

I.PaC and semantic graphs to represent Italian Cultural Heritage

Margherita Porena^{1,2}, Margherita Bartoli¹, Luigi Cerullo¹, Antonella Negri¹

¹Istituto centrale per la digitalizzazione del patrimonio culturale - Digital Library, Italy

²Alma Mater Studiorum - Università di Bologna, Italy

Abstract

I.PaC (Infrastructure and Services for Cultural Heritage) [CN23] is the technological core of the National Digital Ecosystem for Culture - Ecomic, an initiative developed within Italy's National Recovery and Resilience Plan (PNRR). It is designed to support the management, enrichment, and accessibility of digital cultural resources, and it is developed not merely as a data repository but as an advanced infrastructure that enables new models of interaction and valorization of cultural heritage.

I.PaC represents the central hub of the ecosystem and it integrates and connects heterogeneous systems, allowing cultural institutions to ingest, manage, and reuse digital data through a comprehensive range of services. The infrastructure also provides services for digital asset management (DAM) and advanced content processing, allowing institutions to optimize digitization processes and improving the quality of their metadata.

A key feature of I.PaC is the use of domain-specific and cross-domain graphs: these graphs allow to create relationships between cultural objects from different fields (archival, bibliographic, museum, and multimedia), enabling cross-domain navigation and a richer reconstruction of cultural contexts.

On one hand the domain-specific graphs serve as vertical, in-depth models that structure and organize all the relevant information within each cultural sector. They ensure a high level of specialization and interoperability, since they are based on national and international description standard, and they enhance metadata thanks to AI technologies based on entity recognition, disambiguation, and data enrichment. Each domain graph retains its internal logic while benefiting from a shared semantic foundation.

On the other hand, the cross-domain knowledge graph provides a transversal representation of information about cultural heritage, by modeling common entities and their relationships, which can then be reused across all domain-specific graphs. This approach enables the integration of information from traditionally separate disciplines, revealing unexpected connections. By implementing I.PaC's semantic graphs, Italy's cultural institutions can transition from static data repositories to dynamic, knowledge-driven platforms, creating new opportunities for accessibility and valorization.

1. Introduction

The Italian cultural heritage represents one of the largest collections of historical, artistic, and documentary objects in the world. It is composed of cultural properties from various disciplinary domains— bibliographic, archival, and museal—, characterized by their own description practices, cataloguing standards, and management methods. While these differences are important, because they represent the specificity of each domain, they also present a significant challenge: the lack of native semantic interoperability among data limits the ability to integrate, correlate, and contextualize information related to cultural heritage from different sectors.

For example, we might have a work of art, a document attesting to its history and creation, and a publication describing it: although all three refer to the same cultural context, they are described and accessed separately, stored in distinct databases. This fragmentation of information about cultural heritage reduces opportunities for both scientific research and public access.

To address this challenge, I.PaC (Infrastructure and Services for Cultural Heritage) was created as a digital infrastructure conceived within the framework of the National Recovery and Resilience Plan. From this perspective, I.PaC can be seen as an enabling tool for the semantic connection of cultural data, promoting the cross-domain integration of descriptions and the overcoming of traditional domain boundaries.

By adopting knowledge models based on semantic graphs and providing advanced services (including those powered by Artificial Intelligence) for data management, cleansing, and enrichment, I.PaC enables the reconstruction of the cultural contexts of heritage items. In this way, each cultural property is no longer considered an isolated entity, but is positioned within a complex network of historical, geographical, conceptual, and social relationships—offering new perspectives for enhancement and accessibility for researchers, institutions, and the public.

This article presents the I.PaC infrastructure through its key com-

ponents, beginning with a general overview of the system and its objectives. It then provides an in-depth analysis of the adopted semantic model, focused on the use of knowledge graphs and their modeling through UML. The article describes the main graphs that form the conceptual base of the ecosystem: the cross-domain graph, which serves as a transversal integration layer, and the domain-specific graphs (multimedia, archival, bibliographic and museal), each modeled according to the characteristics and standards of its respective disciplinary field. Each section highlights the conceptual, methodological, and practical aspects of the modeling, with the aim of showing how I.PaC transforms fragmented cultural data into a rich, contextualized, and interoperable information.

