

Collective intelligence. AI for expanding the information content offered to users: the SBN Sommerso Project

Irene Piccari^(a)

a) Central Institute for the Union Catalog of Italian Libraries and for Bibliographic Information (ICCU)

Contact: Irene Piccari, irene.piccari@cultura.gov.it

Received: 16 July 2025; **Accepted:** 21 October 2025; **First Published:** 15 January 2026

ABSTRACT

The SBN Sommerso project was developed with the aim of integrating bibliographic heritage not yet described in the centralised database into the collective catalog of the National Library Service (Servizio Bibliotecario Nazionale, SBN). To enrich the catalog, the project involves an automated process based on the use of artificial intelligence techniques for the processing of UNIMARC records which, when compared with the data in the SBN, are recognised as identical, similar or new, thus making new bibliographic data or new information relating to library holdings available, while ensuring that the uniqueness of the bibliographic record within the SBN Index is maintained. The automation of activities, which is necessary considering the amount of data managed, is achieved through the use of specific machine learning algorithms, designed and developed for the I.PaC infrastructure and adapted within SBN Sommerso through training on data and metadata specific to the bibliographic domain.

For the clustering and deduplication of entities represented in the records, namely Manifestation, Agent, Series, Work, Subject and Printer's Mark, the algorithms took into account all the specific metadata for each element. The output required from the system involves presenting the user with groups of identical entities and identifying new ones. The system also groups entities that are assessed as similar and presents them to the domain expert user for verification of the results and correct modulation of the evaluative choices made by the AI, for an increasingly rich, accurate and reliable collective catalog.

KEYWORDS

Collaborative cataloging; Artificial intelligence; National Library Service; Bibliographic information; Machine learning.

The SBN Sommerso project: introduction

The SBN Sommerso project was defined and developed by the Central Institute for the Union Catalog of Italian Libraries and for Bibliographic Information (ICCU)¹ as part of a broader initiative devoted to enriching the SBN union catalog² within the evolutionary developments of the collective database of the Italian library network (SBN Index).³

SBN Sommerso's objective is to feed the SBN Index with UNIMARC⁴ data not yet present in the union catalog, through an automated procedure dedicated to recognizing new, identical, and similar records to import into the SBN database, thereby making new bibliographic information or new holdings information from libraries available to its vast user base. The process was carried out with the support of comparison algorithms developed on the basis of rules defined by domain experts as necessary to make evaluative decisions about the degree of equality between distinct records, and with the development of machine learning models, based on statistical-probabilistic computation. As is well known, collaborative cataloging — on which the cooperative model of SBN is based — implies and requires the uniqueness of each record in the Index and its non-duplication, in order to optimize cataloging of the bibliographic heritage and thus, ultimately, to ensure the availability of cataloging records for users seeking to retrieve information and to enable the development of more efficient services. In SBN, this organizational model corresponds to the relational data architecture chosen from the very beginning of the network's development as the technological support for nationwide cooperation. The technological model made available for comparing SBN data and UNIMARC data provided for the synergistic use of rule-based algorithms and machine learning. Both models—the rule-based and the machine learning-based—are used in the process of identifying new, identical, and similar records, and they are applied to different entities: rule-based algorithms for authority entities and machine learning algorithms for entities of the Manifestation type.

How can we decide if data encoded in two different records refer to the same entity? What are the essential characteristics of the entity being compared? Where, in the specific data encoding format, is the information on which to base the evaluation? Which sources in the cataloging domain are most relevant for selecting reliable information? How do we interpret the specific weight of a certain data element in context? These are some of the pieces of information that the artificial intelligence must be able to learn during training, so that its computational power can effectively collaborate in building an ever more populated, participatory, and reliable national bibliographic

¹ ICCU is the central institute of the Ministry of Culture that, in compliance with the autonomy of libraries, coordinates, promotes and manages the SBN network, the SBN interlibrary loan service (ILL SBN), the Manus and Edit16 censuses, and the Registry of Italian Libraries. It develops standards and guidelines for cataloging and digitization. For information on the Institute's history and functions, see the ICCU website (dedicated page: <https://www.iccu.sbn.it/it/istituto/>).

² The Servizio Bibliotecario Nazionale (SBN) is the network of Italian libraries promoted by the Ministry of Culture with the cooperation of the Regions and Universities, coordinated by the ICCU. For more information on SBN, see the dedicated page on the ICCU website at: <https://www.iccu.sbn.it/it/SBN/>.

³ The SBN Index is the central database of SBN, populated with the bibliographic records produced by the libraries that are part of the network.

⁴ UNIMARC (Universal Machine Readable Cataloging) is a standard format in the international MARC family created to facilitate the exchange of bibliographic records in electronic format through the encoding of bibliographic elements and a defined logical and physical record structure.

database. The SBN Sommerso project thus represents a further step in the long process of automating bibliographic data management that has characterized SBN's history, seeking the support of a powerful technological tool to optimize a procedure that is strategic for increasing the quantity of data in the SBN Index and, in the long term, for improving the overall quality of the data in the union catalog itself.

