SCIRES it
SCIentific RESearch and Information Technology
Ricerca Scientifica e Tecnologie dell'Informazione
Vol 16, Issue 1 (2026), 129-146
e-ISSN 2239-4303, DOI 10.2423/i22394303v16n1p129

CASPUR-CIBER Publishing, http://www.sciresit.it

TRACING INTERPRETATION: A HUMAN-IN-THE-LOOP WORKFLOW FOR DISTILLING CONCEPTUAL SYSTEMS FROM ARCHITECTURAL THEORY WRITINGS

Sharon Giammetta, Pieter Pauwels***

*University of Padua – Padua, Italy.
**Eindhoven University of Technology – Eindhoven, The Netherlands.

Abstract

The reconstruction of conceptual systems emerging from architectural theoretical writings is traditionally based on the interpretive reading of sources. These writings are fundamental for understanding the cultural identity of a historic building. However, this knowledge is expressed in a discursive, often implicit, and unstructured form. Scholars often mentally reconstruct the relationships between concepts, then express them in narrative form. This study proposes a human-in-the-loop workflow to trace, as much as possible, some aspects of the interpretive process and model these distilled conceptual structures. Starting from a research question born from the observation of some recurring clues during the first reading of the sources, computational corpus analysis helps identify lexical signals, which are subsequently interpreted through close reading, traced through TEI encoding, and formalized in RDF to make conceptual relationships explicit.

Keywords

Quantitative Text Analysis, Interpretation, RDF, Knowledge Graph, Semantic Graph, TEI Encoding, Human-In-The-Loop.

1. *Introduction*

Historical archives of contemporary architecture often contain many unstructured textual documents. Alongside drawings and technical documentation, theoretical writings and reflective essays often constitute the conceptual backbone necessary to interpret architectural practice. Architects' writings often contain conceptual formulations and reflections, sometimes even philosophical, through which the authors articulate design principles, theoretical positions, visions of architecture, and abstract concepts that are fundamental for understanding their design thinking and, consequently, the cultural identity of their works.

The digitization of sources has made documents more accessible, but it has not substantially transformed the traditional modes of historiographic analysis. Documents can be queried through keyword searches thanks to Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) technologies, which facilitate the localization of terms but do not allow conceptual access to the contents. The challenge of identifying concepts in large textual corpora has already been a subject of interest in Digital Humanities (DH), where several studies have explored computational approaches that go beyond simple keyword search. These approaches are based on the statistical identification of sets of co-occurring terms derived from selected text passages, which are then used to retrieve conceptually related content within the corpus using probabilistic measures (Alexander et al., 2014; Osadetz et al., 2018; Ruiz Fabo & Poibeau, 2019). The meaning of a concept, in fact, is often scattered across discursive fragments that only make sense through a broader interpretive process which, in traditional historiography, takes place through close reading of the sources, a cognitively demanding activity. Textual passages are linked, documents are compared, and conceptual relationships are progressively distilled within an interpretive theoretical framework. However, the result of this process is expressed as linear prose (essays or monographs). Such discursive reformulation, whether extensive or not, can once again render the underlying relational structure implicit, making it more difficult to observe and reconstruct, both for a human reader and for a machine. Concepts are described and discussed, while the relationships between them are incorporated into the narrative rather than formalized. This implies that the underlying theoretical structure must be at least

partially reconstructed by the subsequent reader through a new interpretive process. Although concept detection and textual analysis have been widely explored in DH, few studies address how such conceptual structures, reconstructed through interpretive close reading, can be traced and made more explicit through structured semantic relations.

In the field of DH, quantitative text analysis has provided useful tools for exploring large corpora at scale (Piper, 2016). Text mining techniques allow the identification of lexical patterns, frequencies, and statistical distributions, making it possible to rapidly detect relevant signals within large sets of documents (McGillivray & Tóth, 2020), which makes them particularly useful for supporting exploratory analysis.

On the contrary, encoding and semantic markup technologies, such as XML-TEI and RDF, support the structuring and explicit representation of information. XML-TEI is an encoding standard developed in the humanities that enables the structuring and annotation of digital humanities texts, thereby enriching their informational content and increasing their machine readability (Burnard, 2014). RDF, a foundational technology of the Semantic Web, structures data as a network of relations between entities using sets of subject-predicate-object triples that form semantic graphs (Berners-Lee et al., 2001), enabling the explicit representation of these relations.

