra i "ferri" del mestiere di storico, entrano i tools di Named Entity Recognition (NER), come spaCy, Flair, Transformers (Hugging Face), Stanza (Stanford NLP), OpenNLP (Apache).

Strumenti, questi tutti, che hanno realmente cambiato gli approcci quantitativi della ricerca storica, ma che richiedono una competenza informatica matura per consentire alle macchine di poter "interagire" con la documentazione storica.

La realtà della ricerca umanistica, però, è stata totalmente rivoluzionata nell'ultimo biennio, quando quello che è sempre stato il progetto "finale" della Computer Science – la creazione di una macchina "pensate" (McCarthy et al. 2006) – ha trovato una sua forma iniziale: l'Intelligenza Artificiale.

Cambiando totalmente gli schemi di interazione con i testi e i documenti digitali (e digitalizzati), le IA superano la potenza di calcolo degli altri tools – precedentemente menzionati e utilizzati –, aprendo la ricerca (sia umanistica che fisico-naturale) verso dimensioni e opportunità prima non facilmente accessibili.

Ma quali opportunità, realmente, possiamo evidenziare nel rapporto tra ricerca storica, Intelligenza Artificiale e mestiere di storico?

La creazione del nuovo database sulla schiavitù ha dato l'opportunità, al gruppo di ricerca, di testare le potenzialità di queste piattaforme, comprenderne i meccanismi e gettare le basi per la creazione di un sistema di prompt che possa tradursi in momenti di Machine Learning per l'addestramento di una IA storica, rispondendo, in tal senso, alla necessità evidenziata da Ennals (1986) di «dire alla [macchina] cosa noi [storici] facciamo».

3.1. La costruzione del dataset. *Named Entity Recognition* e modelli di IA a confronto

Il processo "digitale" si fonda sulle specifiche dell'architettura del database che si deve costruire, quindi sulla sua "ontologia". Questo aspetto determina a quale procedimento computazionale deve essere sottoposta la documentazione archivistica per essere tradotta in *digital twin* (il gemello digitale). Oppure, optare per la sola estrapolazione dei dati, tralasciando di operare la creazione di edizioni digitali o formati che virtualizzano la consultazione – come, ad esempio, il flipbook (Spina 2023b).

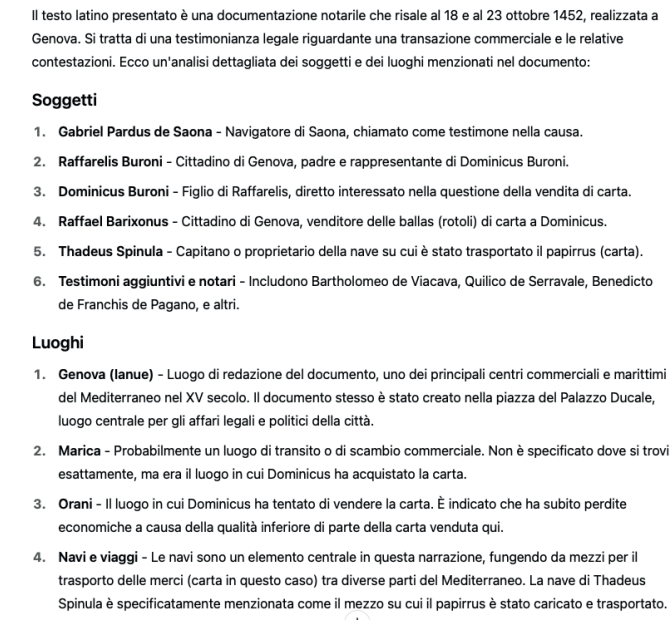
Nel nostro caso, il lavoro di "digitalizzazione" si è fermato alla sola fase di estrapolazione dei dati, senza guardare alla possibilità di inserire le immagini dei documenti all'interno dei vari record creati.

Quindi, il lavoro ha previsto l'uso di strumenti di NER, per l'estrapolazione delle entità nominate (soggetti, luoghi, oggetti), la successiva compilazione di un file CSV con i dati ricavati, e la sua strutturazione per rispondere alle necessità dell'ontologia che sottostà al database.

Tralasciando, fin da subito, l'opzione di utilizzare gli strumenti classici, il ragionamento è stato tutto rivolto su quale LLM utilizzare per l'estrapolazione dei dati, tenendo ben chiaro che le IA generative, pur essendo strumenti potentissimi, in grado di svolgere operazioni diverse – per cui prima si utilizzava un solo tool per ognuna di esse –, senza uscire fuori dalla piattaforma, molte volte producono *outcomes* non esatti, o meglio: allucinati (Maleki et al. 2024; Salvagno et al. 2023).

Questo "bug", soprattutto quando si applicano gli LLMs a complessi documentari molto vasti, scritti in lingue diverse dalle parlate – come il latino, per cui si richiede una specifica competenza –, rischia di creare set di dati non corretti, inficiando il lavoro di ricerca e la messa online di un database le cui informazioni non hanno corrispondenza archivistica e storica con quanto realmente scritto nel documento. Tale fenomeno, ad esempio, si è palesato in una fase di estrapolazione dei dati in cui si è chiesto a ChatGPT-3.5 di effettuare una operazione di NER su un testo latino, tratto da Carlo Taviani (2024, 206), di cui si riporta, qui, solamente la parte interessata: «[…] in dicto loco Orani, in quo loco eum vendidit dictus Dominicus cuidam mauro speciario qui multum frequentat in *terra nigrorum* mercatorio modo; quam venditionem fecit dictus Dominicus mediante ipso teste, ex quo causam habuit videndi dictum papirru, ex quo etiam scit que testificatur. Interrogatus quotenis est, respondit etatis annorum triginta quinque vel circa».

L'outcome dell'IA è il seguente (Fig. 3):



Il testo latino presentato è una documentazione notarile che risale al 18 e al 23 ottobre 1452, realizzata a Genova. Si tratta di una testimonianza legale riguardante una transazione commerciale e le relative contestazioni. Ecco un'analisi dettagliata dei soggetti e dei luoghi menzionati nel documento:

Soggetti

1. **Gabriel Pardus de Saona** - Navigatore di Saona, chiamato come testimone nella causa.
2. **Raffarelis Buroni** - Cittadino di Genova, padre e rappresentante di Dominicus Buroni.
3. **Dominicus Buroni** - Figlio di Raffarelis, diretto interessato nella questione della vendita di carta.
4. **Raffael Barixonus** - Cittadino di Genova, venditore delle ballas (rotoli) di carta a Dominicus.
5. **Thadeus Spinula** - Capitano o proprietario della nave su cui è stato trasportato il papirrus (carta).
6. **Testimoni aggiuntivi e notari** - Includono Bartholomeo de Viacava, Quilico de Serravale, Benedicto de Franchis de Pagano, e altri.

Luoghi

1. **Genova (Ianue)** - Luogo di redazione del documento, uno dei principali centri commerciali e marittimi del Mediterraneo nel XV secolo. Il documento stesso è stato creato nella piazza del Palazzo Ducale, luogo centrale per gli affari legali e politici della città.
2. **Marica** - Probabilmente un luogo di transito o di scambio commerciale. Non è specificato dove si trovi esattamente, ma era il luogo in cui Dominicus ha acquistato la carta.
3. **Orani** - Il luogo in cui Dominicus ha tentato di vendere la carta. È indicato che ha subito perdite economiche a causa della qualità inferiore di parte della carta venduta qui.
4. **Navi e viaggi** - Le navi sono un elemento centrale in questa narrazione, fungendo da mezzi per il trasporto delle merci (carta in questo caso) tra diverse parti del Mediterraneo. La nave di Thadeus Spinula è specificatamente menzionata come il mezzo su cui il papirrus è stato caricato e trasportato.

Figura 3. Outcome ChatGPT 3.5.

Seppur corretti entrambi gli elenchi proposti dall'IA, quello che salta all'attenzione – palesando una "difficoltà" del modello – è l'assenza di una entità nominata, la cui formulazione nel documento non è specificatamente toponomastica, ossia l'espressione *terra nigrorum*, la quale si riferisce all'Africa subsahariana.

Questa è una informazione essenziale, non solo per la valenza storica, quanto per la correttezza del network che devesi creare attraverso i dati del dataset, e la corretta esecuzione della funzione di NER.

La considerazione logica, quindi, è che la "black box", che sottostà all'IA, non contiene informazioni storiche complesse, ma soprattutto, complete. ChatGPT, così come le altre IA, è stata allenata su dataset costruiti sulla base delle informazioni storiche presenti su Wikipedia e su altri siti internet "accessibili" alle macchine. Non esiste, dentro questi modelli, alcun dato storico ricavato dall'enorme e complesso patrimonio storiografico, esegetico, interpretativo, su cui si fonda il "mestiere di storico" (Bloch 1963).

Tale assenza può essere, dunque, colmata solamente con un corretto *training* che coinvolga gli esperti – in questo caso, gli storici –, al fine di "formare" efficacemente l'IA. Nel nostro caso, la semplice istruzione (prompt) «Terra nigrorum = Africa» ha consentito all'IA di effettuare correttamente l'estrapolazione e correggere l'elenco delle entità (Fig. 4).

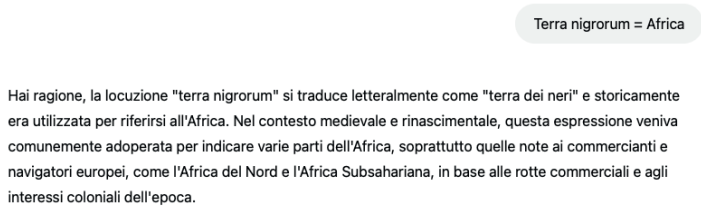


Figura 4. Training attraverso prompt.

Questo aspetto, se vero che mette in evidenza i limiti dei modelli, è ancor più vero che sottolinea la necessità di un controllo costante sul flusso di "ragionamento" dell'IA. Ad oggi, non esistono strumenti per accertare il lavoro della macchina. Agli operatori e agli studiosi è dato l'onere di esaminare continuamente i risultati a cui perviene un LLM. Tutto quello che esso produce, in termini di *outcome*, infatti, non è "costantemente" corretto, *in primis* per il fatto che l'IA ed i modelli si basano su *black boxes* che possono essere descritte come "ambienti stocastici", i quali cercano di portare dentro questa tecnologia un principio basilare dell'"esistenza", ossia quel «dynamical behavior of each process (biological, industrial, economical, cultural) that is not stationary but evolves with the passage of time» (Wang et al. 2023) – si vedano, inoltre, anche Galić et al. (2023); El Haoud e Bachiri (2019). Questo procedimento punterebbe a rendere il processo computazionale sempre più simile a quello umano – progetto a cui OpenAI guarda con i suoi modelli di "reasoning" (OpenAI o3 e o4-mini), i quali vengono allenati per produrre *outcomes* dettagliati e per risolvere problemi complessi –, 'contemplando' la possibilità di migliorare, ad ogni step, le risposte fornite dall'IA. Tutto questo si traduce in un ambiente il cui meccanismo è incomprensibile, e che nei modelli più piccoli comporta che, date specifiche istruzioni, l'IA non le esegue in *routine*. Se al primo prompt, infatti, il sistema risponde efficacemente, la sua capacità computazionale cade di prestazione sui nuovi processi di NER. Per superare tale limite, occorre, infatti, dare alla macchina il prompt di partenza, per evitare che essa "dimentichi" cosa fare. Tutto questo perché il lavoro di elaborazione comporta, per i modelli (sia in remoto che *online*), un elevatissimo consumo di energia e dati, obbligando le aziende a porre dei limiti specifici sul numero di token che l'utente può utilizzare per far eseguire un procedimento computazionale. Nel caso dei modelli gratuiti di OpenAI, ad esempio, il limite è di 15.000 token, superati i quali, tutto quello che è stato inserito come informazione e/o prompt, viene progressivamente cancellato per far spazio ad un nuovo procedimento, oppure il modello restituire l'errore di bloccare per alcune ore l'inserimento di nuovi prompt.

A questo si aggiungano – per quello che concerne la ricerca storica – due aspetti essenziali: (1) dentro i grandi modelli, l'informazione storica è pressoché nulla. I dati su cui vengono allenate le IA sono ricavati da enciclopedie online, e da qualche sito ad accesso aperto (senza la creazione di profili utente); tutto quello che riguarda la grande letteratura e la Storiografia, dunque, non è stato tradotto in dati per il training – a causa di violazione di copyright sulle pubblicazioni, accesso negato ai file delle pubblicazioni digitali, e altri aspetti che impediscono all'IA di "leggere" testi e documenti storiografici. (2) Le fonti archivistiche (su cui si fonda il lavoro di ricerca dello storico), per il loro "logico" *status* cartaceo, non sono consultabili dalle IA; e anche i grandi progetti di digitalizzazione di materiale d'archivio non guardano ad una adeguata "traduzione" dei testi in linguaggio comprensibile dalla macchina, e quindi non utilizzabile da quest'ultima per rispondere adeguatamente ai prompt che le vengono dati.

