I.PaC: the National Data Space for Cultural Heritage

Margherita Porena[1,2,*], Antonella Negri[1] and Luigi Cerullo[1]

[1]Istituto centrale per la digitalizzazione del patrimonio culturale - Digital Library, Via di San Michele, 18, Rome, 00153, Italy
[2]Alma Mater Studiorum - Università di Bologna, Via Zamboni, 33, Bologna, 40126, Italy

Abstract

The article describes I.PaC (Infrastructure and Digital Services for Cultural Heritage), the digital framework designed as a central hub for managing descriptive data and digital objects from cultural institutions at a national level. The paper investigates the use of Artificial Intelligence (AI) within the I.PaC infrastructure to enhance the quality of descriptive data, to add value to digital objects, and to assist users in navigating cultural portals.

Keywords

Cultural Heritage, National data space, Generative AI

1. Introduzione

The great number of Italian cultural properties presents numerous challenges in terms of accessibility, conservation, and enhancement of cultural heritage. To address these challenges, a dedicated digital infrastructure for cultural heritage has been developed with the aim of:

- making cultural heritage accessible to a global audience, enabling the discovery of artworks, monuments, and historical documents from anywhere in the world and improving their accessibility and fruition;
- encouraging the digitalization of cultural properties, ensuring their preservation for future generations;
- promoting education and scientific research, by providing students and researchers with simplified access to valuable materials and information on cultural heritage, which might otherwise be difficult to obtain;
- acting as a catalyst for cultural tourism, stimulating the local economy, and further enhancing cultural heritage.

One of the core components of this ecosystem is *I.PaC - Infrastructure and Digital Services for Cultural Heritage* [1], which serves as a hub for the conservation, management, and enrichment of Italian digital cultural heritage. The platform strives to eliminate barriers to the access to cultural information and to solve issues related to the management of heterogeneous data in terms of format, category and domain.

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
margherita.porena@cultura.gov.it (M. Porena);
antonella.negri@cultura.gov.it (A. Negri);
luigi.cerullo@cultura.gov.it (L. Cerullo)

2. I.PaC: The National Data Space for Cultural Heritage

I.PaC - Infrastructure and Digital Services for Cultural Heritage [2] - is the data space dedicated to the preservation, management, and valorization of the Italian digital cultural heritage.

This digital space collects descriptive data and digital objects related to Italian cultural properties from archives, libraries, museums, and cultural sites across the country. The comprehensive repository ensures that valuable cultural artifacts and their associated metadata are preserved for future generations and made accessible to researchers, educators, and the general public.

The services provided by I.PaC are organized into four main areas: (1) digital assets management and processing: this area offers the necessary tools to preserve, process, and present digital objects linked to cultural heritage. It includes functionalities for the digitization, cataloging, and long-term storage of cultural assets, ensuring their integrity and accessibility over time; (2) domain and (3) cross-domain graphs: these services support the representation, querying, and retrieval of information about cultural entities and their semantic relationships. By constructing detailed graphs, I.PaC enables the recreation of the context and history of cultural objects, providing deeper insights and facilitating complex research queries that span multiple domains; (4) Teca multimediale: this user interface, offered as a Software-as-a-Service (SaaS), allows users to create, modify, search, and delete digital resources within I.PaC. It supports advanced searches, making it easier for users to find and interact with the cultural data they need. The Teca multimediale also integrates multimedia capabilities, enabling the seamless presentation of various digital formats.

For the first three areas, I.PaC is exploring the use of artificial intelligence models to improve, enrich, and extract data. These AI models are designed to enhance the accu-

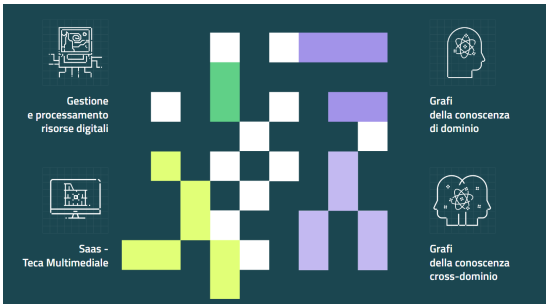

Figure 1: Four main I.PaC service areas

racy and depth of cultural data, promoting continuous evolution in the way cultural information is managed and shared [3]. By leveraging AI, I.PaC aims to facilitate more efficient data processing, uncover hidden connections between cultural entities, and provide users with richer, more contextualized information about Italy's cultural heritage.