The gap therefore lies not so much in the absence of technical tools, but in the lack of replicable approaches capable of making some aspects of the interpretive process traceable and of representing the conceptual structures that emerge from it. The risk, however, is to shift the focus of research from historical analysis to technical tasks, such as data preparation or text encoding, that can effectively become ends in themselves.

The present contribution proposes a human-in-the-loop workflow that integrates quantitative computational exploration and qualitative interpretation to distill, trace, and make more explicit the conceptual system emerging from a corpus of theoretical writings by the Italian architect Francesco Mansutti. The investigation begins with the observation of a stylistic recurrence within the corpus, namely the capitalization of certain terms (Fig. 1), interpreted as a possible clue to relevant theoretical concepts. This approach echoes Ginzburg's indiciary paradigm (Ginzburg, 1986), in which the investigation begins from apparently marginal traces that are not understood as proofs, but as evidence[1] that guides the interpretive process (Ginzburg, 2022). Among the terms identified, the term *Architecture* was selected as a demonstrative case for conceptual modeling, since it is historically characterized by multiple definitions and constitutes a central theoretical concept around which each architect elaborates their own position.

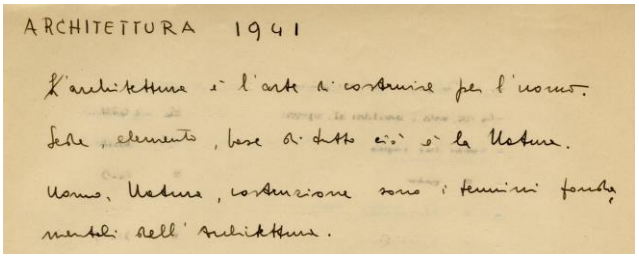


Fig. 1: Detail of a manuscript by Francesco Mansutti showing capitalized terms such as *Uomo* (Human Being), *Natura* (Nature) and *Architettura* (Architecture), interpreted as a preliminary investigative clue. Source: MART. Archivio del '900, Fondo Architetti Francesco Mansutti - Gino Miozzo (hereafter FMM), Man.-Mio.4.6.8.4.

2. State of the Art

2.1 Quantitative and Qualitative Approaches

Quantitative analysis, entrusted to computational calculation, and qualitative analysis, grounded in human interpretation, are often perceived as distinct approaches (Ginzburg, 2019). Within the field of DH, this tension has been widely discussed in relation to the relationship between large-scale computational analysis (*distant reading*) and *close reading* (Ginzburg, 2019; Piper, 2016; Rockwell & Sinclair, 2016).

The use of computers in the humanities has opened significant methodological possibilities. Quantitative approaches make it possible to extend analysis to entire corpora, addressing the problem of the representativeness of evidence by allowing scholars to assess whether observations derived from limited samples recur across larger collections of texts (Piper, 2016). Such

[1] The distinction between *proof* and *evidence* is explicitly discussed in the note to the Italian edition of Ginzburg's work (*History, Rhetoric, and Proof*), where Ginzburg relates it to the Aristotelian distinction between technical and non-technical proofs.

developments are sometimes perceived as a radical innovation capable of transforming research, but also as something that risks marginalizing human interpretation in the process of knowledge construction. As Ginzburg observes, the rigour of quantitative investigations cannot dispense with qualitative analysis, since the computer «*does not think but executes*» and operates within interpretive categories established by the scholar (Ginzburg, 2019).

If computational analysis expands the scale of observation, microhistory and the indiciary paradigm (Ginzburg, 1986) remind us that the reduction of scale is not a retreat into the particular, but a strategy for addressing general questions through a singular case. From this perspective, meaning does not coincide with statistical prominence, but may emerge from marginal details, anomalies, and deviations from the norm (Ginzburg, 2019). Historical knowledge is therefore constructed through a constant tension between series and clue, between recurrence and singularity. The challenge, therefore, is not to choose between *distant reading* and *close reading*, but to articulate workflows that enable an iterative movement between computational exploration and human interpretation. However, in both approaches the interpretive process tends to remain largely implicit: in *distant reading*, in the selection of parameters, thresholds, and analytical categories; in *close reading*, in the contextual inferences and conceptual connections made by the scholar. Computation may turn into a black box (Rockwell & Sinclair, 2016), but interpretation can also remain partially opaque, as inferences, intuitions, and associations that lead from evidence to interpretation are difficult to make fully explicit. If one wishes to prevent technique or hermeneutic judgment from remaining entirely invisible, it becomes necessary to ask how some aspects of the interpretive process can be traced and represented more explicitly.