Tutto questo "obbliga" gli utenti/ricercatori/storici/archivisti a lavorare a fasi, in quanto, al raggiungimento del limite di token, il modello smette di elaborare, chiedendo all'utente di non immettere nuovi comandi, per un certo lasso di tempo – aumentando gli intervalli nel flusso di lavoro, che mostra tutti i limiti di un approccio digitale *tout court*, per cui, ad esempio, non è possibile effettuare una operazione di NER su un lungo testo, con un unico prompt.

Si è deciso, così, di effettuare dei test per capire quale modello avrebbe consentito di diminuire i tempi delle operazioni di riconoscimento delle entità nominate e compilare un file CSV scaricabile con i dati estratti.

Per operare efficacemente, i vari modelli vengono scaricati in locale, attraverso la piattaforma Ollama, che consente di effettuare il download dei modelli open-access, allo scopo di creare una IA "personalizzata". I test vengono effettuati su macchina MAC STUDIO, con chip Apple M2 Max, con sistema operativo macOS Sequoia 15.4.1, 12 Core, 64 GB di Ram, GPU integrata con supporto Metal 3.

Configurata la piattaforma Ollama (Github, n.d.), e la Chatbox AI (n.d.) per l'interazione con gli LLMs, sono stati scaricati i modelli deepseek-v2, openai/hello, openchat, gemma, mistral, olmo2, deepseek-r1, llama3.1 (Figg. 5-12).

Figura 5. Modello Hello.

Figura 6. Modello Deepseek-V2.

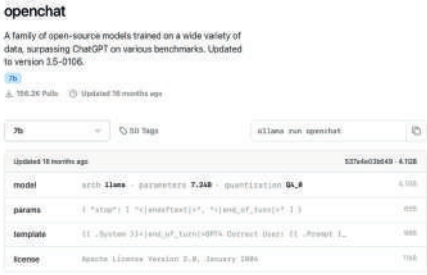


Figura 7. Modello Openchat.

Figura 8. Modello Mistral.

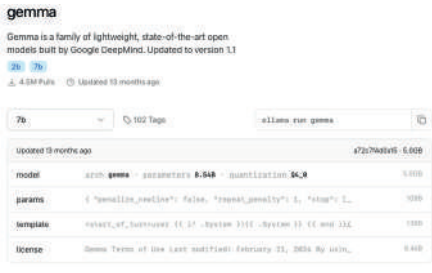


Figura 9. Modello Gemma.

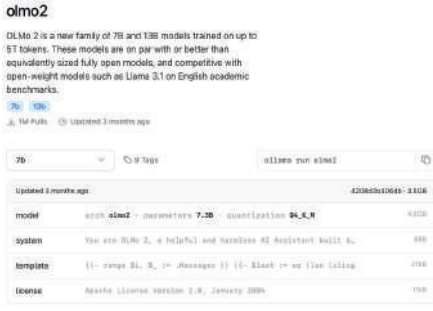


Figura 10. Modello Olmo2.

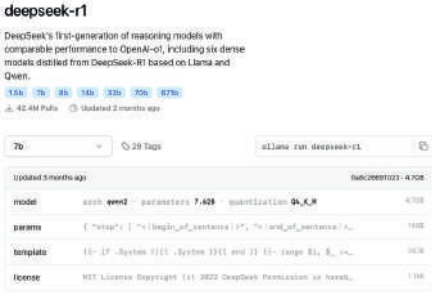


Figura 11. Modello Deepseek-R1.

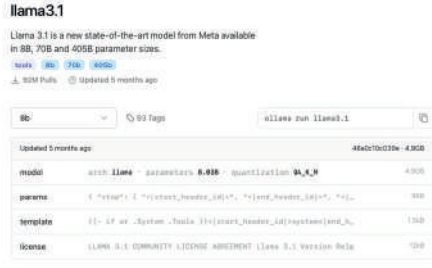
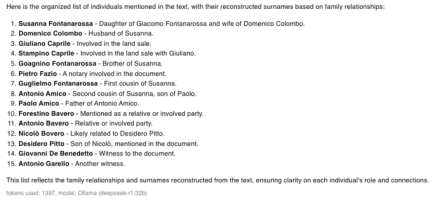


Figura 12. Modello Llama 3.1.

I vari modelli presentano specifiche caratteristiche, le quali sono determinanti nella scelta della versione da scaricare. I modelli più performanti, dal punto di vista dei risultati, sono quelli più grandi (671b, 405b, etc.). Ma più grande è il modello, maggiori sono le prestazioni hardware richieste. Ad esempio, un prompt di NER, nel caso di un modello di 236bv (deepseek-v2), espleta il processo in circa 40 minuti. I tempi aumentano quando ci spostiamo su modelli ancora più grandi. Invece, quelli più piccoli vengono gestiti senza difficoltà dall'hardware, ma i risultati che offrono sono meno dettagliati, e i processi di estrazione delle entità non portano i risultati sperati. Di seguito, nelle figure 13-18, si riportano gli *outcomes* di una NER eseguita su un documento trascritto (e tradotto) in inglese, pubblicato – e a cui si rimanda – in Aldo Agosto (1998, 92), in cui sono presenti queste entità nominate (persone): Susanna, Giacomo Fontanarossa, Domenico Colombo, Giuliano Caprile, Stampino Caprile, Goagnino Fontanarossa, Pietro Fazio, Antonio Amico, Paolo, Antonio Fontanarossa, Guglielmo Fontanarossa, Giovanni Fontanarossa, Bernardo Fontanarossa, Forestino Bavero, Antonio Bavero, Nicolò Bovero, Desidero Pitto, Nicolò, Lord Vicar, Lord Podestà of Genoa, Giovanni De Benedetto, Antonio, Antonio Garello, Gerolamo.

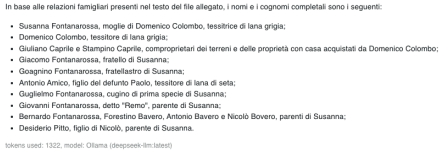

Here is the organized list of individuals mentioned in the text, with their reconstructed surnames based on family relationships:

1. **Susanna Fontanarossa** - Daughter of Giacomo Fontanarossa and wife of Domenico Colombo.
2. **Domenico Colombo** - Husband of Susanna.
3. **Giuliano Caprile** - Involved in the land sale.
4. **Stampino Caprile** - Involved in the land sale with Giuliano.
5. **Goagnino Fontanarossa** - Brother of Susanna.
6. **Pietro Fazio** - A notary involved in the document.
7. **Guglielmo Fontanarossa** - First cousin of Susanna.
8. **Antonio Amico** - Second cousin of Susanna, son of Paolo.
9. **Paolo Amico** - Father of Antonio Amico.
10. **Forestino Bavero** - Mentioned as a relative or involved party.
11. **Antonio Bavero** - Relative or involved party.
12. **Nicolò Bovero** - Likely related to Desidero Pitto.
13. **Desidero Pitto** - Son of Nicolò, mentioned in the document.
14. **Giovanni De Benedetto** - Witness to the document.
15. **Antonio Garello** - Another witness.

This list reflects the family relationships and surnames reconstructed from the text, ensuring clarity on each individual's role and connections.

tokens used: 1397, model: Ollama (deepseek-r1:32b)

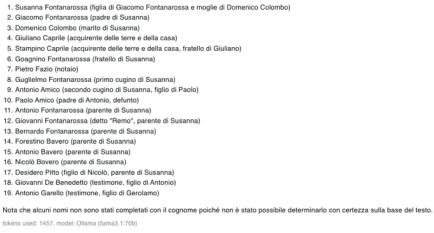
Figura 13. Outcome modello Deepseek-R1: 32b.


In base alle relazioni famigliari presenti nel testo del file allegato, i nomi e i cognomi completati sono i seguenti:

- Susanna Fontanarossa, moglie di Domenico Colombo, tessitrice di lana grigia;
- Domenico Colombo, tessitore di lana grigia;
- Giuliano Caprile e Stampino Caprile, comproprietari dei terreni e delle proprietà con casa acquistati da Domenico Colombo;
- Giacomo Fontanarossa, fratello di Susanna;
- Goagnino Fontanarossa, fratellastro di Susanna;
- Antonio Amico, figlio del defunto Paolo, tessitore di lana di seta;
- Guglielmo Fontanarossa, cugino di prima specie di Susanna;
- Giovanni Fontanarossa, detto "Remo", parente di Susanna;
- Bernardo Fontanarossa, Forestino Bavero, Antonio Bavero e Nicolò Bovero, parenti di Susanna;
- Desiderio Pitto, figlio di Nicolò, parente di Susanna.

tokens used: 1322, model: Ollama (deepseek-llm:latest)

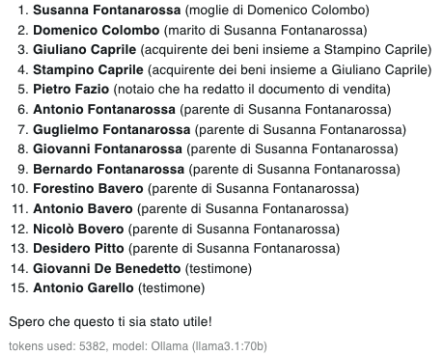
Figura 14. Outcome modello Deepseek-Llm: latest.


1. Susanna Fontanarossa (figlia di Giacomo Fontanarossa e moglie di Domenico Colombo)
2. Giacomo Fontanarossa (padre di Susanna)
3. Domenico Colombo (marito di Susanna)
4. Giuliano Caprile (acquirente delle terre e della casa)
5. Stampino Caprile (acquirente delle terre e della casa, fratello di Giuliano)
6. Goagnino Fontanarossa (fratello di Susanna)
7. Pietro Fazio (notaio)
8. Guglielmo Fontanarossa (primo cugino di Susanna)
9. Antonio Amico (secondo cugino di Susanna, figlio di Paolo)
10. Paolo Amico (padre di Antonio, defunto)
11. Antonio Fontanarossa (parente di Susanna)
12. Giovanni Fontanarossa (detto "Remo", parente di Susanna)
13. Bernardo Fontanarossa (parente di Susanna)
14. Forestino Bavero (parente di Susanna)
15. Antonio Bavero (parente di Susanna)
16. Nicolò Bovero (parente di Susanna)
17. Desidero Pitto (figlio di Nicolò, parente di Susanna)
18. Giovanni De Benedetto (testimone, figlio di Antonio)
19. Antonio Garello (testimone, figlio di Gerolamo)

Nota che alcuni nomi non sono stati completati con il cognome poiché non è stato possibile determinarlo con certezza sulla base del testo.

tokens used: 1457, model: Ollama (llama3.1:70b)

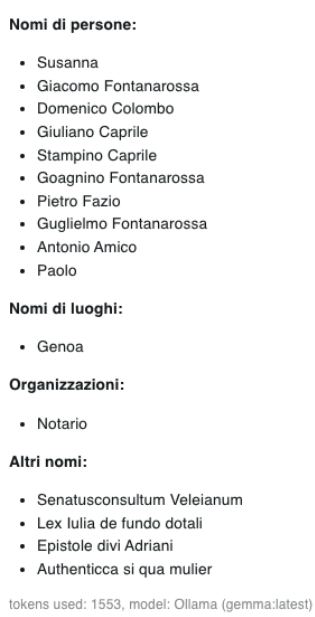
Figura 15. Outcome modello Llama 3.1: 70b.


1. **Susanna Fontanarossa** (moglie di Domenico Colombo)
2. **Domenico Colombo** (marito di Susanna Fontanarossa)
3. **Giuliano Caprile** (acquirente dei beni insieme a Stampino Caprile)
4. **Stampino Caprile** (acquirente dei beni insieme a Giuliano Caprile)
5. **Pietro Fazio** (notaio che ha redatto il documento di vendita)
6. **Antonio Fontanarossa** (parente di Susanna Fontanarossa)
7. **Guglielmo Fontanarossa** (parente di Susanna Fontanarossa)
8. **Giovanni Fontanarossa** (parente di Susanna Fontanarossa)
9. **Bernardo Fontanarossa** (parente di Susanna Fontanarossa)
10. **Forestino Bavero** (parente di Susanna Fontanarossa)
11. **Antonio Bavero** (parente di Susanna Fontanarossa)
12. **Nicolò Bovero** (parente di Susanna Fontanarossa)
13. **Desidero Pitto** (parente di Susanna Fontanarossa)
14. **Giovanni De Benedetto** (testimone)
15. **Antonio Garello** (testimone)

Spero che questo ti sia stato utile!

tokens used: 5382, model: Ollama (llama3.1:70b)

Figura 16. Outcome modello Llama 3.1: 70b.



Nomi di persone:

- Susanna
- Giacomo Fontanarossa
- Domenico Colombo
- Giuliano Caprile
- Stampino Caprile
- Goagnino Fontanarossa
- Pietro Fazio
- Guglielmo Fontanarossa
- Antonio Amico
- Paolo

Nomi di luoghi:

- Genoa

Organizzazioni:

- Notario

Altri nomi:

- Senatusconsultum Veleianum
- Lex Iulia de fundo dotali
- Epistole divi Adriani
- Authenticca si qua mulier

tokens used: 1553, model: Ollama (gemma:latest)

Figura 17. Outcome modello Gemma: latest.