The last section explores how artificial intelligence techniques are applied to the graph-modeled data. These AI-driven services — including reconciliation, cleansing, clustering, and enrichment — leverage the semantic structure of the data to enhance discoverability, interoperability, and user experience, while fully preserving the integrity of original cataloging inputs. This final section illustrates the innovative potential of combining AI with semantic knowledge graphs to optimize the accessibility and reuse of cultural heritage information.

2. I.PaC infrastructure

Within the framework of national policies for the digital transformation of culture, I.PaC – Infrastructure and Services for Cultural Heritage represents the technological core of the Digital Ecosystem for Culture (ECoMic) [CBC*24], developed within the scope of the PNRR sub-investment MIC3 1.1.4. It is configured as the first national data space dedicated to culture, designed to securely and persistently host the entire digital heritage of the country, including both digital resources and descriptive metadata. I.PaC is not merely a centralized archive, but an open semantic platform that integrates advanced services for the management, enrichment, and interoperability of cultural heritage data.

The infrastructure enables cooperation between heterogeneous information systems through standardized interoperability mechanisms, facilitating the reuse and enhancement of content within distributed digital environments. Thanks to a technological setup that includes semantic engines, Digital Asset Management (DAM) services, processing tools based on Artificial Intelligence, and knowledge graphs, I.PaC makes it possible to transform isolated digital collections into connected and queryable information heritage. In this context, semantic relationships become structural elements of representation, enabling the discovery of new connections and the emergence of cross-cutting meanings.

The design of I.PaC aligns with the principles of the National Digitization Plan (PND), promoting an open, modular, and standards-compliant data management approach capable of supporting all phases of the data lifecycle: ingestion, organization, description, enrichment, access, and reuse. The infrastructure provides services aimed both at front-end systems, dedicated to the public use of content, and at back-end systems, dedicated to the production and management of metadata.

One of the distinctive elements of I.PaC is the adoption of

domain-specific and cross-domain semantic graphs, which represent cultural entities and their relationships in a flexible and queryable structure. These graphs form the basis for advanced operations of analysis, semantic reconciliation, and automatic enrichment, enabling a deeper understanding of the heritage. In this sense, I.PaC not only centralizes and harmonizes digital content but transforms it into active knowledge, ready to be reused in scientific, educational, and communicative contexts.

3. Knowledge graphs

In the context of knowledge representation, the graph can be considered as one of the most flexible and scalable way to model entities and relationships, particularly to represent the cultural heritage, since it is characterized by complex data in which each cultural object gains meaning through a network of connections with other entities such as agents, events, places, etc. Formally, a graph consists of nodes (or vertices), which represent entities or concepts, and edges, which express the semantic relationships between those entities.

In I.PaC, the use of knowledge graphs makes it possible to overcome the limitations of traditional relational databases, offering a more natural and navigable representation of cultural information. In this model, relationships take on a primary role, fostering integration, semantic interoperability, and new ways of exploring digital heritage.

Five main graphs have been developed in I.PaC:

A cross-domain graph, designed to overcome the separation between disciplinary fields and to enable transversal navigation across cultural assets from different sectors, and four domain-specific graphs:

- the bibliographic graph, for the representation of books, manuscripts, and bibliographic materials;
- the archival graph, for the modeling of documents and archival collections;
- the museum graph, dedicated to artistic, archaeological, and historical objects from museums and heritage offices;
- the multimedia graph, focused on the description of digital resources (images, audio, video, complex content).

The domain graphs collect the domain-specific descriptive metadata of their respective cultural sectors, modeling highly specialized information in accordance with the relevant cataloging and descriptive standards.

The cross-domain graph integrates and connects these fields, offering a unified and enriched view of cultural heritage, which enables the reconstruction of the historical, social, and geographical contexts in which the assets acquire full meaning.

The complexity of the structuring lies in the integration of different standards and descriptive criteria (MARC, XML, RDF), and in the harmonization of heterogeneous conceptual models, in order to ensure a coherent and interoperable representation of information within the semantic web.

In the sections that follow, all the graphs developed within the I.PaC framework will be analyzed in detail, with the exception of

the museum graph. This latter, in fact, fully adopts the structure of the ArCo (Architecture of Knowledge) model [CGM*19], developed by the Central Institute for Cataloguing and Documentation (ICCD), the national body responsible for cultural heritage cataloging standards. ARCo already constitutes a native graph representation, based on RDF/OWL ontologies, and is fully compliant with Italian descriptive standards. For this reason, the museum graph did not require dedicated modeling within I.PaC, as it could be directly integrated into the system, ensuring both semantic coherence and normative alignment.