2.2 Statistical and Symbolic Approaches in DH

This tension between quantitative and qualitative approaches is also reflected in the computational methods applied to the humanities. Such methods do not constitute a homogeneous set; rather, they reflect a long-standing methodological contrast in computational linguistics and artificial intelligence between statistical approaches, based on probabilistic methods to infer patterns from natural language texts and symbolic or declarative approaches, based on formal languages and hand-built syntactic rules (Indurkhya & Damerau, 2010), which appealed to linguists and computer scientists and contributed to the early development of artificial intelligence.

Statistical approaches, which have become widespread within the so-called *Computational Humanities* (CH)[2], largely employ *Natural Language Processing* (NLP) and *text mining* techniques for the automated analysis of large textual corpora. These techniques not only allow the systematic analysis of large sets of documents, but also enable the generation of new research questions through the identification of patterns, co-occurrences, and latent structures in textual data (Biemann et al., 2014).

Within the symbolic-declarative tradition, structured annotation and markup practices have become established to make textual structures and relations explicit. In this sense, although statistical approaches have provided promising results in data extraction, they often depend on such forms of annotation for processing textual data. Within this second tradition, the Text Encoding Initiative (TEI) is one of the longest-lived and most influential projects in the field of the DH and has become a well-established standard for text encoding in the humanities, including source texts, manuscripts, archival documents, ancient inscriptions, and many others (Burnard, 2014). The project originated in 1987 following a conference promoted by the *Association for Computers and the Humanities* (ACH) (Fiormonte et al., 2020). Based on the extensible markup language XML (eXtensible Markup Language)[3], published by the World Wide Web Consortium (W3C) in 1998 but with origins in document preparation systems of the 1980s, TEI allows a text to be represented as a linear sequence of

[2] The term *Computational Humanities* emerged from the *Dagstuhl Seminar 14301: Computational Humanities - bridging the gap between Computer Science and Digital Humanities* (2014), which brought together researchers to outline this field as a distinct research area focused on the computational analysis of humanities data.

[3] Early approaches to Named Entity Recognition (NER), a subfield of Natural Language Processing (NLP), relied on manually annotated corpora based on markup languages such as SGML (and later XML), where entities were explicitly tagged within the text.

characters enriched with markup elements that make explicit its structure and certain features relevant for interpretation. The TEI guidelines indeed provide an articulated set of elements for describing the logical structure of a document, identifying entities such as persons, places, and dates, and recording phenomena typical of critical editions or linguistic analysis (Fiormonte et al., 2020).

Research projects in the historical-architectural field that have made use of TEI include *Architectura Sinica* (Miller et al., 2023; Vanderbilt University, 2014), dedicated to the study of traditional Chinese architecture through the digital edition and encoding of textual and epigraphic sources integrated with geographic and iconographic data. The recent project *CoenoBIuM* has also applied this tool for the digital edition, semantic annotation, and comparative analysis of early modern monastic sources (Guidarelli et al., 2025; Papa, 2026). In such contexts, TEI allows for the representation and querying of the structure of documentary sources through the explicit marking of textual elements. However, in some TEI-based implementations, information continues to be organized primarily around the document, while conceptual relations often remain embedded within the encoded text.

In this sense, the Resource Description Framework (RDF) offers a complementary model that makes these relationships explicit and queryable as autonomous entities. Developed by the W3C as part of the Semantic Web technologies, RDF introduces a graph-based approach to data modeling in which information is represented through subject-predicate-object triples, that is, elementary assertions describing entities and their relationships independently from the sequential structure of the original document (W3C, 2003). Its logical structure can be represented in two ways: an assertion-based perspective, in which each triple constitutes a textual declaration about a resource, and a graph-based perspective, in which triples correspond to labeled edges connecting nodes within a graph (Fig. 2).