- Susanna Fontanarossa
- Domenico Colombo
- Giuliano Caprile
- Stampino Caprile
- Goagnino Fontanarossa
- Pietro Fazio
- Guglielmo Fontanarossa
- Antonio Amico
- Paolo
- Antonio Fontanarossa
- Guglielmo Fontanarossa
- Giovanni Fontanarossa ("Remo")
- Bernardo Fontanarossa
- Forestino Bavero
- Antonio Bavero
- Nicolò Bovero
- Desidero Pitto

Tabella dei Nominativi:

Nominativo	Descrizione in inglese
Susanna Fontanarossa	Daughter of the deceased Giacomo Fontanarossa
Domenico Colombo	Weaver of woolen cloth
Giuliano Caprile	Purchaser of certain lands and possessions
Stampino Caprile	Purchaser of certain lands and possessions
Goagnino Fontanarossa	Brother of Susanna Fontanarossa
Pietro Fazio	Notary
Guglielmo Fontanarossa	First cousin of Susanna Fontanarossa
Antonio Amico	Second cousin of Susanna Fontanarossa
Giovanni Fontanarossa ("Remo")	Son of the deceased Paolo
Bernardo Fontanarossa	Son of Giacomo Fontanarossa
Forestino Bavero	Son of Giacomo Fontanarossa
Antonio Bavero	Son of Giacomo Fontanarossa
Nicolò Bovero	Son of Giacomo Fontanarossa
Desidero Pitto	Son of Nicolò

tokens used: 1653, model: Ollama (gemma:latest)

Figura 18. Outcome tabella con modello Gemma: latest.

I modelli testati, al di là delle diverse prestazioni in ordine al tempo impiegato per l'elaborazione del prompt, compongono liste di nomi non complete, e la maggior parte di loro non riesce a ricostruire correttamente il quadro delle relazioni famigliari, partendo dalle descrizioni dei legami presenti nel testo del documento.

L'unico modello in grado di soddisfare il prompt è stato il "deepseek-r1:671b", il quale, però, ha impiegato 57 minuti per comporre la lista e le relazioni famigliari.

Tutti i modelli, inoltre, in remoto, non sono in grado di organizzare i dati in un file CSV e consentire il download di quest'ultimo, per la costruzione dei records del database.

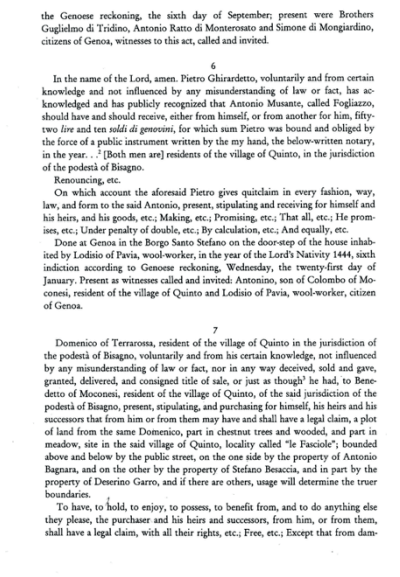
Questo quadro, così come descritto, si traduce in limiti reali nel lavoro dello studioso. Tempi lunghi di elaborazione e computazione non corretta creano un divario tra il mondo della Storia – e dell'Archivistica – e quello delle tecnologie di IA, le quali, seppur promettenti nelle loro prospettive di impiego, non presentano ancora un addestramento adeguato nei due settori di ricerca.

Un ulteriore test è stato effettuato mediante costruzione di un codice Python, con piattaforma Jupyter Notebook (n.d.), attraverso la quale è stato elaborato un codice per l'estrazione delle entità nominate, che ha restituito dettagliatamente le varie entità, ma mostrando un "bug" che necessitava un ul-

teriore lavoro (in termini di tempi di programmazione) di modifica del codice. Nel PDF, ogni documento è diviso da quello precedente, attraverso un sistema tradizionale di numerazione araba (Fig. 19). Tale separazione non è "individuata" dal codice, conseguentemente, l'elenco delle entità estratte comprende tutto quello che viene nominato nella pagina del PDF, creando una lista che, seppur corretta, non consente di estrapolare i dati per singolo documento archivistico, costruire il singolo regesto e compilare il singolo file CSV.

A questo si aggiunga la necessità di una configurazione costante del codice, relativamente alle API Key dei singoli modelli, per cui, pur valutando positivamente il risultato, non è stato implementato alcun codice Python per continuare il lavoro in questa direzione.

Di seguito, l'esempio del layout del PDF, il codice costruito e i risultati del test (Figg. 20-22).



the Genoese reckoning, the sixth day of September; present were Brothers Guglielmo di Tridino, Antonio Ratto di Monterosato and Simone di Mongiardino, citizens of Genoa, witnesses to this act, called and invited.

6

In the name of the Lord, amen. Pietro Ghirardetto, voluntarily and from certain knowledge and not influenced by any misunderstanding of law or fact, has acknowledged and has publicly recognized that Antonio Musante, called Fogliazzo, should have and should receive, either from himself, or from another for him, fifty-two *lire* and ten *soldi di genovini*, for which sum Pietro was bound and obliged by the force of a public instrument written by the my hand, the below-written notary, in the year. . .[2] [Both men are] residents of the village of Quinto, in the jurisdiction of the podestà of Bisagno.

Renouncing, etc.

On which account the aforesaid Pietro gives quitclaim in every fashion, way, law, and form to the said Antonio, present, stipulating and receiving for himself and his heirs, and his goods, etc.; Making, etc.; Promising, etc.; That all, etc.; He promises, etc.; Under penalty of double, etc.; By calculation, etc.; And equally, etc.

Done at Genoa in the Borgo Santo Stefano on the door-step of the house inhabited by Lodisio of Pavia, wool-worker, in the year of the Lord's Nativity 1444, sixth indiction according to Genoese reckoning, Wednesday, the twenty-first day of January. Present as witnesses called and invited: Antonino, son of Colombo of Moconesi, resident of the village of Quinto and Lodisio of Pavia, wool-worker, citizen of Genoa.

7

Domenico of Terrarossa, resident of the village of Quinto in the jurisdiction of the podestà of Bisagno, voluntarily and from his certain knowledge, not influenced by any misunderstanding of law or fact, nor in any way deceived, sold and gave, granted, delivered, and consigned title of sale, or just as though[3] he had, to Benedetto of Moconesi, resident of the village of Quinto, of the said jurisdiction of the podestà of Bisagno, present, stipulating, and purchasing for himself, his heirs and his successors that from him or from them may have and shall have a legal claim, a plot of land from the same Domenico, part in chestnut trees and wooded, and part in meadow, site in the said village of Quinto, locality called "le Fasciole"; bounded above and below by the public street, on the one side by the property of Antonio Bagnara, and on the other by the property of Stefano Besaccia, and in part by the property of Deserino Garro, and if there are others, usage will determine the truer boundaries.

To have, to hold, to enjoy, to possess, to benefit from, and to do anything else they please, the purchaser and his heirs and successors, from him, or from them, shall have a legal claim, with all their rights, etc.; Free, etc.; Except that from dam-

Figura 19. Pagina PDF dei vari documenti, senza separazione tra le parti.

6

In the name of the Lord, amen. Pietro Ghirardetto, voluntarily and from certain knowledge and not influenced by any misunderstanding of law or fact, has acknowledged and has publicly recognized that Antonio Musante, called Fogliazzo, should have and should receive, either from himself, or from another for him, fifty-two *lire* and ten *soldi di genovini*, for which sum Pietro was bound and obliged by the force of a public instrument written by the my hand, the below-written notary, in the year. . .[2] [Both men are] residents of the village of Quinto, in the jurisdiction of the podestà of Bisagno.

Renouncing, etc.

On which account the aforesaid Pietro gives quitclaim in every fashion, way, law, and form to the said Antonio, present, stipulating and receiving for himself and his heirs, and his goods, etc.; Making, etc.; Promising, etc.; That all, etc.; He promises, etc.; Under penalty of double, etc.; By calculation, etc.; And equally, etc.

Done at Genoa in the Borgo Santo Stefano on the door-step of the house inhabited by Lodisio of Pavia, wool-worker, in the year of the Lord's Nativity 1444, sixth indiction according to Genoese reckoning, Wednesday, the twenty-first day of January. Present as witnesses called and invited: Antonino, son of Colombo of Moconesi, resident of the village of Quinto and Lodisio of Pavia, wool-worker, citizen of Genoa.

Figura 20. Testo tratto da Aldo Agosto (1998, 32).

Figura21. Prompt training NER file PDF.

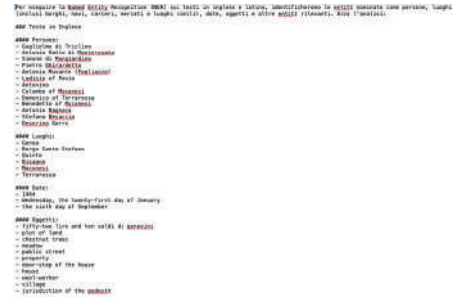
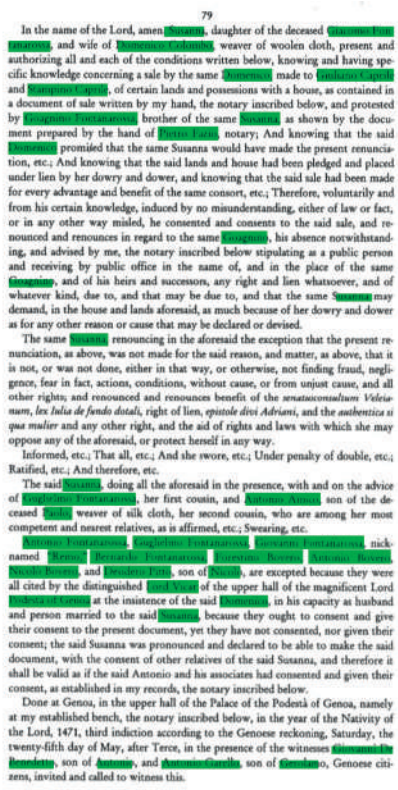


Figura 22. Outcome analisi riconoscimento entità nominate.

Il processo, invece, è totalmente diverso, e realmente promettente, quando si passa alle versioni online delle IA, soprattutto nei modelli a pagamento (in abbonamento). In questo caso, infatti, non si riscontrano errori negli *outcomes* (Figg. 23-24).



79

In the name of the Lord, amen Susanna, daughter of the deceased Giacomo Fontanarossa, and wife of Domenico Colombo, weaver of woolen cloth, present and authorizing all and each of the conditions written below, knowing and having specific knowledge concerning a sale by the same Domenico made to Giuliano Caprile and Stampino Caprile, of certain lands and possessions with a house, as contained in a document of sale written by my hand, the notary inscribed below, and protested by Goagnino Fontanarossa, brother of the same Susanna, as shown by the document prepared by the hand of Pietro Fazio, notary; And knowing that the said Domenico promised that the same Susanna would have made the present renunciation, etc.; And knowing that the said lands and house had been pledged and placed under lien by her dowry and dower, and knowing that the said sale had been made for every advantage and benefit of the same consort, etc.; Therefore, voluntarily and from his certain knowledge, induced by no misunderstanding, either of law or fact, or in any other way misled, he consented and consents to the said sale, and renounced and renounces in regard to the same Guiglielmo, his absence notwithstanding, and advised by me, the notary inscribed below stipulating as a public person and receiving by public office in the name of, and in the place of the same Goagnino, and of his heirs and successors, any right and lien whatsoever, and of whatever kind, due to, and that may be due to, and that the same Susanna may demand, in the house and lands aforesaid, as much because of her dowry and dower as for any other reason or cause that may be declared or devised.

The same Susanna, renouncing in the aforesaid the exception that the present renunciation, as above, was not made for the said reason, and matter, as above, that it is not, or was not done, either in that way, or otherwise, not finding fraud, negligence, fear in fact, actions, conditions, without cause, or from unjust cause, and all other rights; and renounced and renounces benefit of the *senatusconsultum Velleianum*, *lex Iulia de fundo dotali*, right of lien, *epistole divi Adriani*, and the *authentica si qua mulier* and any other right, and the aid of rights and laws with which she may oppose any of the aforesaid, or protect herself in any way.

Informed, etc.; That all, etc.; And she swore, etc.; Under penalty of double, etc.; Ratified, etc.; And therefore, etc.

The said Susanna, doing all the aforesaid in the presence, with and on the advice of Guglielmo Fontanarossa, her first cousin, and Antonio Amico, son of the deceased Paolo, weaver of silk cloth, her second cousin, who are among her most competent and nearest relatives, as is affirmed, etc.; Swearing, etc.

Antonio Fontanarossa, Guglielmo Fontanarossa, Giovanni Fontanarossa, nicknamed "Remo", Bernardo Fontanarossa, Forestino Bavero, Antonio Bavero, Nicolò Bovero, and Desidero Pitto, son of Nicolò, are excepted because they were all cited by the distinguished Lord Vicar of the upper hall of the magnificent Lord Podesta of Genoa at the insistence of the said Domenico, in his capacity as husband and person married to the said Susanna, because they ought to consent and give their consent to the present document, yet they have not consented, nor given their consent; the said Susanna was pronounced and declared to be able to make the said document, with the consent of other relatives of the said Susanna, and therefore it shall be valid as if the said Antonio and his associates had consented and given their consent, as established in my records, the notary inscribed below.