3. AI applied to descriptive data

In the graphs area, one of the main problems is that the data managed by the I.PaC graph comes from various sources, which may assign different identifiers to otherwise identical entities. This can lead to an overabundance of entities that, in reality, refer to the same object. A typical example is "Agent" entities (like Leonardo Da Vinci), which is registered in multiple systems with different identifiers, creating in this way different entities.

To solve this problem, innovative AI algorithms have been employed. These algorithms intelligently analyze the context of each entity, taking into account important details like ates and places of birth, qualifications, and biographical information. By doing so, they can group entities that are nominally different but semantically identical, effectively reducing duplication.

In the context of agent reconciliation, AI faces a particularly challenging task due to the often limited descriptive data available. Frequently, the only information provided is the agent's full name, with no chronological references or additional identifying details. In such cases, the AI must employ advanced techniques to analyze the works associated with the agents. For artworks, the AI can attempt to identify stylistic similarities by examining features such as brushwork, technique, and recurring motifs. For bibliographic works, it can focus on similarities related to the subject matter, comparing themes related to the work. These methods enable the AI to suggest potential matches, overcoming the limitations imposed by the lack of detailed data.

Plans are in place to expand this approach to encompass other types of entities, such as events and literary works. This is essential for ensuring that different records referring to the same object are accurately reconciled, maintaining the integrity and efficiency of the I.PaC graph.

Currently, two additional AI applications within I.PaC are being tested:

- the development of models aimed at linking identified entities to controlled vocabularies and domain-specific terminological tools that describe cultural properties,
- the enrichment of the graph with information extracted from textual contexts.

For the first aspect, many cultural properties are described using unstructured texts that do not refer to standardized vocabularies or thesauri, making access to information less immediate. The project aims to create models that link these descriptions to standard categories from controlled vocabularies, despite the challenge posed by the highly specialized and domain-specific nature of such terminologies.

For the second aspect, the team is working on AI models that extract data from unstructured texts to integrate it into the I.PaC data model in a structured form, thus simplifying the search process and increasing the informative value of the graph.

4. AI applied to digital objects

One of the primary goals of I.PaC is to manage a great number of digital objects that come from various cultural institutions and organizations across the country.

Among the various functionalities offered, I.PaC is experimenting with a content processing system using a range of artificial intelligence techniques, from Machine Learning to generative models. This initiative aims to achieve two primary objectives: on one hand, to generate new digital content or media; on the other, to extract meaningful information from existing content.

In this initial phase, 7 specific use cases have been selected to test how AI can enhance digital resources and enrich the graph. These use cases are:

- (1) Text extraction from ancient and modern monographs: this involves extracting text frome the digitization of the monograph, creating an abstract, extracting named entities, identifying the subject, determining the table of contents, and identifying the physical structure of the resource, ensuring that images are arranged according to the correct pagination or foliation indicated in the resource. The challenge in this case lies in analyzing ancient monographs, which often have

particularly deteriorated text, instances of bleed-through, and highly complex layouts where text is arranged on the page in various shapes.

- (2) Processing of digitized journals: unlike the previous case, this task involves identifying the articles within a periodical, associating each article with its corresponding text section, title and subtitle, and author. The challenge here is the vast variety of layouts that need to be recognized. Additional difficulties include identifying sections that are physically separate but logically part of the same article, handling articles that continue on different, often distant, pages of the resource, and dealing with advertisements that can physically and logically separate various parts of the same article.
- (3) Audio and video elaboration. In this context, the AI must be capable of extracting text from audio files that may be corrupted. For each resource, it will need to generate an abstract: if the resource is musical, the abstract should consider only the descriptive metadata; for spoken resources, the abstract should be based on the content of the extracted text.
- (4) Image processing: Within the I.PaC ecosystem, millions of images related to cultural heritage will be hosted. This use case aims to process these images to identify the main subject and the entities they comprise, mapping this information to nationally recognized controlled vocabularies (such as the Thesaurus del Nuovo Soggettario di Firenze or the Iconclass classification). Each recognized entity must be associated with the coordinates of the section of the resource where it is located, making it easily representable in a IIIF manifest, [4][1]. The goal is to create a description of the image that can also be reproduced via audio files (to improve information accessibility) and to identify similar images, including a similarity score for each recognized similar image.