3.1. Graph modeling methodology

The creation of knowledge graphs in I.PaC followed a methodology structured into several phases:

the first phase concerned the identification of objectives (I): each graph within the I.PaC infrastructure was designed to meet specific goals, addressing different needs that cannot be unified under a single perspective. The cross-domain graph has as its primary objective the integration and semantic reconciliation of information from heterogeneous disciplinary fields, in order to build a unified and interconnected view of cultural heritage. This makes it possible, for example, to link museum objects, bibliographic resources, and archival documents referring to the same cultural context.

In contrast, the domain-specific graphs are oriented toward the accuracy and richness of scientific representation within their respective fields. In these cases, it is essential to preserve the conceptual distinction between similar but non-equivalent entities, avoiding any forced unification that would weaken their descriptive significance. The goal in developing these graphs is to maintain a modeling approach that adheres to disciplinary standards, enhancing the semantic and cataloging specificity of each domain.

The second phase (II) focused on the analysis of domain models: once the objectives for each graph were defined, an in-depth analysis was conducted of the conceptual models and languages used in the various sectors, taking into account the variety of standards (MARC, XML, RDF) and the different information structures. While the ABAP sector already had a graph-based model, the bibliographic, archival, and multimedia sectors required significant adaptation.

The process then moved on to the creation of the model (III): a unified conceptual model was defined, based on entities and relationships common to the various cultural domains, integrating specific entities from individual sectors where necessary in order to preserve descriptive richness and depth. The design of the graphs in I.PaC was carried out using the UML (Unified Modeling Language), adopted to formally define the system's entities, properties, and relationships. In particular, UML was used to:

- define classes corresponding to the main types of cultural entities (heritage items, agents, events, places, concepts, digital resources)
- specify the semantic relationships between classes, indicating their cardinality, direction, and navigability rules;
- describe the attributes associated with each entity, distinguishing between mandatory and optional;

- formalize specializations and compositions between classes, to accurately represent conceptual articulations and hierarchies.

Once the models were established, it was necessary to map the data onto these models (IV): in this phase, the information structures of the individual sectors were harmonized with the cross-domain model, ensuring that the information could be navigated consistently, regardless of the source domain.

Finally, the data were normalized and enriched (V): this final phase involved the normalization of heterogeneous information and subsequent semantic enrichment, based on artificial intelligence techniques such as entity reconciliation, clustering, and the creation of new implicit relationships.

3.2. The cross-domain graph

One of the distinguishing elements of the I.PaC infrastructure is the cross-domain graph, designed to overcome traditional boundaries between archives, libraries, museums, and other cultural sectors, and to coherently integrate heterogeneous data. Based on a flexible and scalable semantic representation, this graph enables transversal navigation across cultural objects, involved agents, places, events, and concepts, thus promoting contextualization and interoperability between different domains.

The graph is based on a shared semantic model, which represents a conceptual vocabulary common to the various domains (bibliographic, archival, museal, multimedia), thus forming the core of the ecosystem's interoperability.

At the cross-domain graph level, a generic entity (Entity) was first defined, from which all other entities inherit. This generic class provides shared attributes—such as labels, notes, or descriptions—and relationships such as those to names, identifiers, or temporal entities. The main types of entities modeled are described below.

3.2.1. Cultural Entity

At the center of the graph is the conceptual class *Cultural Entity*, representing any cultural object of interest, whether tangible or intangible, physical or conceptual. It is the broadest and most transversal category, capable of encompassing books, artworks, archival documents, artifacts, images, editions, collections, as well as digital resources or abstract entities related to culture. Each cultural entity is identified by a unique identifier and described through textual labels, descriptive notes, links to controlled vocabularies, and, when available, associated digital resources.

Cultural entities are designed to be highly relational: each one can be linked to responsible agents, reference places, time contexts, and concepts that describe or classify its content. This structure allows for the reconstruction of the semantic network in which the asset is embedded, facilitating not only access to content but also understanding of its meaning and cultural context.