Done at Genoa, in the upper hall of the Palace of the Podestà of Genoa, namely at my established bench, the notary inscribed below, in the year of the Nativity of the Lord, 1471, third indiction according to the Genoese reckoning, Saturday, the twenty-fifth day of May, after Terce, in the presence of the witnesses Giovanni De Benedetto, son of Antonio, and Antonio Garello, son of Gerolamo, Genoese citizens, invited and called to witness this.

Figura 23. Testo tratto da Aldo Agosto (1998, 92).

Ecco la lista riorganizzata con i cognomi attribuiti sulla base delle relazioni familiari:

1. **Susanna Fontanarossa** – figlia del defunto Giacomo Fontanarossa e moglie di Domenico Colombo.
2. **Giacomo Fontanarossa** (defunto) – padre di Susanna.
3. **Domenico Colombo** – marito di Susanna, tessitore di panni di lana.
4. **Giuliano Caprile** – acquirente delle terre e possedimenti.
5. **Stampino Caprile** – acquirente delle terre e possedimenti.
6. **Goagnino Fontanarossa** – fratello di Susanna, protestante della vendita.
7. **Pietro Fazio** – notaio che redasse un documento riguardante la protesta di Goagnino.
8. **Guglielmo Fontanarossa** – primo cugino di Susanna, presente per consigliarla.
9. **Antonio Amico** – figlio del defunto Paolo Amico, secondo cugino di Susanna, presente per consigliarla.
10. **Paolo Amico** (defunto) – padre di Antonio Amico.
11. **Antonio Fontanarossa** – parente di Susanna, citato ma non consenziente.
12. **Giovanni Fontanarossa**, soprannominato "Remo" – parente di Susanna, citato ma non consenziente.
13. **Bernardo Fontanarossa** – parente di Susanna, citato ma non consenziente.
14. **Forestino Bavero** – parente di Susanna, citato ma non consenziente.
15. **Antonio Bavero** – parente di Susanna, citato ma non consenziente.
16. **Nicolò Bovero** – parente di Susanna, citato ma non consenziente.
17. **Desidero Pitto** – figlio di Nicolò Pitto, parente di Susanna, citato ma non consenziente.
18. **Nicolò Pitto** – padre di Desidero Pitto, parente di Susanna.
19. **Signore Vicario della sala superiore del Podestà di Genova** – funzionario che gestì la questione.
20. **Magnifico Signore Podestà di Genova** – autorità sotto cui si svolse l'atto.
21. **Giovanni De Benedetto** – figlio di Antonio De Benedetto, testimone.
22. **Antonio De Benedetto** – padre di Giovanni De Benedetto.
23. **Antonio Garello** – figlio di Gerolamo Garello, testimone.
24. **Gerolamo Garello** – padre di Antonio Garello.
25. **Notaio** (autore del documento) – non nominato esplicitamente, ma identificato come colui che ha redatto l'atto.

Figura 24. Risultato della NER con ChatGPT-4.5.

Testate le varie piattaforme, dunque, il lavoro di estrazione dati e compilazione del CSV viene affidato completamente a ChatGPT-4 (successivamente 4.5).

Entrambi i modelli soddisfano le aspettative, ed i risultati sono eccellenti in tutte le operazioni richieste, consentendo di procedere all'analisi dei documenti utili alla creazione del database.

I documenti sono trascrizioni digitali tratti dalla relazione storica di Emilio Marengo, *Genova e Tunisi 1388-1515*, scritta nel 1901 per gli *Atti della Società Ligure di Storia Patria, Volume XXXII*, in cui sono confluiti 64 documenti, conservati presso l'Archivio di Stato di Genova, relativi all'oggetto di studio del progetto PNRR.

Il PDF del testo è stato trattato con software OCR, Adobe PDF Pro, per ottenere la trascrizione dei documenti, in formato DOC. Ottenute le trascrizioni, i testi sono stati processati, come già detto, con i modelli ChatGPT-4 e 4.5.

La scelta, oltre che motivata dagli errori negli *outcomes* dei modelli in remoto, è stata puntualizzata anche dal fatto che il modello di OpenAI riesce a

produrre file CSV correttamente compilati e scaricabili direttamente dall'interfaccia di dialogo. Inoltre, i due modelli riescono, con specifici comandi, a realizzare processi di Network Analysis che possono essere tradotti in chiave "visual" (Fig. 25) (Bleichmar e Schwartz 2019; Bradley et al. 2018; Ebbrecht-Hartmann et al. 2023; Staley 2015; Theibault 2016).

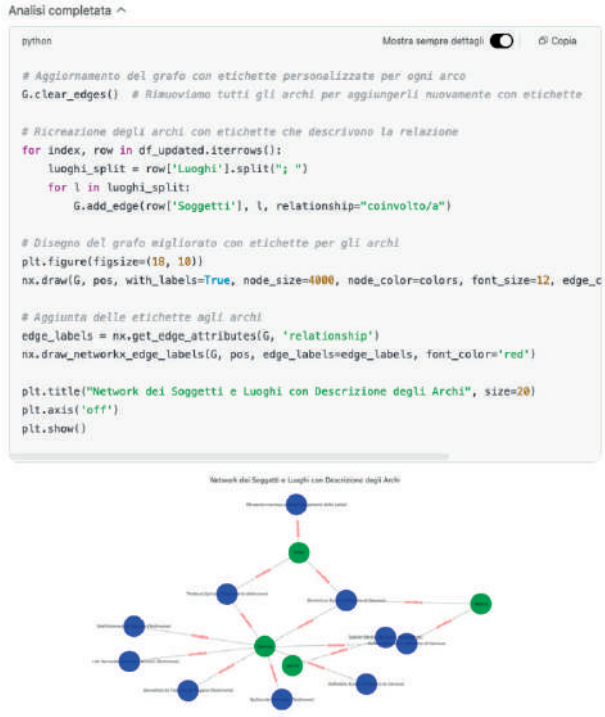


Figura 25. Esempio di network analysis.

Si aggiunga l'enorme capacità, di tale IA, di sviluppare codici in diversi linguaggi (es. Python), che possono essere utilizzati in processi di training – ad esempio, con TensorFlow (TF) –, al fine di potenziare il modello con i dati che esso stesso ha ricavato, in un processo che, se da un lato estrae dati, da un altro lato, li utilizza in una successiva fase di training, che riesce a tradurre quei dati in informazioni storiche.

Nel nostro caso, il codice TF compilato ha consentito di "allenare" il modello sui significati delle espressioni toponomastiche relative ai termini *Terra nigrorum* e su altre che sono state individuate in ulteriori documenti (Fig. 26).

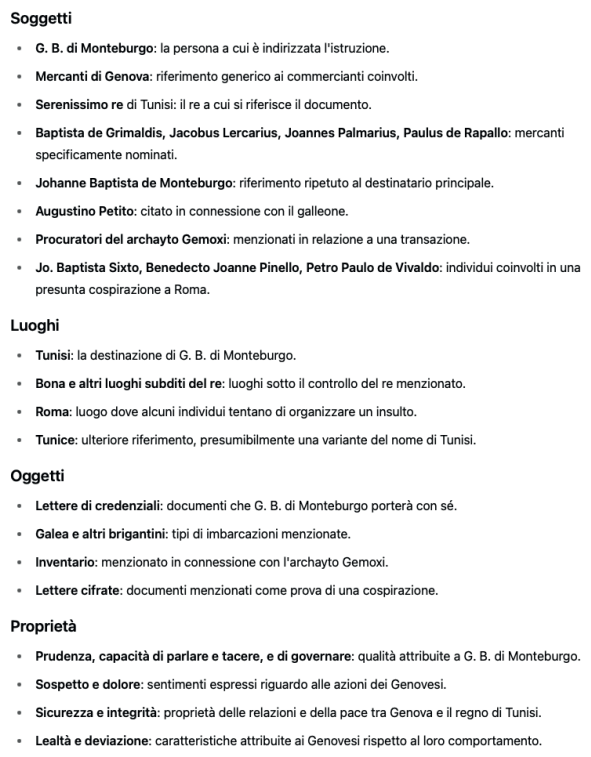
Figura 26. Training con TensorFlow per entità nominate (toponomi non specifici).

Tutto questo, però, richiede un controllo specifico da parte dell'utente e la compilazione di specifici prompt, che devono istruire la macchina a non creare *outcomes* "allucinati".

Il lavoro di configurazione dei prompt, infatti, è lungo e capzioso, in quanto l'utente deve predisporre le istruzioni in una vera e propria sequenza, in cui l'IA non deve essere lasciata libera. Ogni singolo step computazionale deve essere correttamente eseguito dal modello, e, per questo, il prompt deve essere esplicito, logico e discreto.

Dopo varie fasi e diversi test sulla struttura del prompt, che andava dal semplice «Estrai entità nominate» – che lasciava all'IA la libertà di individuare *entities* non esatte, o di crearne altre –, si addiviene ad un comando esplicito: «Estrai entità nominate, nello specifico: persone, luoghi, oggetti, funzioni, tralasciando le date».

In questo caso, ChatGPT è riuscita ad estrarre correttamente tutte le entità e a dividerle per categoria (Fig. 27).



Soggetti

- **G. B. di Monteburgo**: la persona a cui è indirizzata l'istruzione.
- **Mercanti di Genova**: riferimento generico ai commercianti coinvolti.
- **Serenissimo re** di Tunisi: il re a cui si riferisce il documento.
- **Baptista de Grimaldis, Jacobus Lercarius, Joannes Palmarius, Paulus de Rapallo**: mercanti specificamente nominati.
- **Johanne Baptista de Monteburgo**: riferimento ripetuto al destinatario principale.
- **Augustino Petito**: citato in connessione con il galleone.
- **Procuratori del archayto Gemoxi**: menzionati in relazione a una transazione.
- **Jo. Baptista Sixto, Benedecto Joanne Pinello, Petro Paulo de Vivaldo**: individui coinvolti in una presunta cospirazione a Roma.

Luoghi

- **Tunisi**: la destinazione di G. B. di Monteburgo.
- **Bona e altri luoghi subditi del re**: luoghi sotto il controllo del re menzionato.
- **Roma**: luogo dove alcuni individui tentano di organizzare un insulto.
- **Tunice**: ulteriore riferimento, presumibilmente una variante del nome di Tunisi.

Oggetti

- **Lettere di credenziali**: documenti che G. B. di Monteburgo porterà con sé.
- **Galea e altri brigantini**: tipi di imbarcazioni menzionate.
- **Inventario**: menzionato in connessione con l'archayto Gemoxi.
- **Lettere cifrate**: documenti menzionati come prova di una cospirazione.

Proprietà

- **Prudenza, capacità di parlare e tacere, e di governare**: qualità attribuite a G. B. di Monteburgo.
- **Sospetto e dolore**: sentimenti espressi riguardo alle azioni dei Genovesi.
- **Sicurezza e integrità**: proprietà delle relazioni e della pace tra Genova e il regno di Tunisi.
- **Lealtà e deviazione**: caratteristiche attribuite ai Genovesi rispetto al loro comportamento.

Figura 27. Divisione delle entità nominate per classi (Soggetti, Luoghi, Oggetti, Proprietà).

Successivamente, è stato costruito il prompt relativo alla creazione del file CSV. Anche in questo caso, dal semplice comando «Crea file CSV con i dati estratti dal testo analizzato», è stato necessario costruire un prompt specifico, in quanto l'IA organizzava il foglio in maniera non funzionale alla compilazione del record per il database finale. La maggior parte degli *outcomes*, infatti, si limitava alla semplice lista dei nomi e la relativa categoria, oppure – a causa del fenomeno di "allucinazione" – alla creazione di una vera e propria immagine, che rappresentava un foglio CSV (Fig. 28).

Figura 28. Fenomeno "allucinazione" di ChatGPT nella fase di creazione del file EXCEL.

Per questi motivi, il prompt è stato costruito tenendo conto di alcune specifiche informazioni, che dovevano essere necessariamente evidenziate nelle colonne del foglio di calcolo (Fig. 29):

> crea un file CSV dove metti nella prima colonna il nome dell'entity, nella seconda colonna il nome esatto ricavato dal testo latino, nella terza colonna la categoria di entity (person, place, physical object), nella quarta colonna il ruolo dell'entity all'interno del documento, rispettando queste istruzioni: se l'entity è una persona, devi inserire il lavoro e i legami famigliari con altri soggetti del documento, se menzionati. Se invece è un luogo, inserisci che funzione ha quel luogo nel contesto del documento. Se invece è un oggetto, inserisci che funzione ha quell'oggetto nel contesto del documento.

Figura 29. Prompt dettagliato per la costruzione e compilazione del file CSV.

Questo ultimo prompt ha consentito alla macchina di configurare le colonne del CSV e, successivamente, compilare il file pronto per il download (Fig. 30).

Figura 30. Esecuzione in codice del prompt per la costruzione e compilazione del file CSV.

Il lavoro si completa con il comando in cui si istruisce l'IA alla compilazione di un "regesto", partendo da un breve training che obbligava il modello a tenere fermi dei limiti di carattere/parole, e normalizzare il testo, al fine di costruire un regesto in linguaggio moderno (Fig. 31).