[1]The International Image Interoperability Framework (IIIF) is a standard developed to facilitate the access and sharing of digital images by libraries, archives, museums, and other institutions with image collections. IIIF enables interoperability between different platforms and viewing systems, allowing users to access, view, and annotate high-resolution images uniformly and consistently, regardless of their origin. A manifest in this context is a JSON document that provides detailed information about a digital resource, such as an image or a collection of images. The manifest contains metadata that define various aspects of the resource, such as bibliographic information, structure (e.g., pages of a manuscript), and coordinates for annotating specific sections of the image. Through the manifest, IIIF-compatible applications can present and manage images in a standardized way, supporting advanced functionalities like zooming, magnification, page navigation, and collaborative annotations.

Figure 2: Example of metadata extraction from maps

- (5) Metadata extraction from maps: this use case involves developing technologies capable of extracting data from digitized maps, such as place names, the scale used, and any symbols marked on the map along with their legend associations. Specifically, for cadastral maps, the AI must also recognize the cadastral parcels indicated in the image. The challenge in this task is that many ancient and modern maps have handwritten data, making it difficult to recognize different types of handwriting. Additionally, there is often no extended textual context available that could help the AI correct extraction errors using semantic context.
- (6) Extraction of musical notation from digitized sheet music: This use case focuses on developing technologies capable of extracting musical notation from digitized sheet music and enabling the playback of the extracted notation. The AI must accurately recognize and interpret various musical symbols, notes, and annotations present in the sheet music. This involves dealing with challenges such as varying quality of digitized images, handwritten annotations, and different musical notation styles. The goal is to create a digital representation of the music that can be easily read, edited, and played back, preserving the integrity and accuracy of the original sheet music.
- (7) Extraction of information from catalog records: Over time, numerous paper catalog records have been created to describe cultural heritage items, representing a valuable informational resource that needs to be recovered. In many cases, the only information available about certain cultural heritage items is contained in these paper records. This use case involves extracting information from these digitized catalog records to map the extracted metadata to the current national information representation models. The challenge here lies in the significant variation in the layouts used

in these catalogue records and the differing information each type of records requires. It is not possible to identify a specific layout or consistently recurring data (except for some basic information, such as the catalog number or the classification of the item). Therefore, the technology must be capable of extracting the information, recognizing its semantics, and mapping it to the relevant descriptive data model.

Technologies for the last three use cases have already been successfully tested, demonstrating the feasibility and effectiveness of the proposed solutions. However, in the coming months, these successfully tested technologies will require fine-tuning to improve performance and achieve increasingly precise results. For the other use cases, a proof of concept (PoC) is currently being carried out by two competing companies. Upon completion of this phase, the best results will be evaluated, and the most suitable solution will be selected. The final choice will consider both the technologies used and the developed pipeline, which must be capable of processing the resource in an automatic manner, ensuring all required outputs. Human intervention will only be necessary for result validation, thus ensuring an efficient and scalable process for managing cultural heritage resources.

5. Generative AI to enhance information retrieval

I.PaC provides also services to enhance information retrieval in the form of chatbots that use generative artificial intelligence to assist users in navigating portals dedicated to cultural heritage.