As of now, the National Library Service involves 103 hubs (Poli) with the participation of more than seven thousand libraries spread throughout the country, comprising a cataloged collection of about 21 million documents and 115 million location holdings. The cooperative cataloging project demonstrates its vitality through the entry of new hubs and libraries into the network. SBN Sommerso aims to meet the need for automation dictated by the volume of data managed by the Index and the necessity of maintaining control over the uniqueness of records, while at the same time making projects to import UNIMARC databases not yet included in the union catalog more feasible (or even possible). This applies both to data from hubs and libraries already participating in the Service and to entities not yet part of it but which could join, contributing their cataloged records to the Index from the initial stage of participation, ensuring no duplication of entities described in bibliographic records and populating the catalog with new, accurate, and reliable descriptions.

Current data management procedures in SBN already make use of algorithms to define rules capable of identifying and suggesting to the user lists of similar or identical records within the Index database to allow their potential merging. The advances over the current situation include, as will be discussed later, the ability to take into account UNIMARC records external to the SBN Index for comparison, the adoption of AI (artificial intelligence)–based algorithms in addition to rule-based ones,⁵ and the use of the I.PaC graph⁶ as part of the AI service applied to descriptive metadata, of which the SBN Sommerso project is a concrete application.

⁵ Artificial intelligence (AI) is the ability of a machine to exhibit human capabilities such as reasoning, learning, planning, and creativity. The EU Artificial Intelligence Act provides, in Article 3, the following definition of an “AI system”: “‘AI system’ means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.”, as a field of study and as a set of computing technologies, has a long history whose origin can be chronologically placed in the mid-1950s. Traditionally, the beginning of AI as a scientific discipline is traced to the Dartmouth Conference, held in 1956 at Dartmouth College by a group of American mathematicians, among whom was John McCarthy—who coined the term “artificial intelligence” to refer to the research project defined at that event. This project originally aimed to create an artificial intelligence capable of solving problems in all domains, modeled on the abilities attributed to humans. That early ambition, for a so-called *artificial general intelligence* (AGI), was later scaled down, and research focused on systems dedicated to solving problems confined to a single or narrower domain. A rule-based algorithm is one of the earliest strategies used in AI systems, where the steps for solving a problem are first determined and then encoded into a set of rules and sequential steps that make up the algorithm. The limitations of this approach, along with studies conducted in other fields such as neuroscience and cognitive science, led to a shift in research methods with the development of neural networks, whose key feature—and game-changing aspect—is their capacity for *self-learning*. In this approach, the solution to a given problem is not specified a priori by an algorithm; instead, learning occurs on the basis of many examples provided of the correct solution. For this reason, techniques based on neural networks are called *machine learning*. For a clear and effective treatment accessible even to non-experts, see Francesca Rossi (2024), from which the information summarized here is drawn, and Gino Roncaglia (2023a), where, for the topics introduced above, reference is made to chapter 9.

⁶ I.PaC, *Infrastruttura e servizi digitali per il Patrimonio Culturale*, is the national data space designed to archive, manage, and enrich digital cultural heritage.

This paper therefore aims to report on the project activities carried out by ICCU in defining the SBN Sommerso project, which is tied both to the goal of achieving a quantitative expansion of the bibliographic heritage made available to users and—looking to the medium and long term, respectively—to increasing the qualitative level of the data present in the Index and to supporting experiments of potentially strategic value relating to training AI models on rules, criteria, and practices specific to the cataloging domain. These experiments are aimed at determining whether two bibliographic records describe the same entity and at presenting the probabilistic computation of similarity between entities, in a perspective of collaborative integration of AI techniques in the work of librarians.

AI in the Library Context: Theory and Practice – a Brief Overview of the State of the Art

Within the library community, there is a strong interest in technologies related to artificial intelligence, both in terms of analyzing their technical specificities and engaging in more theoretical reflections on the impact of their application in the field, as well as in terms of launching projects to experiment with AI technologies applied to library processes and services. In line with the general trend, attention to the topic received further momentum following the availability of generative AI technologies to the general public after the launch of ChatGPT by OpenAI in November 2022.⁷ In the international library landscape, the IFLA Advisory Committee on Freedom of Access to Information and Freedom of Expression (IFLA FAIFE) published in 2020 the *IFLA Statement on Libraries and Artificial Intelligence* (IFLA 2020) with the intent to outline key issues related to the use of AI technologies in libraries and to highlight the role libraries could assume in a society that was envisioned—indeed expected—to be increasingly permeated by these technologies.⁸ The Statement opens by acknowledging the transformative power inherent in AI technologies and offers a positive outlook on their societal and innovation impacts. Moreover, the vision proposed indicates that AI technologies can be effectively employed by libraries precisely to strengthen their social mission (IFLA 2020, 1), not only by integrating artificial intelligence and machine learning techniques into daily systems and workflows, but also by actively working to educate users in the conscious use of AI. Libraries are encouraged to broaden their aims on this front by providing algorithmic literacy pathways in addition to digital literacy (IFLA 2020, 2), and by working in synergy with the research community to offer necessary expertise in the ethical and safe management and evaluation of data. IFLA thus seems to assign to libraries and their professionals a role that is not only active but also fundamentally important in particularly critical areas related to the use of AI. These areas range from the possibility of mitigating the worst social impacts on employment—through training and dissemination of the skills needed to face ongoing changes—to influencing the design process of AI technologies by intervening upstream on the quality of training data.