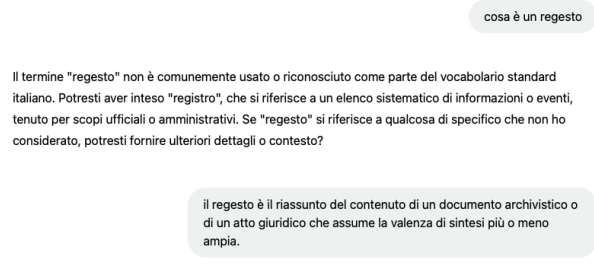


Figura 31. Training per la compilazione del regesto.

Costruito il prompt, quest'ultimo è stato testato sul testo che viene di seguito riportato – in formato immagine, per garantire la riproducibilità del test –, in cui sono presenti anche errori di trascrizione, per verificare, contestualmente, la capacità del modello ChatGPT-4 di normalizzare il testo e correggere i vari errori (Figg. 32, 33):

«In eterni Dei nomine amen. Magnificus dominus Anthoniotus Adurnus pro serenissimo domino rege Francorum gubernator Januensium et communis et populi defensor. Et suum venerabile consilium dominorum decem octo sapientium antianorum. In pleno, integro et totali numero congregatorum, quorum nomina sunt hec: Dom inus Antonius Justinianus miles, prior; Gentilis de Grimaldis ; Raffus Lecavellum ; Antonius de Lacastanea; Leonel de M ari; Janinus de Serra de Pulcifera; Raffael de Facio; Abraymus Pilavicinus; Johannes Ususm aris Petri; Lucianus Spinula Cepriani ; Andreas M am iffus; Francus de Francis; Jacobus de Auria; Dom inicus Bosonus de Struppa; Cataneus Cigala; Jacobus de S alv o ; Jnoflinos de Solario de Cogoleto et Petrus de Vivaldis. Agentes nomine et vice communis fanue et pro ipso communi. Confisi de discrecione et probitate nobilis viri Caroli G rilli civis Januensis. Omni via, iure, modo et forma, quibus m elius potuerunt ct possunt, fecerunt, constituerunt, creaverunt et ordinaverunt eorum dicto nomine et dicti communis Janue certum verum legiptimum et indubitatum actorem, sindicum, procuratorem et ambassatorem ac nuncium specialem, et quidquid et prout de jure melius fieri et esse potesi loco ipsorum dicto nom ine et dicti communis posuerunt et ponunt dictum Carolum G rillum , licet absentem, duraturum usque ad annum unum proxime venturum . A d eundum et se personaliter conferendum ad sublimem conspectum serenissimi et illustrissimi principis et domini domini Mulev Buffers (i) , Dei gratia regis Tunexis et Buzee etc., ct coram quibuscumque consiliariis, auditoribus vel officialibus ipsius constitutis vel constituendis. Et ad confirmandum, ratificandum et approbandum ac de novo, si expediet, faciendum pacem, composicionem seu conventioncm et pacta, vigentem et rigentia inter dictum serenissimum regem si\e digne memorie quondam dominum....(2) olim regem patrem suum, ex una parte, et dictos constituentes dicto nomine sive dictum commune Janue, ex altera, sub illis pactis, promissionibus, obligationibus et cautellis, sub quibus fuit ultimo dicta pax et conventio com posita et firmata, existentibus ambaxatoribus pro dicto communi Janue ad predictam nobilibus et discretis viris Gentile de G rim aldis ei Luchino de Bonavey. E t ad petendum, requirendum et postulandum restitutionem seu satisfactionem et emendam omnium et singulorum dampnorum. robariarum et depredacionum Januensibus seu subditis communis Janue illatorum seu commissarum contra aliquos Januenses vel eorum bona per quosvis subditos prefati serenissimi domini regis Tunexis a tempore dicte alias firmate pacis per supradictos Gentilem et Luchinum ambaxatores dicti communis Janue citra. Nec non ad requirendum, instandum et prosequendum liberacionem et relaxacionem quorumcumque in regno ipsius captivorum Januer.sium vel subditorum communis Janue, tam qui restarent dicto tempore firmate dicte pacis captivi, quam qui postea captivati fuissent. Item ad quitandum, liberandum et absolvendum, nomine dictorum constituentium et dicti communis, prefatum serenissimum dominum regem, heredes et successores eiusdem, de omni eo et toto quod pro emendea, restitucione vel alia satisfacione dictorum dampnorum Januensibus illatorum receperit vel habuerit, seque de eo bene quietum et solutum vocandum nomine sepedicto. Iit ad unum ct plura instrumentum et instrumenta de et super predictis omnibus et singulis. Cum illis omnibus confessionibus, renuntiationibus, promissionibus, obligationibus, clausulis et cautellis, de quibus dicto sindico ambaxatori et procuratori dicto nomine videbitur et placuerit faciendum et fieri seu confici faciendum. Et demum generaliter ad omnia et singula gerendum, faciendum, procurandum et administrandum in predictis omnibus et singulis, et in dependentibus, accessoriis, annexis et connexis predictis et a predictis, ct cuilibet et a quolibet predictorum, et circa predicta que opportuna necessaria vel utilia videantur, queque per quemcumque verum, legitimum et indubitatum procuratorem et sindicum plena et omnimoda potestate sussultum fieri possent, et que ipsimet constituentes possent facere si personaliter interessent, etiam si talia forent que mandatum exigerent speciale. Dantes et concedentes dicto nomine eidem Carolo, sindico et procuratori predicto, in predictis omnibus et singulis et circa predicta, ac in dependentibus, accessoriis, annexis et connexis predictis ct a predictis, et cuilibet et a quolibet predictorum, plenum largum libcium ct generale mandatum cum plena larga libera ct generali administratiooe. Promittentes michi Antonio de Credentia notario et dicti com munis Janue cancellario infrascripco, tamquam publice persone officio publico stipulanti et recipienti nomine et vice prefati sci diissim i domini regis Tunexis et heredum et successorum suorum et cuius et quorum interest, intererit vel interesse poterit, sese dicto nomine habituros et quod dictum commune habebit et tenebit perpetuo ratum gratum et firmum quidquid per dictum eorum sindicum, ambaxatorem et procuratorem, in predictis om nibus et singulis actum, gestum, vel factum fuerit, seu etiam procuratum . Sub ypotheca et obligatione bonorun ipsorum constituentium dicto nomine seu dicti communis Janue habitorum et habendorum. M andantes etiam de predictis omnibus confici debere publicum instrumentum per me supradictum Antonium de Credentia not-irium et dictorum constituentium ac dicti communis Janui cancellarium infrascriptum . Actum Janue in palacio communis, videlicet in camera cubiculari dicti magnifici domini gubernatoris iuxta turrim, anno dominice nativitatis millesimo trecentesimo nonagesimo septimo, indicione quarta secundum cursum Janue, die mercurii X III mensis Marcii, in sero, presentibus Conrado Mazurro notario et communis Janue cancellario, Antonio de Montaldo d^ Gavio quondam Priani, et A ntonio Pellerano de Camulio dicto longo, macerio dicti magnifici dom ini gubernatoris, testibus ad predicta vocatis specialiter et rogatis» .

Figura 32. Documento usato per testare il prompt per la compilazione dei regesti, tratto da Marengo (1901, 139).

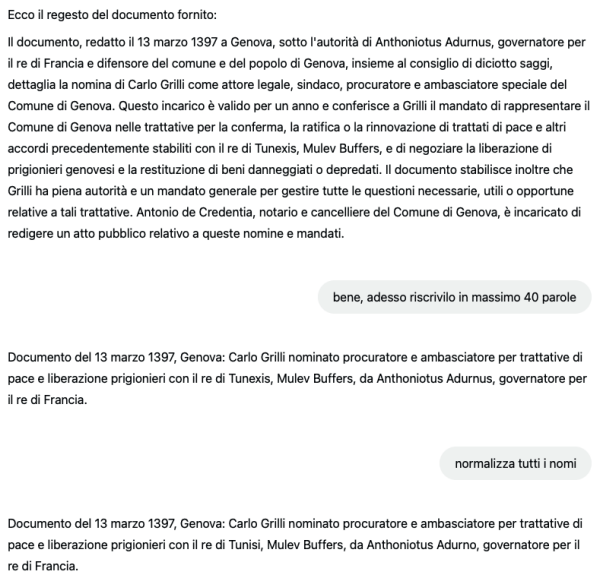


Figura 33. *Outcome* della compilazione del regesto.

Ultimati i vari test, il prompt finale, utilizzato nelle varie fasi di estrazione dei dati e compilazione dei records è il seguente (Fig. 34):

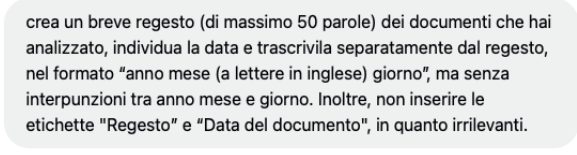


Figura 34. Prompt dettagliato per la compilazione dei regesti.

Il complessivo lavoro di estrazione e compilazione ha consentito di individuare – in un workflow di 450 ore complessive –, su 64 documenti d'archivio trascritti in digitale, 754 *entities*:

- 406 persone
- 8 gruppi
- 3 istituzioni
- 2 organizzazioni
- 144 oggetti
- 198 luoghi

- 20 proprietà
- 2 n.c.

Tali entità, nel corpo del file CSV finale (Figg. 35, 36), vengono organizzate sulla base di un ID univoco, che consente, da un lato, di individuare correttamente l'info richiesta, e, da un altro lato, permettere al sistema "Arches" di disporre la visualizzazione delle informazioni, per i web users, in maniera univocamente correlata.

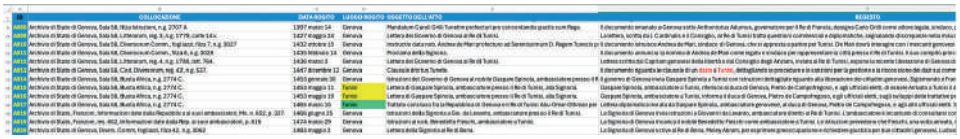
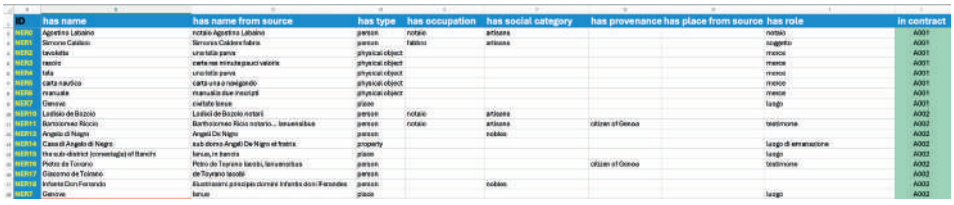


Figura 35. Struttura finale del file CSV del dataset (foglio relativo alla "Collocazione").



ID	has name	has name from source	has type	has occupation	has social category	has provenance	has place from source	has role	in contract
N100	Agostino Lalsalvo	notaio Agostino Lalsalvo	person	notaio	artisans			notaio	A001
N101	Simone Calderi	Simone Calderi fabro	person	fabbro	artisans			maestro	A001
N102	tavoletta	una tabla parva	physical object					merce	A001
N103	rastro	carta nas minutegauci valoris	physical object					merce	A001
N104	tela	una tabla parva	physical object					merce	A001
N105	carta nautica	carta una o navegando	physical object					merce	A001
N106	manuale	manuale due inscripti	physical object					merce	A001
N107	Genova	civitate Ianue	place					luogo	A001
N1010	Ludisio de Bozolo	Ludisio de Bozolo notarii	person	notaio	artisans				A002
N1011	Bartolomeo Riccio	Bartholomeo Ricio notario... Ianuensibus	person	notaio	artisans	citizen of Genoa		testimone	A002
N1012	Angelo di Nagno	Angeli De Nigro	person		nobles				A002
N1014	Casa di Angelo di Nagno	sub dicmo Angeli De Nigro et fratris	property					luogo di ammarazione	A002
N1015	the sub-district [correntaglia] of Bianchi	Ianua, in banchi	place					luogo	A002
N1016	Pietro de Torsino	Petro de Taynino Iacobi, Ianuensibus	person			citizen of Genoa		testimone	A002
N1017	Giacomo de Tolomo	de Tayano Iacobi	person						A002
N1018	Infanta Dom Fernando	illustrissimo principale domini Infantis dom Fernandes	person		nobles				A002
N107	Genova	Ianua	place					luogo	A003

Figura 36. Struttura finale del file CSV del dataset (foglio relativo alle entità nominate).

4. Conclusioni

La costruzione del "Database on the Slave Trade between the Mediterranean and the Atlantic (15th-16th centuries)" si colloca al crocevia tra innovazione tecnologica e riflessione epistemologica, restituendo centralità all'*histoire sérielle* come metodo e rilanciando la funzione dell'archivio come dispositivo cognitivo, interpretativo e politico. L'introduzione di strumenti di Intelligenza Artificiale generativa nei processi di estrazione, codifica e strutturazione dei dati documentari non rappresenta semplicemente un salto tecnologico, ma impone una ridefinizione profonda della semantica archivistica, delle pratiche storiografiche e delle modalità di costruzione della memoria storica. In questa prospettiva, gli *invented archives* non sono solo ambienti digitali, ma spazi di negoziazione epistemica, in cui i dati sono selezionati, modellati e interrogati in relazione a specifici bisogni interpretativi, teorici e politici.