The first project to have been realized, still in public experimentation, is *Alphy*, designed with the goal of assisting users in navigating and accessing information in *Alphabetica* [5], the portal of Italian libraries created by the Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane (ICCU). The application of generative artificial intelligence is crucial in three key phases of the interaction process between the chatbot and the user:

- user intent interpretation: during this phase, the AI analyzes the user's input to accurately identify their intentions;
- mapping intentions to three search templates: in this phase, the system guides the user's intentions towards three key templates: Works, Protagonists, and Themes;
- analysis and enrichment of results: in the third phase, the chatbot reviews the results obtained from the Alphabetica indexes, enriching the response with additional information from its knowledge base, making the user experience more informative and engaging. All information generated by the AI is highlighted in the chat, ensuring compliance with current regulations.

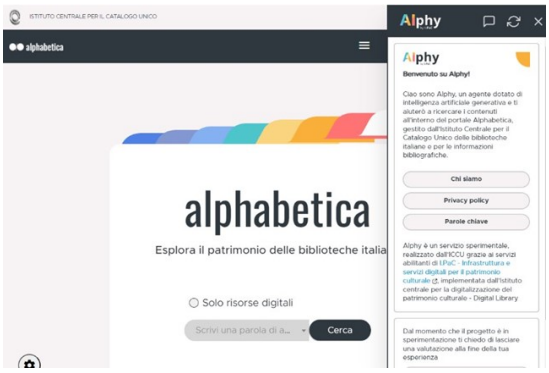

Figure 3: Alphy, the AI-powered generative chatbot for the Alphabetica portal navigation

Currently, another chatbot is being developed for navigating the *General Catalog of Cultural Heritage* [6], which contains data on cultural properties from Italian museum and other cultural institutions. Unlike the first case, this experiment aims to use generative AI to process RDF data, organized according to the ArCo ontology network [7] and accessible through SPARQL queries. The goal is to convert natural language questions into SPARQL queries, thus facilitating access to information. In this context, the generative AI must use Retrieval-Augmented Generation (RAG) [8] because it needs to comprehend the semantics of the ontology and suggest research paths. This approach allows the AI to provide more accurate and contextually relevant responses by dynamically integrating and retrieving pertinent information from the knowledge graph, thereby enhancing the overall user experience in accessing and exploring the vast cultural heritage data.

6. Conclusions

In conclusion, the development and implementation of the I.PaC - Infrastructure and Digital Services for Cultural Heritage - represents an important advancement in the management and valorization of Italian cultural heritage. Through leveraging cutting-edge artificial intelligence technologies, from machine learning to generative models, I.PaC not only aims to preserve and make accessible cultural properties but also to innovate the way these treasures are studied and known. The exploration into AI-driven enhancements, including descriptive data analysis and digital object processing, can bridge the gap between

historical legacy and modern accessibility. The introduction of AI-powered chatbots like Alphy for navigating cultural portals points out the commitment to enhancing user experience and information retrieval. Thanks to the continuous refinement of AI applications and extension of digital services, I.PaC is a powerful example of how culture, technology, and education come together, ensuring that cultural heritage is not only preserved but made accessible in new ways for generations to come.

References

[1] I.pac - infrastruttura e servizi digitali per il patrimonio culturale, 2024. URL: https://ipac.cultura.gov.it/.

[2] L. Cerullo, A. Negri, L'infrastruttura software per il patrimonio culturale (ispc) come abilitatore di un ecosistema digitale nazionale del patrimonio culturale, DigItalia 18 (2023).

[3] R. Parry, Recoding the Museum: Digital Heritage and the Technologies of Change, Routledge, 2007.

[4] R. S. Stuart Snydman, T. Cramer, The international image interoperability framework (iiif): A community technology approach for web-based images, Archiving conference 12 (2015).

[5] Alphabetica, 2021. URL: https://alphabetica.it/.

[6] Catalogo generale dei beni culturali, 2021. URL: https://catalogo.beniculturali.it/.

[7] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, Arco: The italian cultural heritage knowledge graph, in: Proc of ISWC, 2019, pp. 36–52.

[8] M. D. D. M. Hamed Zamani, Fernando Diaz, M. Bendersky, Retrieval-enhanced machine learning, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (Madrid, Spain) (SIGIR '22), Association for Computing Machinery, New York, NY, USA, 2022, p. 2875–2886. doi:https://doi.org/10.1145/3477495.3531722.