3.2.2. Temporal Entity

The temporal dimension is modeled through the class *Temporal Entity*, which allows chronological information to be associated with

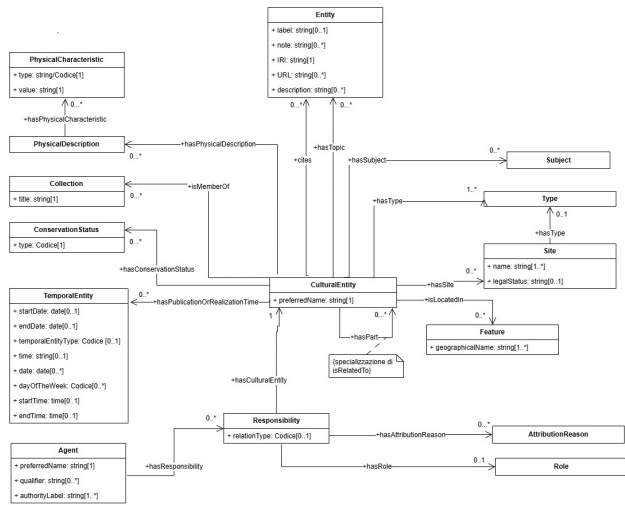


Figure 1: Cultural entity model.

any element of the graph. The structure is flexible enough to represent both precise dates (e.g., creation or publication dates) and complex or approximate time intervals (e.g., "first half of the 19th century", "between 1870 and 1880"). Each *Temporal Entity* can be described using textual labels, defined by start and end dates, and characterized by a certainty level, useful in cases where the information is incomplete or ambiguous.

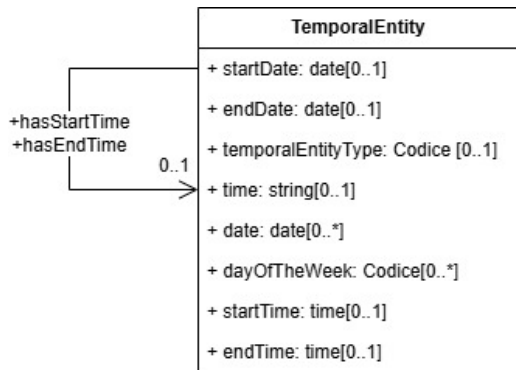


Figure 2: Temporal entity model.

To support this flexibility, the *Temporal Entity* class includes two recursive properties: *has Start Time* and *has End Time*. These relationships allow multiple temporal entities to be connected, for instance, to explicitly represent the endpoints of a time span. This enables specific metadata to be associated with just one endpoint — for example, indicating that a start date is certain while the end date is uncertain. In this case, two separate temporal entities can be created and linked to the main entity via *has Start Time* and *has End Time*, each with its own level of reliability or approximation.

The introduction of these autonomous temporal entities and their connecting relationships allows for the chronological linking of

cultural assets, events, responsibilities, and places, enabling time-based querying and supporting cross-temporal analysis. This is particularly useful for comparing historical trajectories, production phases, or exhibition cycles.

3.2.3. Responsibility

Another fundamental axis of the modeling concerns the subjects involved in the creation, preservation, or promotion of cultural assets, represented by the class *Agent*. This includes individuals, public or private organizations, historical families, research groups, and even software or information systems — whenever they play a documented role in the life of a cultural object.

The relationship between a *Cultural Entity* and an *Agent* is not direct but mediated by an intermediate class, *Responsibility*, which explicitly defines the type of relationship: author, publisher, curator, cataloger, conservator, etc. This approach documents not only who played a role, but also what the role was and when it occurred, making the attribution explicit and offering a model suitable for representing complex or overlapping information over time.

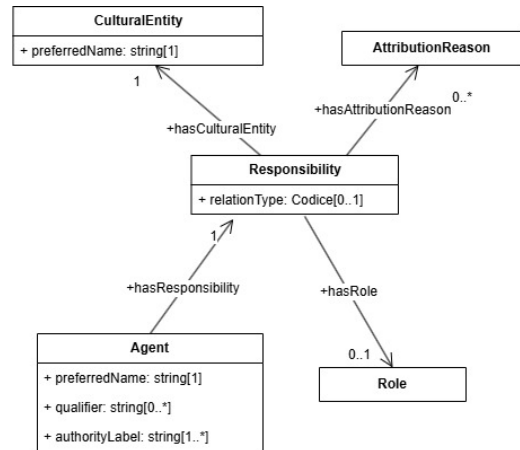


Figure 3: Responsibility model.