L'esperienza condotta ha mostrato come i modelli di IA, se opportunamente addestrati e guidati da prompt costruiti con rigore metodologico, siano oggi in grado di affrontare le complessità linguistiche, strutturali e semantiche della

documentazione storica. Tuttavia, le criticità emerse – dalla possibilità di generare output allucinati, all'opacità della logica inferenziale dei modelli – rendono evidente la necessità di mantenere un controllo umano costante, capace di orientare l'interazione uomo-macchina entro un quadro di responsabilità epistemica. L'intervento dello storico, in tal senso, non può essere surrogato dall'automazione: è nella progettazione ontologica, nella calibrazione dei criteri di estrazione, nella verifica contestuale e nell'analisi qualitativa che si gioca la scientificità del risultato.

Al tempo stesso, il progetto ha rivelato le ambivalenze di un'archivistica digitale ancora parzialmente ancorata a logiche indicizzative tradizionali, incapaci di attivare pienamente le potenzialità computazionali offerte dalla Digitalità. La trasformazione del documento in *cybertesto* (Aarseth 1997), la costruzione di ambienti interrogabili e relazionali, l'integrazione tra approccio seriale e Machine Learning, sono obiettivi ancora da consolidare, soprattutto in una prospettiva disciplinare in cui Storia e Archivistica, spesso, procedono su binari paralleli. In questo contesto, l'Intelligenza Artificiale si configura come agente tecnico, ma anche come oggetto critico, capace di rimettere al centro il dibattito su autorità archivistica, gerarchie semantiche, invisibilità strutturale e *agency* dei soggetti rappresentati.

Il database, pensato come piattaforma aperta, scalabile e FAIR-oriented, intende contribuire anche a tale ripensamento, offrendo uno strumento per la riconfigurazione storica e computazionale dei legami tra Africa, Mediterraneo e Atlantico nella lunga durata del primo capitalismo globale. Restituire visibilità a soggetti subalterni, riconoscere il ruolo strategico delle reti finanziarie e mercantili premoderne, valorizzare la complessità documentaria delle fasi arcaiche della tratta schiavista: sono questi i compiti di una Digital History che, lungi dall'essere ancella della tecnica, si pone come frontiera metodologica ed etica, orientata alla costruzione critica della memoria e alla trasformazione consapevole degli strumenti della conoscenza storica.

Bibliografia

Aarseth, Espen J. J. 1997. *Cybertext: Perspectives on Ergodic Literature*. Baltimore: Johns Hopkins University Press.

Agosto, Aldo. 1998. *Christopher Columbus and His Family: The Genoese and Ligurian Documents*. Turnhout, Belgium: Brepols Publishers.

Ahmadi-Assalemi, Gabriela, Haider Al-Khateeb, Carsten Maple, et al. 2020. "Digital Twins for Precision Healthcare." In *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, edited by Hamid Jahankhani, Stefan Kendzierskyj, Nishan Chelvachandran, and Jaime Ibarra, 133-58. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-35746-7_8.

Armenteros-Martinez, Ivan. 2022. "Enslaved Africans in Medieval and Early Modern Iberia." In *Oxford Research Encyclopedia of African History*. https://doi.org/10.1093/acrefore/9780190277734.013.864.

Ayers, Edward L., and Anne S. Rubin. 2000. *The Valley of the Shadow. Two Communities in the American Civil War - The Eve of War*. Har/Cdr edizione. New York: WW Norton & Co. Inc.

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43 (3-4): 465-82.

Banfi, Fabrizio, Elena Dellù, Chiara Stanga, et al. 2023. "Representing Intangible Cultural Heritage of Humanity: From the Deep Abyss of the Past to Digital Twin and XR of the Neanderthal Man and Lamalunga Cave (Altamura, Apulia)." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-M-2–2023 (giugno):171-81. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-171-2023.

Bleichmar, Daniela, and Vanessa R. Schwartz. 2019. "Visual History. The Past in Pictures." *Representations* 145 (1): 1-31. https://doi.org/10.1525/rep.2019.145.1.1.

Bloch, Marc. 1963. *The Historian's Craft: Ou metier d'historien*. Tradotto da Peter Putnam. New York: Knopf.

Bonazza, Giulia. 2023. "Slavery in the Mediterranean." In *The Palgrave Handbook of Global Slavery throughout History*, edited by Damian A. Pargas and Juliane Schiel, 227-42. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-13260-5_13.

Bono, Salvatore. 2017. "Schiavitù mediterranea." *Nuova informazione bibliografica* 56 (4): 675-98. https://doi.org/10.1448/87956.

Bradley, Adam James, Mennatallah El-Assady, Katharine Coles, et al. 2018. "Visualization and the Digital Humanities." *IEEE Computer Graphics and Applications* 38 (6): 26-38. https://doi.org/10.1109/MCG.2018.2878900.

Braudel, Fernand. 2003. *Scritti sulla storia*. Milano: Bompiani.

Braudel, Fernand. 2015. *Storia, misura del mondo*. Il Mulino.

Braudel, Fernand. 2017. *La Méditerranée*. Paris: FLAMMARION.

Braudel, Fernand, e Alfredo Salsano. 1974. *La storia e le altre scienze sociali*. Roma: Laterza.

Brown, Vincent. 2021. "Slave Revolt in Jamaica, 1760-1761." http://revolt.axismaps.com/.

Bush, Vannevar. 1945. "As We May Think." *Atlantic* 176 (1).

Chatbox AI. n.d. Consultato il 19 maggio 2025. https://chatboxai.app/it-IT.

Chaunu, Pierre. 1959. *Séville et l'Atlantique, 1504-1650*: *Structures et conjoncture de l'Atlantique espagnol et hispano-américain (1504-1650). Tome II, volume 1 : La conjoncture (1504-1592)*. 12 voll. Travaux et mémoires. Paris: SEVPEN. https://books.openedition.org/iheal/5807.

Ciofalo, Giovanni, e Silvia Leonzi. 2013. *Homo communicans. Una specie di/in evoluzione*. Roma: Armando Editore.

Clerici, Alberto, e Maurizio De Pra. 2012. *Informatica e web*. EGEA spa.

Correia, Luiz Felipe Ribeiro, Roberto dos Santos Bartholo, Aline Winckler Brufato, and Edney Christian Thomé Sanchez. 2023. "Toward a Digital Twin for Cultural Heritage." In *Advances in Tourism, Technology and Systems: Selected Papers from ICOTTS 2022, Volume 1*, edited by João Vidal Carvalho, António Abreu, Pedro Liberato, and Alejandro Peña, 419-30. Singapore: Springer Nature. https://doi.org/10.1007/978-981-99-0337-5_35.

Darcy, Robert, e Richard C. Rohrs. 1995. *A Guide to Quantitative History*. Westport, Conn.: Praeger.

Delpiano, Patrizia. 2021. *La schiavitù in età moderna*. Gius.Laterza & Figli Spa.

Digital Library. 2021. "Principi F.A.I.R. – Interoperabilità." 28 Ottobre. https://digitallibrary.cultura.gov.it/notizie/principi-f-a-i-r-interoperabilita/.

Dollar, Charles M., and Richard J. Jensen. 1974. *Historian's Guide to Statistics; Quantitative Analysis and Historical Research*. Huntington, N.Y.: R.E. Krieger Pub. Co.

Ebbrecht-Hartmann, Tobias, Noga Stiassny, and Lital Henig. 2023. "Digital visual history: historiographic curation using digital technologies." *Rethinking History* 27 (2): 159-86. https://doi.org/10.1080/13642529.2023.2181534.

Eijnatten, Joris van, Toine Pieters, and Jaap Verheul. 2013. "Big Data for Global History: The Transformative Promise of Digital Humanities." *BMGN - Low Countries Historical Review* 128 (4): 55-77. https://doi.org/10.18352/bmgn-lchr.9350.

El Haoud, Naima, and Zineb Bachiri. 2019. "Stochastic Artificial Intelligence benefits and Supply Chain Management inventory prediction." In *2019 International Colloquium on Logistics and Supply Chain Management (LOGISTIQUA)*, 1-5. https://doi.org/10.1109/LOGISTIQUA.2019.8907271.

El Saddik, Abdulmotaleb. 2018. "Digital Twins: The Convergence of Multimedia Technologies." *IEEE MultiMedia* 25 (2): 87-92. https://doi.org/10.1109/MMUL.2018.023121167.

Eltis, David, and David Richardson. 2008. *Extending the Frontiers: Essays on the New Transatlantic Slave Trade Database*. Yale University Press.

Ennals, Richard. 1985. *Artificial Intelligence: Applications to Logical Reasoning and Historical Research*. Ellis Horwood.

Erwin, Brittany. 2020. "Digital Tools for Studying Empire: Transcription and Text Analysis with Transkribus." *Not Even Past*, 6 November 2020. https://notevenpast.org/digital-tools-for-studying-empire-transcription-and-text-analysis-with-transkribus/.

Fiume, Giovanna. 2012. *Schiavitù Mediterranee. Corsari, Rinnegati e Santi Di Età Moderna*. Bruno Mondadori.

Franzosi, Roberto. 2017. "A third road to the past? Historical scholarship in the age of big data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 50 (4): 227-44. https://doi.org/10.1080/01615440.2017.1361879.

Fuller, Aidan, Zhong Fan, Charles Day, and Chris Barlow. 2020. "Digital Twin: Enabling Technologies, Challenges and Open Research." *IEEE Access* 8:108952–71. https://doi.org/10.1109/ACCESS.2020.2998358.

Furet, François. 1968. «Sur quelques problèmes posés par le développement de l'histoire quantitative.» *Social Science Information* 7 (1): 71-82.

Furet, François. 1971. «L'histoire quantitative et la construction du fait historique.» *Annales. Histoire, Sciences Sociales* 26 (1): 63-75.

Furet, François. 1981. «Il quantitativo in Storia.» In *Fare storia. Temi e metodi della nuova storiografia*, a cura di Jacques Le Goff e Pierre Nora. Torino: Einaudi.

Gabellone, Francesco. 2022. "Digital Twin: A New Perspective for Cultural Heritage Management and Fruition." *Acta IMEKO* 11 (1): 1-7. https://doi.org/10.21014/acta_imeko.v11i1.1085.

Galić, Radoslav, Elvir Čajić, Zvezdan Stojanovic, and Dario Galić. 2023. "Stochastic Methods in Artificial Intelligence." *Research Square*. https://doi.org/10.21203/rs.3.rs-3597781/v1.

Gautier, Dassonneville, Adèle Huguet, Marie-Laure Massot, et al. 2022. «Compte-rendu de la journée d'étude 'Point HTR 2022' Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d'archives.» Report, CAPHES - UMS 3610 CNRS/ENS; AOROC. https://hal.science/hal-03692413.

Gioffrè, Domenico. 1962. "Il commercio d'importazione genovese alla luce dei registri del dazio, 1495-1537." In *Studi in onore di Amintore Fanfani*, vol. 5. Giuffrè.

Gioffrè, Domenico. 1971. *Il mercato degli schiavi a Genova nel secolo XV*. Bozzi.

Github. n.d. "Ollama." Consultato il 19 maggio 2025. https://ollama.com/.

Gonzalez, Teofilo, Jorge Diaz-Herrera, and Allen Tucker. 2014. *Computing Handbook, Third Edition: Computer Science and Software Engineering*. CRC Press.

Graham, Shawn, Ian Milligan, and Scott Weingart. 2015. *Exploring Big Historical Data. The Historian's Macroscope*. Imperial College Press. https://www.perlego.com/book/839949/exploring-big-historical-data-the-historians-macroscope-pdf.

Grieves, Michael W. 2023. "Digital Twins: Past, Present, and Future." In *The Digital Twin*, edited by Noel Crespi, Adam T. Drobot, and Roberto Minerva, 97-121. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-21343-4_4.

Hutson, James, Joe Weber, and Angela Russo. 2023. "Digital Twins and Cultural Heritage Preservation: A Case Study of Best Practices and Reproducibility in Chiesa dei SS Apostoli e Biagio." *Art and Design Review* 11 (February). https://doi.org/10.4236/adr.2023.111003.

Itzcovich, Oscar. 1989. "Lo storico e il database." *Quaderni storici* 24 (70 (1)): 321-25.

Itzcovich, Oscar. 1993. *L'uso del calcolatore in storiografia*. Milano: FrancoAngeli.

Itzcovich, Oscar. 1996. "Dal mainframe al personal, il computer nella storiografia quantitativa." *Storia & computer*.

Jupyter. n.d. Consultato il 19 maggio 2025. https://jupyter.org.

Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. "Transkribus. A Service Platform for Transcription, Recognition and Retrieval of Historical Documents." In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 19-24. https://doi.org/10.1109/ICDAR.2017.307.

Kaplan, Frédéric, and Isabella di Lenardo. 2017. "Big Data of the Past." *Frontiers in Digital Humanities* 4. https://doi.org/10.3389/fdigh.2017.00012.

Kennedy, S. Wright, Jessica C. Kuzmin, and Benjamin Jones. 2017. "New Methods in the History of Medicine: Streamlining Workflows to Enable Big-Data History Projects." *Medical History* 61 (3): 477-80. https://doi.org/10.1017/mdh.2017.54.

Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. "eScriptorium: An Open Source Platform for Historical Document Analysis." In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2:19-19. https://doi.org/10.1109/ICDARW.2019.10032.

Lu, Jessica H., and Caitlin Pollock. "Design, Development, and Documentation: Hacking TEI for Black Digital Humanities." Maryland Institute for Technology in the Humanities (MITH). https://mith.umd.edu/digital-dialogues/dd-fall-2019-jessica-h-lu-caitlin-pollock/.

Maleki, Negar, Balaji Padmanabhan, and Kaushik Dutta. 2024. "AI Hallucinations: A Misnomer Worth Clarifying." In *2024 IEEE Conference on Artificial Intelligence (CAI)*, 133-38. https://doi.org/10.1109/CAI59869.2024.00033.

Marengo, Emilio. 1901. "Genova e Tunisi (1388-1515). Relazione storica del socio avv. Emilio Marengo, sotto-archivista nel R. Archivio di Stato di Genova." In *Atti della Società Ligure di Storia Patria, XXXII. Genova*.

Massot, Marie-Laure, Arianna Sforzini, and Vincent Ventresque. 2019. "Transcribing Foucault's handwriting with Transkribus." *Journal of Data Mining and Digital Humanities* Atelier Digit_Hum (marzo). https://hal.archives-ouvertes.fr/hal-01913435.

McCarthy, John, Nathaniel Rochester, Marvin L. Minsky, and Claude E. Shannon. 2006. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." AI Magazine 27 (4). https://doi.org/10.1609/aimag.v27i4.1904.

Milioni, Nikolina. 2020. "Automatic Transcription of Historical Documents. Transkribus as a Tool for Libraries, Archives and Scholars." Student Thesis, Master's Programme in Digital Humanities. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-412565.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, et al. 2019. "Transforming scholarship in the archives through handwritten text recognition. Transkribus as a case study." *Journal of Documentation* 75 (5): 954-76. https://doi.org/10.1108/JD-07-2018-0114.

National African-American Reparations Commission, n.d. "The Atlantic Slave Trade in Two Minutes." Consultato il 19 maggio 2025. https://reparationscomm.org/reparations-news/atlantic-slave-trade-in-two-minutes/.

North American Slave Narratives. n.d. Consultato il 19 maggio 2025. https://docsouth.unc.edu/neh/index.html.

Northeastern University Library. n. d. "Early Caribbean Digital Archive." Consultato il 19 maggio 2025. https://ecda.northeastern.edu/#fl-main-content.

Nowatzki, Robert. 2021. "From Datum to Databases: Digital Humanities, Slavery, and Archival Reparations." *The American Archivist* 83 (2): 429-48. https://doi.org/10.17723/0360-9081-83.2.429.

Rabus, Achim. 2019. "Recognizing Handwritten Text in Slavic Manuscripts, a Neural-Network Approach Using Transkribus." *Scripta & E-Scripta*, fasc. 19, 9-32.

Rau, Virginia. 1957. "A family of Italian merchants in Portugal in the XVth century: the Lomellini." In *Studi Armando Sapori*, 1:715–26.

READ-COOP. s.d. "Transkribus." Consultato 30 agosto 2022. https://readcoop.eu/it/transkribus/.

Reilly, Edwin D., Anthony Ralston, and David Hemmendinger. 2000. *Encyclopedia of Computer Science*. Nature Publishing Group.

Rosenzweig, Roy. 1997. "Brave New World or Blind Alley? American History on the World Wide Web." *Journal of American History* 84 (1): 132-55.

Rosenzweig, Roy. 2001. "The Road to Xanadu: Public and Private Pathways on the History Web." *Journal of American History* 88 (2): 548-79. https://doi.org/10.2307/2675105.

Ruis, Andrew R., and David Williamson Shaffer. 2017. "Annals and Analytics: The Practice of History in the Age of Big Data." *Medical History* 61 (1).

Salvagno, Michele, Fabio Silvio Taccone, and Alberto Giovanni Gerli. 2023. "Artificial Intelligence Hallucinations." *Critical Care* 27 (1): 180. https://doi.org/10.1186/s13054-023-04473-y.

Santoro, Raffaele. 2015. "I grandi archivi: un patrimonio di Big Data." In *La grande bellezza dei grandi dati. Frontiere di fruizione del patrimonio archivistico per studiosi e società civile*. Roma: Archivio di Stato di Roma.

Schiuma, Giovanni, and Daniela Carlucci. 2018. *Big Data in the Arts and Humanities. Theory and Practice*. Auerbach Publishers.

Schlagdenhauffen, Régis. 2020. "Optical Recognition Assisted Transcription with Transkribus: The Experiment concerning Eugène Wilhelm's Personal Diary (1885-1951)." *Journal of Data Mining & Digital Humanities* Atelier Digit_Hum. https://doi.org/10.46298/jdmdh.6249.

Schwartz, Stuart B. 2004. *Tropical Babylons: Sugar and the Making of the Atlantic World, 1450-1680*. Univ of North Carolina Press.

Sgroi, Salvatore. 1979. Recensione di *La schiavitù [domestica] in Sicilia nel XVI secolo*, Corrado Avolio. Catania, Bacina 1978.

Sicking, Louis, and Alain Wijffels. 2020. "Introduction Flotsam and Jetsam in the Historiography of Maritime Trade and Conflicts." In *Conflict Management in the Mediterranean and the Atlantic, 1000-1800: Actors, Institutions and Strategies of Dispute Settlement*. Brill Nijhoff. https://brill.com/display/title/39048.

Slave Voyages. n.d. "Explore the Origins and Forced Relocations of Enslaved Africans Across the Atlantic World." Consultato il 19 maggio 2025. https://www.slavevoyages.org/.

Spina, Salvatore. 2020. *Archivi nell'era delle Digital Humanities, dei Big Data e della Genetica*. Viagrande (Catania): Algra.

Spina, Salvatore. 2021. "Datificazione delle fonti storiche per la digital history delle pandemie." *Umanistica Digitale* 10.

Spina, Salvatore. 2022a. *Digital History. Metodologie informatiche per la ricerca storica*. Napoli: Edizioni Scientifiche Italiane.

Spina, Salvatore. 2022b. "Historical Network Analysis & Htr Tool. Per un approccio storico metodologico digitale all'archivio Biscari di Catania." *Umanistica Digitale*, fasc. 14, 163-81. https://doi.org/10.6092/issn.2532-8816/15159.

Spina, Salvatore. 2023a. "Artificial Intelligence in archival and historical scholarship workflow: HTR and ChatGPT." *Umanistica Digitale*. https://doi.org/10.48550/arXiv.2308.02044.

Spina, Salvatore. 2023b. "Biscari Epistolography. From Archive to the Website." *DigItalia* 2.

Spina, Salvatore. 2024. "Digitality as a longue durée historical phenomenon." *Umanistica Digitale* 18:1-25.

Staley, David J. 2015. *Computers, visualization, and History. How New Technology Will Transform Our Understanding of the Past*. 1 voll. London and New York: Routledge, Taylor & Francis Group.

Taviani, Carlo. 2024. "The Genoese Merchant Network and Gold (Ea.1450-1 5 30)." In *Global Gold: Aesthetics, Material Desires, Economies in the Late Medieval and Early Modern World*, edited by Thomas B. F. Cummins. Florence: Villa I Tatti.

The Maryland State Archives, n.d. "Legacy of Slavery in Maryland." Consultato il 19 maggio 2025. https://slavery.msa.maryland.gov/.

Theibault, John. 2016. "Visualizations and Historical Arguments." In *Writing History in the Digital Age*, edited by Jack Dougherty and Kristen Nawrotzki. University of Michigan Press.

Trasselli, Carmelo. 1969. "Genovesi in Sicilia." In *Atti della Società Ligure di Storia Patria, nuova serie*, 153-78. Genova: Società Ligure di Storia Patria. https://www.storiapatriagenova.it/Scheda_vs_info.aspx?Id_Scheda_Bibliografica=711.

Trasselli, Carmelo. 1972. "Considerazioni sulla schiavitù in Sicilia alla fine del Medioevo." *Clio* 8:67-90.

Trinkle, Dennis A, and Scott A. Merriman. 2000. *The History Highway 2000, a Guide to Internet Resources*. Armonk, N.Y.: M.E. Sharpe.

Verlinden, Charles. 1970. *The Beginnings of Modern Colonization: Eleven Essays*. Ithaca: Cornell Univ Pr.

Wang, Fuzhang, Ayesha Sohail, Wing-Keung Wong, Qurat Ul Ain Azim, Shabieh Farwa, and Maria Sajad. 2023. "Artificial intelligence and stochastic optimization algorithms for the chaotic datasets." *Fractals* 31 (06): 2240175. https://doi.org/10.1142/S0218348X22401752.

World Wide Web. n.d. Consultato il 19 maggio 2025. http://info.cern.ch/hypertext/WWW/TheProject.html.

Xu, Zhiwei, and Jialin Zhang. 2022. *Computational Thinking: A Perspective on Computer Science*. Springer Nature.

Rubriche

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-549-3
DOI 10.57574/59655493XX
pag. 137-140 (gennaio-giugno 2025)
Open access article licensed under CC-BY-NC-SA

Recensione del volume *ClaG – Classificazione dei giochi per ludoteche e biblioteche*

Antonietta Folino*

Il volume di Carlo Bianchini e Paolo Munini (2024) si propone di illustrare le caratteristiche, gli scopi e le modalità di applicazione di *ClaG*, un sistema classificatorio per l'organizzazione dei giochi nell'ambito delle ludoteche e delle biblioteche.

Seppur applicata ad oggetti diversi dalle risorse informative più consuete e tradizionali per le quali le necessità di organizzazione e di recupero da parte degli utenti finali sono note e appaiono stringenti e indubbie, la classificazione si propone le medesime finalità di ogni altro sistema orientato ad identificare le caratteristiche rilevanti degli oggetti da classificare in base alle quali costituire degli insiemi omogenei che li raggruppino. Nello specifico, *ClaG* si connota come uno strumento pratico, diretto alla ricerca, all'identificazione e alla consegna di uno o più giochi – in base alle caratteristiche che gli stessi possiedono – da parte del personale dei luoghi di conservazione e degli utenti finali interessati al loro utilizzo.

La predisposizione del sistema di classificazione ha richiesto anche in questo caso la preliminare definizione del suo oggetto di interesse al fine di delimitare la propria applicabilità e il proprio ambito di intervento. La distinzione tra "attività ludica" e "gioco", a partire dalla trasposizione concettuale degli equivalenti in lingua inglese *play* e *game* – spesso tradotti in italiano con il solo termine "gioco", la cui sfera semantica non tiene traccia delle sfumature di significato tra i due – permette di identificare come oggetto di interesse di *Clag* i «giochi che si trovano più frequentemente nelle collezioni delle ludoteche e delle biblioteche» (Bianchini e Munini 2024, 11). L'individuazione delle proprietà intrinseche che caratterizzano i giochi, funzionali alla loro classificazione, si è basata su un'attenta rassegna della letteratura e delle definizioni che diversi studiosi del dominio hanno fornito per circoscrivere e distinguere

* Dipartimento di Culture, Educazione e Società, Università della Calabria, Rende (CS), Italia. antonietta.folino@unical.it. ORCID: 0000-0002-5882-6150.

le differenti tipologie di giochi esistenti, così come le loro finalità e le abilità di volta in volta richieste.

In merito alle scelte di classificazione compiute dagli autori, l'applicazione di un approccio a faccette risulta vincente proprio in considerazione delle molteplici caratteristiche in base alle quali i giochi possono essere classificati e/o cercati. Le proprietà di pluridimensionalità, flessibilità e scalarità proprie di tale metodo (Rosati 2003), infatti, consentono, nel caso di specie, una descrizione e una ricerca a partire da molteplici punti di vista, incontrando le esigenze degli utenti, potenzialmente molto eterogenei, e le esigenze dei professionisti, soprattutto in risposta alla necessità di classificare nuovi giochi sfruttando lo schema esistente e non introducendo modifiche che ne stravolgerebbero l'impianto di base.

L'applicazione della classificazione a faccette proposta nello specifico contesto si distingue per il rispetto, oltre che dei principi sopra menzionati, anche di quei criteri propri della teoria classificatoria di Ranganathan (2006), alla quale gli autori fanno esplicito riferimento, e che non sempre contraddistinguono l'architettura dell'informazione negli ambienti soprattutto digitali che pur sembrano aderire ad una tale metodologia. Si tratta nello specifico, del principio di esclusività – che garantisce che le categorie individuate siano mutuamente esclusive evitando sovrapposizione concettuale tra le stesse e quindi favorendo una univoca collocazione di ogni elemento – e del principio di crescente complessità che determina l'ordine con il quale le categorie stesse vengono elencate e utilizzate per la classificazione.

In particolare, le faccette individuate dagli autori sono: a) *Spazio*, direttamente assimilabile all'omonima categoria del PMEST definita dal bibliotecario indiano e qui relativa all'identificazione dello spazio di gioco; b) *Materiali*, ovvero gli oggetti necessari alla pratica di un gioco, la cui applicazione prevede la preliminare individuazione del materiale di volta in volta prevalente o più importante; c) *Ambientazione*, che richiede l'identificazione di un tema chiaro e preciso associabile al gioco; d) *Esito*, ovvero la conclusione che ci si attende al termine del gioco; e) *Genere*, basata sull'analisi dell'atteggiamento del giocatore verso il gioco con un chiaro riferimento alle diverse Intelligenze definite da Howard Gardner per la classificazione dei giochi di abilità e f) *Età*, la cui possibile soggettività viene superata facendo ricorso a fonti autorevoli e oggettive e superando il giudizio del singolo classificatore.