In this sense, relations are no longer contained within a linear textual structure but are externalized into an explicit and queryable network, like a conceptual map. This network is formalized through declarative languages such as RDFS and OWL, which allow the definition of classes, properties, and logical constraints, introducing semantic typing of entities and relations.

```
1   :architecture at:is :artOfBuilding .
2   :artOfBuilding at:for :humanBeing .
```

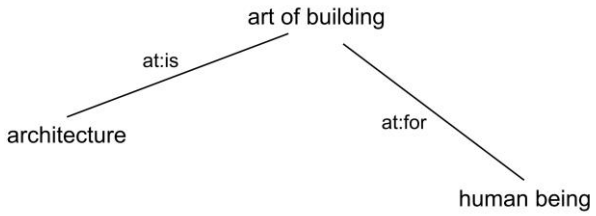


Fig. 2: Turtle serialization and graph-based representation of part of the definition «*Architecture is the art of building for Human Being*».

Unlike procedural programming languages, which describe sequences of instructions to be executed, declarative languages do not prescribe actions but model the meaning of a domain by defining entities and the semantic relations that link them (Halpin, 2008). An ontology is, in fact, a shared formal conceptualization of a domain (Borst, 1997; Gruber, 1993). These structures form the basis for representing information in a machine-readable format and for supporting inference processes. In this sense, they are a fundamental component of the so-called *Knowledge Graphs* (KGs), broad systems that integrate, link, and make queryable information from diverse sources (Ehrlinger & Wöß, 2016). However, in this work, the term semantic graph is preferred, referring to the original interpretation by Sowa (Sowa, 1992).

Semantic graphs have progressively been adopted in Cultural Heritage (CH) contexts to interconnect heterogeneous data sources and enable forms of semantic navigation. Initiatives such as the *Venice Time Machine* (Kaplan, 2015), which uses data extraction and modeling techniques to analyze at scale the documents of the Venetian State Archives and reconstruct the historical evolution of the city, or projects such as *Venice's Nissology* (Galeazzo et al., 2024), aimed at building a semantic infrastructure for integrating museum, archival, and territorial metadata using the *CIDOC Conceptual Reference Model* (CIDOC-CRM), demonstrate how graph modeling enables the connection of people, places, events, and objects through formalized and queryable relations. CIDOC-CRM, developed by the *International Committee for Documentation* (CIDOC) of the *International Council of Museums*

(ICOM), is a reference ontology for the semantic modeling of CH that supports interoperability between heterogeneous datasets by providing a shared conceptual model for describing entities and their relationships over time (Bekiari et al., 2021). In such cases, graphs generally operate on relatively stable entities, such as historical actors, artifacts, and documented events. Consequently, most existing semantic models are designed to represent relatively stable entities and factual relations.

2.3 Semantic Graphs as Conceptual Modeling Tools

While in cultural heritage contexts, semantic graphs are used primarily to represent relatively stable entities and historical events, applying such approaches to theoretical texts raises a different challenge. In these cases, the entities to be modeled do not necessarily correspond to identifiable actors or objects, but rather to abstract concepts expressed through theoretical terms. The formalization of these conceptual domains has been addressed within the framework of Semantic Web technologies through models such as the *Simple Knowledge Organization System* (SKOS), developed within the W3C Semantic Web activities and formalized as a recommendation in 2009 (W3C, 2009). Unlike declarative languages such as OWL, which are oriented toward the formal definition of classes and logical constraints, SKOS is designed as a more lightweight conceptual model for representing *Knowledge Organization Systems* (KOS) such as thesauri, taxonomies, and classification schemes in an interoperable and reusable format on the Semantic Web (Miles et al., 2005; W3C, 2009). For this reason, it is widely used to formalize specialized vocabularies, as in the case of the *Getty Art & Architecture Thesaurus*, developed through collaborative processes involving domain experts and user communities (Miles et al., 2005).