The modeling of responsibilities addresses both descriptive and analytical needs, enabling, for example, the study of a cultural property's history through the various actors who influenced its trajectory.

3.2.4. Feature

The spatial dimension is modeled through the class *Feature*, which describes places associated with cultural assets. This is fundamental for the territorial contextualization of digital content, especially in a national system like I.PaC, which aims to integrate contributions from across the country. The class is structured into several subtypes representing different levels of geographic detail: from specific addresses to zones, cities, provinces, regions, and states. It also supports the representation of specific cultural or natural sites, such as historic buildings, through the class *Site*.

Features can be georeferenced through coordinates, connected to other places via hierarchical relationships (e.g., an address within a

textual network that highlights intertextual relationships, transmissions, derivations, and rewritings.

The *WorkInstance* class, on the other hand, represents the concrete manifestations of the work: printed editions, manuscript copies, facsimiles, critical or digital editions. This class is intended to describe all the elements that define a publication within a specific editorial context, including the traditional ISBD areas (title, responsibility, edition, publication, physical description, notes), as well as information on version, paratextual elements, use of color, writing system, and internal structure. The model also establishes an association between *WorkInstance* and *Item*, the entity that describes the actual physical or digital copy held, including details about location, call number, extent, provenance, conservation status, acquisition method, and access profiles.

One of the most advanced aspects of the bibliographic graph is its handling of serial publications, managed through the classes *Series*, *SerialWork*, and *Issue*. *Series* aggregate publications tied by a common editorial line or series title; within them, *SerialWork* represents the conceptual dimension of seriality (e.g., a journal), while *Issue* models each published issue, describing its number, volume, publication date, periodicity, and links to previous or subsequent issues. Each *Issue* may contain multiple *WorkInstances*, which in turn are linked to their respective *Items*, allowing for highly granular content descriptions.

Particular attention is devoted to the representation of manuscripts, treated not merely as material variants but as bibliographic objects with specific historical, artistic, and codicological value. The *Manuscript* class, designed as a logical extension of *WorkInstance* and *Item*, enables a detailed description that includes information on incipit, explicit, support, book format, musical or liturgical notation, decorations and miniatures, scripts used, languages and alphabets, bindings, and restorations. This modeling allows the reconstruction of complex material and cultural trajectories, such as those of medieval, modern, and contemporary manuscripts, and enables direct linkage to institutions, historical catalogs, documentary sources, and producing or holding entities.

To strengthen the semantic layer of the graph, auxiliary classes such as *Classification* are included, enabling the assignment of subject categories to works and manifestations (e.g., according to Dewey, DDC, LCC, or local systems), and *Extent*, which describes the physical dimensions of resources (pages, volumes, plates)—useful for organization, digitization, and preservation.

All bibliographic entities can be linked to agents (e.g., people or institutions) via the *Responsibility* class, which specifies the role of the agent in relation to the work or publication: author, editor, translator, illustrator, publisher, printer. Furthermore, each resource can be linked to one or more *DigitalResources*, enriched by logical and physical structures (in the case of digitized or born-digital items), and optionally connected to *Concepts* or *Subjects* for semantic categorization.

3.6. Applying AI to Cultural Heritage Graphs: Reconciliation, Cleansing, Clustering, and Enrichment

Within the I.PaC infrastructure, one of the most innovative elements is the integration of AI-based services applied to graph data.

The data in I.PaC, semantically modeled as knowledge graphs, can undergo four main types of processing: reconciliation, cleansing, clustering, and enrichment. These processes are always carried out in full compliance with current regulations on artificial intelligence, ensuring transparency regarding the origin of any generated information. It is important to stress that no automatic process replaces or modifies the original data entered by the cataloguer, who is regarded as the primary and irreplaceable source. On the contrary, the goal is to enhance and structure these data to facilitate access, searchability, and reuse.

The reconciliation process involves mapping free-text strings to controlled vocabularies, thereby reducing lexical ambiguity caused by manual data entry. This is particularly useful in domains where standards prescribe the use of specific terminology while still allowing for human input. Typos, unrecognized synonyms, or uncatalogued variants can compromise data quality. For instance, a cataloguer may enter a term such as “anforetta miniaturistica” (miniature amphora), which is not found in any official vocabulary. AI can map this to an existing term (“anfora”) and retain “miniaturistica” as a qualifier—preserving meaningful information while still aligning with the controlled vocabulary.