⁷ November 22, 2022 is the date on which the U.S. company OpenAI made available an application interface (a chatbot) through which to interact with its generative AI language model, ChatGPT.

⁸ For an in-depth reading of the IFLA Statement in light of current AI developments, see (Lana 2023).

The IFLA Statement therefore presents libraries with a challenging adaptive scenario, but overall a positive assessment of the mediating role that libraries have historically played in the information society. These assumptions are confirmed in a more recent IFLA initiative on AI with the publication of *IFLA Entry Point to Libraries and AI* (2025) by Cox and De Brasdefer (2025),⁹ released as an appendix to the updated 2024 IFLA Internet Manifesto (IFLA 2024). The document opens with an overall positive view of the possibility that AI technologies can support and promote library values in terms of access to and creation of knowledge, and it presents a strong point of view on the role of libraries and on the responsibilities of librarians in safeguarding a responsible, equitable, distributed, and sustainable use of these tools. It is intended, primarily, as an agile support tool for librarians in evaluating the ethical use of AI and as guidance for professional debate (Cox and De Brasdefer 2025, 2). At least two specific points of interest can be highlighted from the document cited. The first concerns the confirmation of a key role assigned to library institutions regarding the spheres of influence they can exercise through activities such as training, advising on choices of different AI models, user education in AI use, and advising on and promoting the responsible use of AI. This strongly reaffirms libraries' mediating capacity in access to information and in the use of information tools, even in the era of generative AIs¹⁰ accessible to a general audience. The second point of interest is identified (by the author of this paper) in the very nature of the document as a practical guide. In listing the possible applications of AI in libraries, its potential risks, and the questions that should guide professionals in considering whether to adopt AI technologies, the document implicitly conveys the premise that libraries have the opportunity to initiate projects to experiment with AI use and to stimulate critical thinking and, above all, a collaborative decision-making process about the role of AI in the library field and the sharing of experiences gained (IFLA 2025, 3). This is set in a context in which libraries are active participants in the collective creation of an artificial intelligence designed for the public good.¹¹ In parallel with reflections on the opportunities and risks arising from the adoption of AI applications from ethical, social, economic, and organizational perspectives—and on strategies libraries can use to contribute in this scenario—the number of projects introducing the use of AI technologies in library activities is growing. The literature, in the form of essays, reports, surveys, and interviews, reflects the wealth of ongoing experiences in a variety of library institutions and

⁹ The document constitutes Appendix I of the updated *IFLA Internet Manifesto* (IFLA 2024).

¹⁰ Generative AIs are neural networks built with advanced machine learning techniques that can even generate content not pre-defined (unlike “traditional” neural networks), starting from an initial input (prompt). They are also characterized by being trained with unsupervised learning techniques, meaning that no human is present to indicate the correct solution to provide as output. ChatGPT is an example of a generative AI. “ChatGPT is a system for interacting with a so-called large language model, that is, a model of human language. In this specific case, with ChatGPT we are conversing with GPT. ... GPT means Generative Pre-trained Transformer; it is generative, i.e. it can generate content (in this case text), it is trained (pre-trained) before being used, and it uses a machine learning technique based on the concept of a transformer, i.e. an innovative architecture based on neural networks.” (Rossi 2024, 63).

¹¹ The concept of “AI for good” appears several times in the IFLA Statement (IFLA 2020, 2 and 9) to indicate various areas in which librarians can apply their skills to define what it means to use artificial intelligence oriented toward the public good. The notion of AI developed as a human-centric technology with the ultimate goal of improving human well-being is present in Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, which establishes harmonized rules on artificial intelligence (AI Act). Finally, for the concept of *AI for social good* see (Floridi et al. 2018).

the diversity of implementations of AI technologies in library products and services.¹² The introduction of artificial intelligence (including generative AI) and machine learning in libraries has opened new prospects for automated creation, extraction, and enrichment of metadata and catalog records. Recent experiments documented in professional and academic literature show both enthusiasm for the technology's potential and a cautious attitude toward its current limitations, identified primarily in issues of accuracy and reliability¹³ in record creation, concerns about conducting descriptive and representational practices in an ethical manner (Corrado 2021), and the instability of results obtained in record generation when using different input data (prompts) or making requests at different times, as seen in experiments with generative AIs like ChatGPT (Taniguchi 2024).