Uno sviluppo interessante del sistema di classificazione consiste nella parallela definizione del dataset *Cla-G*, un'istanza di Wikibase – il software su cui si basa Wikidata – per favorire l'integrazione con le tecnologie e i linguaggi del web semantico, quali i linked open data, il cui impiego è sempre più diffuso nel dominio della descrizione bibliografica con il vantaggio di aprire e connettere i dati provenienti da più database in una fitta rete di relazioni che favorisce sensibilmente il recupero delle informazioni. Tale interoperabilità sfrutta

altresì le connessioni reciproche tra *ClaG* e i cataloghi locali delle singole ludoteche o biblioteche consentendo agli utenti un'esperienza di navigazione completa e integrata.

Dal punto di vista dell'applicazione pratica l'uso delle tavole e delle regole di classificazione si basa, in maniera del tutto simile alla *Colon Classification*, su una notazione specifica per ciascuna faccetta e su un sistema combinatorio che consente di costruire di volta in volta il codice più appropriato per ogni specifico gioco, applicando la fase di sintesi del noto metodo analitico-sintetico proprio di questo approccio classificatorio e che lo contraddistingue da quello enumerativo dei sistemi gerarchici. Nella costruzione della sequenza delle notazioni gli autori raccomandano la massima aderenza ai principi e alle regole definite per disincentivare utilizzi personalizzati che – seppur possibili – potrebbero compromettere la cooperazione nell'uso della classificazione da parte dei singoli attori a vario titolo coinvolti nel processo.

Per quanto sia la fase di analisi che la fase di sintesi possano presentare un'intrinseca complessità derivante dalla corretta, coerente e completa analisi degli oggetti di classificazione, la trattazione della tematica da parte degli autori presenta – a fronte di un adeguato apparato bibliografico relativo alle tematiche e agli studi inerenti i giochi e le attività ludiche – un livello di tecnicismo adatto ad un pubblico non necessariamente esperto in discipline biblioteconomiche e avvezzo alla pratica della classificazione. La commistione di competenze e conoscenze esperte e il diverso background degli autori hanno senz'altro contribuito a raggiungere tale obiettivo e a rappresentare entrambe le sfere d'azione, quella biblioteconomica e quella socio-pedagogica.

Infine, sebbene non tutto il sistema sia stato al momento sviluppato e necessiti quindi di aggiunte e precisazioni in alcune tavole di classificazione, la sua evoluzione appare già ben definita e viene anticipata dagli autori, che ne prefigurano le successive espansioni – quali ad esempio la specificazione della precisa ambientazione di un gioco che allo stato attuale prevede solo il valore "gioco simulato" ricorrendo probabilmente alla Classificazione Decimale Dewey per l'impiego dei codici relativi a soggetti potenzialmente generici e molto eterogeni – proprio grazie alla già menzionata analisi della letteratura in materia di giochi e soprattutto all'analisi delle specificità dei giochi stessi e delle caratteristiche funzionali ad una loro più completa descrizione e classificazione.

Riferimenti bibliografici

Bianchini, Carlo, e Paolo Munini. 2024. *ClaG – Classificazione dei giochi per ludoteche e biblioteche*. Comune di Udine, Archivio italiano dei Giochi, Centro per la documentazione della cultura ludica.

Ranganathan, Shiyali Ramamrita. 2006. *Colon Classification. Sixth edition*. Reprinted and Published by Ess Ess Publications for Sarada Ranganathan Endowment for Library Science. Bangalore.

Rosati, Luca. 2003. "La classificazione a faccette fra Knowledge Management e Information Architecture (parte I)." *it.Consult*.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-549-3
DOI 10.57574/59655493XX
pag. 141-144 (gennaio-giugno 2025)
Open access article licensed under CC-BY-NC-SA

Il macellaio pragmatista

Claudio Gnoli*

Organizzare la conoscenza è per prima cosa individuare, nel mondo "tutto attaccato" di cui dicevamo tre puntate fa, delle classi di entità che possano fungere da unità dei nostri sistemi. Naturalmente vorremmo farlo in modo non arbitrario, vorremmo cioè «dividere per specie seguendo le articolazioni naturali e cercare di non spezzare alcuna parte, alla maniera di un cattivo macellaio» come dice un celebre passo del *Fedro* di Platone. Se una massa muscolare si sviluppa nello spazio tra due ossa, non sembra una buona idea tagliare trasversalmente rompendo le ossa. Metafora sgradevole per i vegetariani ma efficace: intuitivamente ci aspettiamo che il mondo abbia delle articolazioni determinate, e che stia a noi scoprirle e rifletterle nei nostri Knowledge Organization Systems (KOS).

Sempre che tali articolazioni davvero esistano "là fuori" e che le nostre facoltà ci consentano di coglierle… Qui sta il perenne dibattito filosofico fra oggettivismo e soggettivismo, che si può applicare anche ai concetti di informazione e documento (Ridi 2015). Gran parte della filosofia evita visioni completamente idealistiche secondo le quali il mondo starebbe solo nella nostra testa, e oscilla piuttosto fra il *realismo* e il *costruttivismo*: un mondo là fuori sembra in effetti esistere, ma i modi in cui lo classifichiamo dipendono anche dalle nostre particolari strutture culturali; culture diverse dalla nostra possono classificare le donne, il fuoco e le cose pericolose in una stessa categoria, come nel titolo della monografia di George Lakoff.

Di fatto possiamo "fare a pezzi il mondo" (per usare le parole dell'ontologo Roberto Poli) in base a innumerevoli criteri alternativi: possiamo cioè dividere le cose secondo certe proprietà come il colore, o secondo tutt'altre come il numero atomico o l'abbondanza. Come distinguere allora i tagli buoni da quelli cattivi?

* Biblioteca della scienza e della tecnica, Università di Pavia, claudio.gnoli@unipv.it. ORCID: 0000-0002-4721-7448.

Anche limitandoci al senso comune, ci sono svariati modi leciti di classificare, ognuno con un suo senso. Posso dividere le piante in commestibili e tossiche, in dicotiledoni e monocotiledoni, in tropicali e artiche, a seconda del mio obiettivo. Come facciamo allora a sapere quale di queste classificazioni è quella che meglio aderisce alla realtà?

La moderna epistemologia della classificazione prende avvio da autori ottocenteschi come John Stuart Mill, secondo il quale

> le proprietà secondo cui gli oggetti vengono classificati dovrebbero essere, se possibile, quelle che sono causa di molte altre proprietà [...]. Una classificazione così formata è propriamente scientifica o filosofica, ed è comunemente chiamata una classificazione o disposizione Naturale, in contrapposizione a Tecnica o Artificiale": è la classificazione che ci interessa realizzare "quando stiamo studiando gli oggetti non per un particolare scopo pratico, ma per nell'interesse di estendere la nostra conoscenza riguardo all'insieme delle loro proprietà e relazioni (Mill 1843, 7, 126).

Nell'ambito del realismo contemporaneo, come spiega Zdenka Brzović (n.d.) nella voce "Natural kinds" dell'*Internet encyclopedia of philosophy*, esistono tre posizioni: dall'essenzialismo, secondo cui ogni classe è caratterizzata da un'essenza di proprietà fisse ben individuabili (l'oro è la classe delle sostanze i cui atomi hanno 79 protoni: nulla più e nulla meno); al meno rigido riconoscimento di *grappoli di proprietà* (una specie animale è caratterizzata da un insieme di proprietà, qualcuna delle quali può anche mancare: sebbene i gatti abbiano in genere la coda, anche quelli dell'isola di Man sono comunque gatti); fino al *realismo promiscuo* o *pluralistico* di John Dupré, secondo il quale tutti i criteri di raggruppamento alternativi sono ugualmente validi in quanto ognuno intercetta un certo aspetto della realtà. In particolare, P.D. Magnus (2012) argomenta che ognuno dei raggruppamenti è valido nell'ambito di un certo *dominio* di applicazione.

Siamo ancora nel realismo, ma nella sua variante pragmatista, di cui nel nostro settore è seguace Birger Hjørland, caposcuola appunto dell'*analisi di dominio*. A questo punto tutti i KOS adottati nei diversi dominî della conoscenza assumono lo stesso valore, non potendo esistere la "visuale da nessun luogo" immaginata da Thomas Nagel. La nostra epistemologia non può che essere un'*epistemologia sociale* (Hjørland 2024), i cui principî, più che nelle facoltà razionali dei singoli ricercatori, si trovano nei contesti storici che collettivamente danno luogo ai paradigmi conoscitivi nelle diverse epoche, regioni e scuole. Oggi classifichiamo come mammifero la stessa balena che in passato classificavamo come pesce, ed ancora oggi possiamo accettare di chiamarla "pesce" nel dominio dell'alimentazione o in quello dell'etnografia, che la può considerare come pesce in una filastrocca tradizionale.

Se tutte le proprietà di volta in volta prese in considerazione hanno la stessa importanza, non c'è dunque modo di ambire ad una classificazione naturale

come quella definita da Mill? Forse una speranza ci resta nell'individuare delle proprietà "che sono causa di molte altre proprietà", ossia in un approccio storico-genealogico. Anche Herbert Spencer (1864) precisò che i raggruppamenti migliori sono quelli che hanno in comune più proprietà possibile. Nei KOS le proprietà sono spesso espresse con faccette, sicché potremmo dire che le classi migliori sono quelle che condividono un maggior numero di faccette: è vero che tanto i sassi quanto gli animali hanno un colore, però gli animali hanno anche faccette come organi, stadi di crescita, metabolismo e comportamenti, che non sono applicabili ai sassi, il che indica che sia utile avere una classe degli animali distinta da quella dei sassi.

Spencer aggiunge che dovremmo considerare le proprietà più "radicali", dalle quali le altre dipendono. Dopotutto è il fatto di avere 79 protoni ed altrettanti elettroni a fare dell'oro la sostanza che è, determinando attraverso il gioco delle attrazioni e repulsioni elettromagnetiche anche le sue proprietà macroscopiche. È vero che possiamo classificare l'oro anche fra i beni conservati dalle banche o fra le materie prime della gioielleria, ma la sua "sede di definizione univoca", come si espresse Jason Farradane nelle leggendarie riunioni del Classification Research Group, sarà la classe delle sostanze chimiche. La balena è certo anche un materiale per gli irresponsabili produttori di olio, un'attrazione per i visitatori degli acquari e un personaggio nel *Pinocchio* della Disney, ma non potrebbe essere tutte queste cose se prima non fosse un mammifero cetaceo. Possiamo dunque concepire un KOS in cui il concetto di balena sia definito nella classe degli animali, e possa quindi essere riutilizzato anche in classi successive combinando con queste la notazione per gli animali balene oppure – nel caso che nel nuovo dominio convenga introdurre una notazione diversa, ad esempio per porre le balene letterarie vicine ai pesci letterari – istituendo dei rinvii "vedi anche" verso la sede di definizione univoca.

In questi esempi, a ben guardare, le diverse classificazioni si differenziano perché ciascuna considera un diverso livello di realtà: quelli atomico, economico e artistico nel caso dell'oro; quelli vivente, tecnologico, turistico e artistico nel caso della balena. Normalmente la sede di definizione univoca sta al livello di realtà più basso che spiega tutti i livelli successivi, ossia quello atomico per l'oro e quello vivente per la balena.

Al tempo stesso, però, la sede univoca non sta nemmeno ad un livello ancora più basso: non sarebbe corretto classificare le balene in base agli atomi che le compongono, perché le proprietà specifiche delle balene, come avere dei fanoni e uno sfiatatoio, compaiono solo a partire dal livello vivente – gli atomi non hanno sfiatatoi. «La biologia a livello molecolare è chimica organica», ripeteva un po' ossessivamente un mio docente nel sottolineare l'importanza del suo corso universitario, per l'appunto quello di Chimica organica: non aveva torto, ma peccheremmo di riduzionismo se per questo classificassimo le balene in base ai loro atomi…

Riferimenti bibliografici

Brzović, Zdenka. n.d. "Natural Kinds." Internet Encyclopedia of Philosophy. https://iep.utm.edu/nat-kind/. Consultato il 15 maggio 2025.

Hjørland, Birger. 2024. "Social epistemology." In *Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli. https://www.isko.org/cyclo/se.

Magnus, P.D. 2012. *Scientific enquiry and natural kinds*. Palgrave-Macmillan.

Mill, John Stuart. 1843. *A System of Logic, Ratiocinative and Inductive*. New York: Harper and Brothers Publishers.

Ridi, Riccardo. 2015. "Livelli di irrealtà. Oggettività e soggettività nell'organizzazione della conoscenza." *Bibliotime* 18 (2). https://www.aib.it/aib/sezioni/emr/bibtime/num-xviii-2/ridi.htm.

Spencer, Hebert. 1864. *The classification of the sciences*. New York: D. Appleton and Company.