3. Methodology

From this perspective, semantic graphs can be employed not only as data integration infrastructures, but also as tools to make explicit the conceptual structures underlying Mansutti's theoretical discourse. The adopted workflow (Fig. 3) was structured as an iterative human-in-the-loop process, in which computational procedures support the corpus exploration and graph generation, while human interpretation guides both the analytical process and the interpretive construction of meaning. Following Ginzburg's indiciary paradigm, the investigation begins with a stylistic clue (capitalized terms), giving rise to the initial interpretive question and guiding the subsequent phases. Quantitative analysis was then used to systematically explore the distribution of these terms within the corpus, followed by qualitative interpretation, selective XML-TEI encoding to preserve traces of the interpretive process, the distillation of concepts and relationships, the automatic generation of an initial RDF semantic graph, and its subsequent interpretive refinement. In essence, the workflow moves from natural language to conceptual structures through interpretation, and then to RDF assertions through semantic modeling, followed by manual refinement.[4]

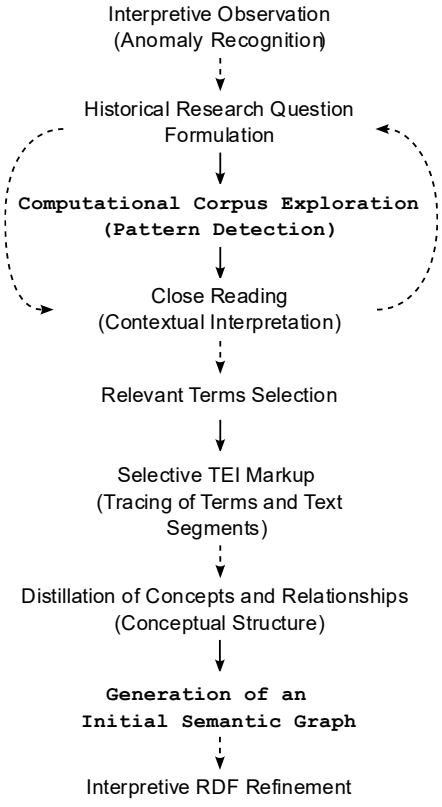


Fig. 3: Human-in-the-loop workflow for reconstructing conceptual systems from architectural theory writings. Different typographical styles distinguish interpretive and computationally supported phases, while dashed arrows indicate heuristic transitions that cannot be fully formalized.

[4] All materials developed for this study are available on Zenodo at: https://doi.org/10.5281/zenodo.20827827

3.1 Corpus Selection and Preparation

A corpus of the architect's theoretical writings was selected for the experimentation. The archive includes both manuscripts and typewritten documents. Since handwritten materials present a higher degree of noise and variable legibility, and many of them also survive in typewritten copies, a sample composed predominantly of typewritten documents was chosen, supplemented by a limited number of manuscripts without the corresponding typewritten copy.

The preparation of the corpus required the production of a queryable textual version of the documents. The digitized materials were uploaded to the Internet Archive platform, which during the upload process automatically generates a series of derivative files, including textual transcriptions based on optical character recognition (OCR) technologies[5]. The system mainly uses the open-source OCR engine Tesseract and produces transcriptions in hOCR format, which also preserve layout information useful for full-text search (Bromley, 2021; Wajer, 2023). This infrastructure, increasingly adopted by cultural institutions for the online publication of digital collections, was used as a practical environment for obtaining an initial textual basis for the documents. This made it possible to verify the potential use of such platforms not only as environments for access and dissemination, but also as operational tools in the preparation of textual corpora.

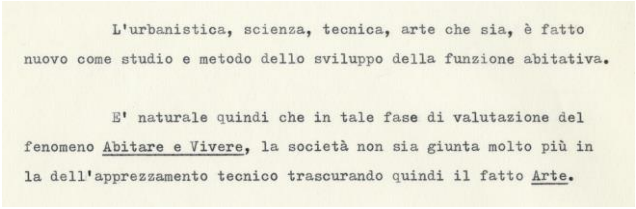
The automatically generated transcriptions were subsequently subjected to a preliminary cleaning phase aimed at reducing typographical noise introduced by the OCR process. The cleaning was conducted by distinguishing between safe structural corrections, which could be performed automatically, and limited manual interventions applied only in cases where segmentation or punctuation errors compromised the integrity of the tokens used for lexical analysis[6].

Manual intervention was deliberately kept minimal and limited to cases where punctuation or segmentation errors compromised the integrity of individual tokens. This approach made it possible to speed up the corpus preparation phase, avoiding excessively pervasive normalization procedures that would risk shifting the focus of the work from historical study to systematic text editing.