Finally, in April 2024 the *Final Report* produced by the Task Group on AI and Machine Learning for Cataloging and Metadata of the Program for Cooperative Cataloging (PCC)¹⁴ was published (PCC Task Group on AI and Machine Learning for Cataloging and Metadata 2024). The report contains data from a survey conducted to map projects launched to use AI in cataloging and metadata creation activities within the PCC community, and to evaluate their impacts. Based on the survey results, the working group identified a set of themes of particular concern within the community—from the need to clearly dispel notions of an easy or immediate use of AI and machine learning technologies (which in fact require careful consideration of implementation choices and the collaboration of expert catalogers to be effectively integrated as part of technical services in libraries), to concerns about the general lack of resources specifically allocated to AI study and experimentation. The identified themes also formed the basis for developing recommendations and possible actions to coordinate and support the activities undertaken to experiment with AI technologies in the field of cataloging, thus promoting a strategic approach to integrating these tools into library processes and services. The document also provides a map—limited in geographic and chronological scope—of the places and areas of activity in this sector. Among these, a project particularly relevant to this paper is briefly mentioned, due to the substantial similarity of its objectives and technologies used. The reference is to the use of machine learning techniques by the OCLC¹⁵ to improve the quality of the catalog and information retrieval by reducing duplicate

¹² By way of example of the extensive professional literature available on the subject, two contributions are cited here that bring together theoretical reflection on the opportunities and risks inherent in AI technologies with a broad and up-to-date overview of existing projects (Cox and Mazumdar 2022; Balnaves et al. 2025). Additionally, by way of illustration, some contributions can be mentioned regarding the many research efforts in the field of automated subject indexing and classification of documents, using tools ranging from NLP techniques to the use of ChatGPT for the creation of subject headings, highlighting results achieved and limitations found. Although not falling within the specific scope of this work, it was considered useful to cite at least a few of the many works on the topic given the relevance of this area of experimentation in the application of AI to the library sector. Volume 59 (8) (2021) of *Cataloging & Classification Quarterly* is a special issue of the journal entirely dedicated to the theme; see also (Chow, Kao, and Li 2024; Chou and Chu 2022; Saccucci and Salaba 2021; Suominen, Inkinen, and Lehtinen 2022; Manzoni 2023).

¹³ See, for example, (Brzustowicz 2023; Chen and Li 2024).

¹⁴ The Program for Cooperative Cataloging is an international program coordinated by the Library of Congress to promote cooperative cataloging.

¹⁵ Online Computer Library Center (OCLC) is a cooperative library organization that offers products, services, and information systems to the libraries and other institutions in its network. It created and maintains WorldCat, the catalog that records the cataloging data of the libraries participating in the OCLC network.

entries in bibliographic data records and ensuring the persistence of correct data. In the opening of Toves' article on OCLC's *Hanging Together* blog, we read

"Any system aggregating data from thousands of sources needs sophisticated processes that mitigate duplication and ensure the correct data remain".¹⁶

Based on the same premises, the SBN Sommerso project was developed for the automatic deduplication—using machine learning—of similar and identical records, and for the recognition of new entities to be incorporated into the union catalog of the Italian library network.

In the Italian professional context as well, one can note a growing interest in issues related to artificial intelligence and its applications, as evidenced by contributions in the professional literature across various areas of reflection¹⁷ and application of AI technologies. This latter aspect, in particular, was highlighted by a recent comparative survey carried out as part of the European project TLIT4U,¹⁸ dedicated to the theme of artificial intelligence literacy and to the role of librarians in AI literacy pathways.¹⁹ The comparative survey (conducted in four countries: Italy, Bulgaria, Mexico, and Bangladesh) started from the observation that data on the levels and modes of AI adoption in libraries were lacking, and aimed to investigate the perceived value and impact of AI, as well as the state of the art both in actual practices of AI use in libraries and in training—both the training received and that considered necessary for librarians to acquire new competencies. The analysis concludes that there is limited use of AI technologies by librarians and institutional support that is still insufficient to meet staff training and infrastructural requirements needed for the stable integration of AI in libraries, while at the same time underlining the urgency of adopting these technologies in order to maintain the social role of libraries (Tammaro, Zanichelli, and Del Carlo 2024, 13).