By using LLM-based models, the system attempts to reconcile text strings with existing vocabulary terms. If reconciliation is not possible—because a new term has been introduced—the system flags it for review by the authority managing the reference vocabulary, enabling potential future updates.

Cleansing is the process to normalize heterogeneous textual data, such as dates, which are often expressed in different formats (e.g., roman vs arabic numerals, centuries with or without abbreviations, or time periods in literal form). In this case the AI process is useful to produce a more coherent dataset and to improve chronological searches.

Clustering involves the identification and merging of duplicate entities. For example, the same author—such as Michelangelo Merisi, known as Caravaggio—might be described differently across domains. In a domain the descriptive profile might be richer, (including pseudonyms, birth dates, artistic schools, etc.), while in another domain cataloguers use a more simplified model. The AI process allows these variations to be logically unified, while preserving the provenance of each record. This is crucial also for the enrichment process, which aggregates complementary metadata from multiple domains to create more complete and accessible representations. For instance, metadata from a system that does not record an author’s gender or school of affiliation can be enriched using information from other sources, thus enhancing the end-user experience.

Another area of enrichment involves the extraction of structured data from unstructured texts, such as descriptive notes or biographies. From these, the system derives entities and relationships that improve semantic navigation and querying

3.7. Conclusions

The I.PaC infrastructure represents a decisive step toward the digital transformation of Italian Cultural Heritage, offering an integrated and semantically rich view of archival, bibliographic,

museal, and multimedia resources. Through the use of knowledge graphs and advanced conceptual modeling, I.PaC overcomes the traditional fragmentation of data, promoting interoperability across domains and enhancing the value of cultural heritage within a unified and contextualized perspective. The ability to navigate information transversally, to reconstruct the historical, geographical, and cultural contexts of heritage assets, and to enable intelligent services for content access and reuse opens new opportunities for scientific research, public dissemination, and citizen engagement.

References

- [BMP24] BODO S., MASCHERONI S., PANIGADA M.: *Fare nuove le cose: Patrimonio culturale e narrazione, uno sguardo pluridisciplinare*. Mimesis Edizioni, 2024. URL: <https://books.google.it/books?id=uRoREQAAQBAJ>.
- [CBC*24] CERULLO L., BARTOLI M., COPPOLA L. A., LANDINO C., MADONNA A. D., NEGRI A., PESCARMONA G., FAUDA PICHET C., PORENA M., ROSSETTI V.: Verso la creazione di un ecosistema digitale nazionale per la cultura. *Digitalia* 19, 2 (dic. 2024), 11–48. URL: <https://digitalia.cultura.gov.it/article/view/3074>, doi:10.36181/digitalia-00101. 2
- [CGM*19] CARRIERO V. A., GANGEMI A., MANCINELLI M. L., MARINUCCI L., NUZZOLESE A. G., PRESUTTI V., VENINATA C.: Arco: The italian cultural heritage knowledge graph. In *The Semantic Web – ISWC 2019* (Cham, 2019), Ghidini C., Hartig O., Maleshkova M., Svátek V., Cruz I., Hogan A., Song J., Lefrançois M., Gandon F., (Eds.), Springer International Publishing, pp. 36–52. 3
- [CN23] CERULLO L., NEGRI A.: L'infrastruttura software per il patrimonio culturale (ispc) come abilitatore di un ecosistema digitale nazionale del patrimonio culturale. *Digitalia* 18, 1 (ago. 2023), 38–50. URL: <https://digitalia.cultura.gov.it/article/view/3008>, doi:10.36181/digitalia-00059. 1
- [Hyv12] HYVÖNEN E.: Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology* 2 (10 2012), 1–159. doi:10.2200/S00452ED1V01Y201210WBE003.
- [MEH19] MERGEL I., EDELMANN N., HAUG N.: Defining digital transformation: Results from expert interviews. *Government Information Quarterly* 36 (06 2019), 101385. doi:10.1016/j.giq.2019.06.002.
- [SSC15] SNYDMAN S., SANDERSON R., CRAMER T.: The international image interoperability framework (iiif): A community amp; technology approach for web-based images. *Archiving Conference* 12, 1 (2015), 16–16. URL: <https://library.imaging.org/archiving/articles/12/1/art00005>, doi:10.2352/issn.2168-3204.2015.12.1.art00005. 5