SBN Sommerso: Architectural Framework for Project Development

Operationally, the SBN Sommerso project is an application case of services developed in the context of I.PaC (Infrastructure and Digital Services for Cultural Heritage).²⁰ In particular, the project makes use of a specific subset of advanced processing services based on artificial intelligence applied to descriptive metadata, dedicated to deduplicating entities represented in catalog records. I.PaC is the national data space designed to archive, manage, and enrich the digital cultural heritage of the National Digital Ecosystem for Culture. Its development is coordinated by the Central Institute for the Digitization of Cultural Heritage (Digital Library) of the Ministry of Culture, as

¹⁶ <https://www.oclc.org/en/news/announcements/2023/leveraging-machine-learning-for-worldcat-de-duplication.html>; <https://hangingtogether.org/machine-learning-and-worldcat-improving-records-for-cataloging-and-discovery/>.

¹⁷ Among the many works published in recent years, relevant to the area of interest and context to which this work refers (with no claim to exhaustiveness), the following contributions can be mentioned: (Lana 2023; Ridi 2023; Morriello 2023; Roncaglia 2023b) in *Biblioteche oggi Trends*, 9 1 (2023), 1, *L'intelligenza artificiale per le biblioteche*, an issue entirely dedicated to the theme.

¹⁸ TLIT4U – Improving Transliteracy Skills through Serious Games.

¹⁹ The survey and the detailed results for Italy are described in (Tammaro, Zanichelli, and Del Carlo 2024).

²⁰ For an introduction to the I.PaC project, see the page available at: https://ipac.cultura.gov.it/#cose_ipac.

part of the investments in culture provided by the PNRR (Italy's National Recovery and Resilience Plan).²¹ I.PaC is one of the components of the ecosystem and is intended to enable cooperation among all the other environments that make up the ecosystem itself (Bartoli et al. 2024, 17–18).²² Developed as an infrastructure for the integrated management of Italy's cultural heritage, I.PaC is configured not only to archive and preserve cultural resources but also as a dynamic environment for data enrichment aimed at sharing, using, and reusing of knowledge. As the central core of the national digital ecosystem, I.PaC offers a catalog of services organized into four functional macro-areas (24 ff). The first is dedicated to the ingestion of descriptive data and digital resources (assets); the second is designed to manage and process digital assets; the third and fourth macro-areas are dedicated respectively to Domain Knowledge Graphs, i.e.

“semantic representations that describe the relationships between cultural heritage entities within specific domains of heritage, such as Archival, Bibliographic, ABAP and Multimedia” (25).

and to Cross-Domain Knowledge Graphs that

“facilitate the representation of semantic relationships between digital entities belonging to different heritage domains” (26).

Domain Knowledge Graphs therefore aim to model highly specialized domains of knowledge and to capture the specific semantics of relevant entities and relationships, whereas the modeling of the Cross-Domain Graph²³ serves to overcome the descriptive specificities of disciplinary domains in order to foster the integration of information from different sources, to allow the reconstruction of contexts within which cultural heritage objects acquire meaning, and to connect museum objects, bibliographic resources, and archival documents (Bartoli et al. 2024, 32). To ensure cross-domain interconnection and navigation, data are also normalized and enriched through procedures that leverage AI techniques. Domain and Cross-Domain Graphs are fed by descriptive and cataloging metadata contributed by cooperating systems.

“The informational objects sent to the infrastructure will undergo a series of processing steps in order to obtain enriched ‘derivatives’. In other words, through processes of de-duplication, clustering, and

²¹ For a detailed reconstruction of the historical and project context of the national digital ecosystem for culture, see pages 11–15 of (Bartoli et al. 2024).

²² The ecosystem is composed of 4 environments: D.PaC (Digitization of cultural heritage), I.PaC (Infrastructure and digital services for cultural heritage), D.PaaS (Data product as a service), and the Access and Co-creation Platform.

²³ A *knowledge graph* is a structured representation of information that models data, concepts, entities, and relationships among them, representing the entities as nodes and the logical or semantic relationships between the nodes as edges. Below are the definitions of *Domain Knowledge Graph* and *Cross-Domain Knowledge Graph* in the I.PaC context, taken from the I.PaC Glossary. A *domain knowledge graph* is “obtained from a local knowledge graph enriched through rules. If no optimization and normalization is performed in I.PaC, it may coincide with the local knowledge graph. The domain knowledge graph conforms to the schema described in the target descriptive record format, of which it represents the extensional part.” A *Cross-Domain Knowledge Graph* is “a graph of semantic relationships between entities (cultural heritage objects, authority records, controlled vocabularies) belonging to different domains of knowledge and descriptions...”. The I.PaC Glossary is available online at: https://ipac.cultura.gov.it/wp-content/uploads/2024/08/Allegato-4-Glossario_I.PaC_202408.pdf.

mesh-up based on rules and AI algorithms, I.PaC extends the dynamic and relational character of cultural information heritage” (Bartoli et al. 2024, 33).

At this point in the process—after having been subjected to the sophisticated “enhancement” elaborations described above—the data from participants in the national data infrastructure indeed contribute to the formation and growth of the I.PaC knowledge base.²⁴ At the center of the national digital ecosystem for culture, and as its technological engine, lies a complex information network fed by the local databases of participating systems. Within this network, the data are made interoperable through a model that standardizes their representational structures by identifying, via a process of semantic modeling, the key entities (concepts, objects) and defining meaningful connections between them. In other words, the original data formats or schemas,²⁵ which differ because they belong to different contexts or domains, are “converted” into a uniform graph structure. The representational paradigm used also makes the data available for interconnection with other entities present in the broader context of the semantic web.

Within this architectural framework, I.PaC applies AI techniques to the structured data in order to maximize interconnection, intra- and inter-domain navigation, and enrichment.²⁶ I.PaC applies AI in three areas: digital objects, navigation and user interaction,²⁷ and descriptive metadata (Bartoli et al. 2024, 34–37). In the area related to descriptive metadata, AI is used to perform clustering, reconciliation, data cleanup, and enrichment operations necessary to improve a data set that is quantitatively significant yet qualitatively heterogeneous, since it is fed by heterogeneous sources. For the purposes of this paper, attention will be focused only on the clustering algorithm because, as mentioned above, it is within the clustering process that algorithms designed to avoid data duplication are employed.²⁸

SBN Sommerso: Machine Learning Applied to Bibliographic Data²⁹

In the I.PaC context, machine learning algorithms support the production of clusters of duplicate entities, aimed at creating *Super Entities* useful for cross-domain data navigation. These algorithms are employed in the SBN Sommerso project to automate the process of identifying identical, sim-

²⁴ The I.PaC knowledge base is also fed through the contribution of SBN Index data, an information asset managed by ICCU.

²⁵ In the case of data exchanged within SBN, this refers to the UNIMARC record format.

²⁶ There is a strong temptation at this point to invoke the vivid and now inescapable image of *the architect and the oracle* used by Roncaglia to describe, respectively, the models known and practiced so far for implementing knowledge organization systems—linked to an encyclopedic structuring of knowledge—and the emerging statistical/probabilistic paradigms typical of how generative artificial intelligences operate, characterizing the scenario just described as one in which they might coexist. The reference here is to (Roncaglia 2023a).

²⁷ Through the development of chatbots deployed on portals for cultural heritage access and use.

²⁸ “To avoid such duplications, AI algorithms were developed that are capable of interpreting the context of entities by analyzing elements such as dates and places, qualifiers and other associated historical information, in order to group entities that are distinct only nominally but are semantically identical.” (Bartoli et al. 2024, 36).

²⁹ SBN Sommerso is an ongoing project; in writing this section the author made use of technical project descriptions shared by the working group during periodic meetings, which have in part been included in the technical documentation. The author gratefully acknowledges those members of the working group who provided support and clarifications regarding the more technical project aspects utilized in this section.

ilar, or new entities and generating the respective lists, including the percentage of semantic similarity detected in lists of similar entities.

From a technical perspective, the processing of UNIMARC records takes place in an architectural context structured into several components, each dedicated to managing different functionality corresponding to various phases of the process. The main phases can be summarized as: data acquisition of the records to be compared, metadata extraction from the UNIMARC records, application of deduplication and clustering algorithms, and finally presentation of results. These steps are outlined below. At the operational level, two distinct procedures have been set up for the acquisition of the bibliographic data to be compared. The system provides, on the one hand, for the upload and loading of SBN Index records via data acquisition modules made available by the I.PaC platform, and on the other hand a workflow dedicated to uploading packages of data in UNIMARC format external to the Index database—essentially those corresponding to SBN Sommerso. Once loaded, the UNIMARC data from the Sommerso are subjected to a formal check through a pre-validation process aimed at ensuring that the records meet integrity and completeness requirements, containing all data designated as mandatory. The system also performs data correctness checks on certain key fields and provides feedback on the results of the pre-validation process, which operators can use to monitor and analyze errors and to make informed choices about how to handle any records flagged as problematic. Once loaded into the system, the UNIMARC records have rules applied to them for extracting the metadata needed for subsequent clustering. For each entity³⁰ involved in the analysis and comparison process of the “sommerso” data, specific rules have been defined—for example, for identifying the key information to extract and for assigning a weight to a data element associated with the reference entity. The extracted metadata, structured appropriately to facilitate record comparison, are then used—depending on the case—either by the machine learning procedures or by the rule-based algorithms to form groups based on semantic or logical similarity. The system in fact applies both rule-based algorithms and a probabilistic approach involving AI semantic processing³¹ to achieve clustering (i.e. grouping based on similarity criteria) as well as deduplication of records.³² Finally, the system will provide users with feedback on the outcomes of the processing. The main result presented to users will consist of three lists: one for the records judged to be identical (i.e., those for which the algorithms found an exact match between the pairs examined); the list of similar records, which have partial matches and are accompanied by the percentage of similarity computed by the algorithms to assist the experts in the final decision;³³ and the list of new records, i.e. those with no correspondence in the Index database and therefore need to be integrated.

³⁰ The entities (or nodes, in the I.PaC context) considered in the SBN Sommerso project are Author (Person, Corporate Body, Conference), Series, Work, Manifestation, Printer's Mark, and Subject, defined in the I.PaC Graph relating to the bibliographic domain.

³¹ Semantic similarity is defined as the distance between two points in a vector database. The closer the points are, the more likely it is (in the case of the SBN Sommerso project) that they represent the same thing.

³² Clustering and deduplication are two distinct processes. In brief, one can say that clustering groups similar records based on measures of semantic/statistical distance, whereas deduplication is a specific action that identifies exact or nearly exact duplicates in order to eliminate or consolidate redundancies in the data.

³³ The percentage of semantic similarity returned to the operators helps to assist the experts in the final decision, as final control over the evaluations presented by the system is in any case entrusted to them.

Conclusions

The SBN Sommerso project described here is conceived as a service both for the institutions that wish to integrate their collections—since it is intended to facilitate their entry into the Italian bibliographic infrastructure and to optimize data loading processes—and for the general public, understood as all those who rely on SBN catalog data as a tool for accessing recorded knowledge and for pursuing their various research objectives. It should be noted, however, that the project activities also serve as an experiment useful for the application of artificial intelligence techniques to bibliographic data, now implemented through machine learning. In this respect, the project represents a contribution in the context of an approach that heeds the calls of those in the library community who deem the participation of librarians in AI experiments to be fundamentally important, in order to actively take part in the process of constructing the emerging AI technologies under development in recent years. Being present and actively participating can be interpreted both in the specific context of SBN Sommerso—as acting as the “human in the loop”³⁴ by integrating one’s own judgments and evaluations with those made by the AI system, to improve it in a highly specialized context—and as leveraging one’s own expertise and values in the concrete shaping of technologies that serve the collective well-being of humanity and beyond.³⁵

³⁴ “The concept of human in the loop—the human in the process—dates back to the dawn of computing and automation. It refers to the importance of incorporating human judgment and experience into the operations of complex systems (the ‘loop’ being automated). Today the term describes the training of AIs that incorporates human judgment. In the future, it may become more difficult to remain within the AI’s decision-making process.” (Mollick 2025, 52–53). On the theme of the human–machine relationship, I would refer to the stimulating reflections offered by Valentina Tanni in her work (Tanni 2025) where she provides, in the opinion of this author, a contribution with a novel vision and breadth in the current debate on the topic: “The holistic approach of cybernetics, which pushed artists and engineers to establish conversations with the machine, is particularly interesting today, at a historical moment in which the discourse on artificial intelligence is instead focused exclusively on the machine’s learning and its supposed capacity to think, understand, create. On the one hand there is the computer, inert matter ready to be activated; on the other is the human creator, grappling with the task of making it become intelligent. Dialogue—a process through which it is possible to learn new, indispensable forms of coexistence—has been pushed into the background, supplanted by an ever more widespread rhetoric of substitution.” (Tanni 2025, 11–12). And further: “To construct an alternative to the antiquated myth of the machine as an antagonistic subject, it is necessary to recover the perspective of relationship, to explore dialogue, to imagine and test possible forms of coexistence . . . It is only through the practice of conversation, accompanied by an in-depth study of the political and economic dynamics within which machines come into being, that we can regain some form of agency within the frenzy of technodeterminism.” (Tanni 2025, 18).

³⁵ I would like to evoke here, in closing, the insights offered in (Bridle 2022) by quoting a short passage: “In this book I will also argue for the capacity for intervention and individuality of technology itself—or perhaps of technology yet to come: the moment, prophesied by many, when machines will become self-sufficient, self-aware, perhaps autonomous. Such a moment does not relieve us humans of responsibility or the ability to change our attitudes and behaviors. On the contrary, considering the agency of technology is an opportunity to have a serious and concrete reflection on how to ensure greater justice and equality to all the inhabitants of the planet: human, non-human, and machines. Technology remains, for now, chiefly in our hands, and we are able to repair, restore, and regenerate its engagement with the world and its effects on it.” (Bridle 2022, 27).

References

- Bartoli, Margherita, Luigi Cerullo, Lina Antonietta Coppola, Costantino Landino, Antonio Davide Madonna, Antonella Negri, Giovanni Pescarmona, Chiara Fauda Pichet, Margherita Porena, and Valentina Rossetti. 2024. "Verso la creazione di un Ecosistema digitale nazionale per la Cultura." *DigItalia* 19 (2): 11–48. <https://doi.org/10.36181/digitalia-00101>.
- Balnaves, Edmund, Leda Bultrini, Andrew Cox, and Raymond Uzwyshyn. 2025. *New Horizons in Artificial Intelligence in Libraries*. Berlin–Boston: De Gruyter Saur. <https://doi.org/10.1515/9783111336435>.
- Bridle, James. 2022. *Modi di essere. Animali, pianeta e computer: al di là dell'intelligenza umana*. Milano: Rizzoli.
- Brzustowicz, Richard. 2023. "From ChatGPT to CatGPT: The Implications of Artificial Intelligence on Library Cataloging." *Information Technology and Libraries* 42 (3). <https://doi.org/10.5860/ital.v42i3.16295>.
- Chen, Suzhen, and Mingyan Li. 2024. "AI for Cataloging and Metadata Creation: Perspectives and Future Opportunities from Cataloging and Metadata Professionals." *Technical Services Quarterly* 41 (4): 317–32. <https://doi.org/10.1080/07317131.2024.2394919>.
- Chow, Eric H. C., T. J. Kao, and Xiaoli Li. 2024. "An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations." *Cataloging & Classification Quarterly* 62 (5): 574–88. <https://doi.org/10.1080/01639374.2024.2394516>.
- Corrado, Edward M. 2021. "Artificial Intelligence: The Possibilities for Metadata Creation." *Technical Services Quarterly* 38 (4): 395–405. <https://doi.org/10.1080/07317131.2021.1973797>.
- Cox, Andrew M., and Suvodeep Mazumdar. 2022. "Defining Artificial Intelligence for Librarians." *Journal of Librarianship and Information Science* 56 (2). <https://doi.org/10.1177/09610006221142029>.
- Cox, Andrew M., and Maria De Brasdefer. 2025. *IFLA Entry Point to Libraries and AI*. IFLA. <https://repository.ifla.org/handle/20.500.14598/4034>.
- IFLA FAIFE (IFLA Committee on Freedom of Access to Information and Freedom of Expression). 2020. *IFLA Statement on Libraries and Artificial Intelligence*. IFLA. <https://repository.ifla.org/handle/20.500.14598/1646>.
- Lana, Maurizio. 2023. "Leggere l'IFLA Statement on Libraries and Artificial Intelligence al tempo di ChatGPT." *Biblioteche oggi Trends* 9 (1): 4–12.
- Manzoni, Laura. 2023. "L'indicizzazione semantica in automatico e semiautomatico nelle principali biblioteche europee." *Biblioteche oggi Trends* 9 (1): 57–64.
- Mollick, Ethan. 2025. *L'intelligenza condivisa. Vivere e lavorare insieme all'AI*. Roma: Luiss University Press.
- Morriello, Rossana. 2023. "Dati e metadati bibliotecari per l'intelligenza artificiale: per un'agency delle biblioteche." *Biblioteche oggi Trends* 9 (1): 38–45.

PCC (Program for Cooperative Cataloging). 2024. “Final Report: PCC Task Group on Strategic Planning for AI and Machine Learning.” PCC. <https://www.loc.gov/aba/pcc/taskgroup/TG-Strategic-Planning-AI-final-report.pdf>.

Ridi, Riccardo. 2023. “Intelligenza artificiale e web semantico: nessi reciproci, ambiguità e definizioni.” *Biblioteche oggi Trends* 9 (1): 27–37.

Roncaglia, Gino. 2023a. *L'architetto e l'oracolo: forme digitali del sapere da Wikipedia a ChatGPT*. Bari–Roma: Laterza.

Roncaglia, Gino. 2023b. “Intelligenze artificiali generative e mediazione informativa: una introduzione.” *Biblioteche oggi Trends* 9 (1): 13–26.

Rossi, Francesca. 2024. *Intelligenza artificiale: come funziona e dove ci porta la tecnologia che sta trasformando il mondo*. Bari–Roma: Laterza.

Saccucci, Caroline, and Athena Salaba. 2021. “Introduction to Artificial Intelligence (AI) and Automated Processes for Subject Access.” *Cataloging & Classification Quarterly* 59 (8): 699–701. <https://doi.org/10.1080/01639374.2021.2022058>.

Suominen, Osma, Juho Inkinen, and Mona Lehtinen. 2022. “Annif and Finto AI: Developing and Implementing Automated Subject Indexing.” *JLIS.it* 13 (1): 265–82. <https://doi.org/10.4403/jlis.it-12740>.

Tammaro, Anna Maria, Francesco Zanichelli, and Susanna Del Carlo. 2024. “L’IA in biblioteca: quale impatto? I risultati di un’indagine in Italia.” *Biblioteche Oggi* 42 (8): 3–14. <https://doi.org/10.3302/0392-8586-202408-003-1>.

Tanni, Valentina. 2025. *Conversazioni con la macchina. Il dialogo dell’arte con le intelligenze artificiali*. Roma: Tlon.