At this stage, the objective was not to produce a philologically precise transcription of the entire corpus, but rather to obtain a textual representation sufficiently reliable for the analysis of capitalization patterns. This pragmatic approach reflects common conditions in historical research, where textual corpora often need to be progressively constructed from heterogeneous archival materials rather than derived from already structured digital datasets.

3.2 Interpretive Starting Point: Capitalization as Investigative Clue

In the present case, the process did not originate from a generic application of statistical techniques, but from the observation of a recurring stylistic clue within the writings, namely the systematic capitalization of certain terms (Fig. 4). This writing style, already perceptible from an initial reading, suggested a possible conceptual value attributed by the author to specific words. Capitalization was therefore adopted as a clue to investigate the hypothesis that these terms corresponded to central theoretical concepts in the author's discourse.


L'urbanistica, scienza, tecnica, arte che sia, è fatto nuovo come studio e metodo dello sviluppo della funzione abitativa.

E' naturale quindi che in tale fase di valutazione del fenomeno Abitare e Vivere, la società non sia giunta molto più in la dell'apprezzamento tecnico trascurando quindi il fatto Arte.

Fig. 4: Example of systematic capitalization in a typewritten document. Terms such as *Abitare* (Dwelling), *Vivere* (Living), and *Arte* (Art) are intentionally capitalized by the author, suggesting a potential conceptual function within the theoretical discourse. Source: MART. FMM, Man.-Mio.4.6.5.63.

3.3 Quantitative Signal Detection

Starting from the interpretive clue identified in the previous phase, a computational analysis was

[5] A sample of documents was also tested with LLM-based transcription tools (e.g., Caracal), which showed promising results. However, for this experiment the OCR transcriptions generated by Internet Archive were used to assess the quality of a widely adopted infrastructure for publishing and managing digital collections.

[6] Automatic corrections were limited to non-semantic operations, such as removing line-break hyphenation (-\n) and normalizing multiple spaces. More invasive edits (e.g., punctuation changes, removal of corrupted lines, or normalization of ambiguous characters) were avoided to prevent lexical distortions in the corpus.

conducted to detect capitalized terms within the corpus.

The objective was not to automatically determine the conceptual centrality of the terms, but rather to identify potentially significant lexical patterns requiring subsequent interpretation.

The analysis considered exclusively cases in which the same term appears both in capitalized form and in ordinary form, to exclude false positives resulting from the initial position of the word within a sentence. Through Python-based scripts, the target terms were extracted and their absolute frequency and distribution within the corpus were measured.

Rather than considering frequency as a direct indicator of conceptual centrality, the quantitative data were used comparatively within the corpus. In particular, the difference between capitalized usage and the ordinary usage of the same terms was examined to identify lexical deviations that could indicate intentional conceptual marking.

Unlike text analysis tools that automatically apply lexical normalization procedures, no normalization of linguistic forms was performed. In this case, the graphical variation between uppercase and lowercase itself constituted the phenomenon under observation.

From this perspective, quantitative analysis was not used as a tool for determining meaning, but as a means of identifying candidate tokens to be subjected to subsequent phases of close reading and contextual interpretation.

3.4 Iterative Close Reading and Interpretation

The computational outputs were subsequently examined through close reading of the sources. This phase aimed to verify the semantic context of the terms identified in the quantitative analysis and to distinguish between purely stylistic recurrences and potentially relevant occurrences from a theoretical perspective. The close-reading phase included:

- interpretive reading of the context of candidate tokens within the paragraphs;
- identification and elimination of false positives.

Through this interpretive reading of the context, candidate concepts considered significant for the analysis were selected. This operation involved the elimination of false positives and the progressive definition of a restricted set of terms that, within the corpus, showed a usage consistent with a possible conceptual function.

In this sense, the close-reading phase also operated as an initial form of interpretive distillation, aimed at reducing the set of terms identified by the quantitative analysis to only those conceptual candidates considered relevant.

The analysis was conducted iteratively. Quantitative results provided further clues for the qualitative reading of the sources, while contextual interpretation progressively refined the selection of relevant concepts.

3.5 Selective XML TEI Encoding for Tracing the Interpretive Observations

The concepts considered relevant in the previous phase were subsequently annotated through XML-TEI. Rather than proceeding with an exhaustive encoding of the entire corpus, a selective annotation strategy guided by the research question was adopted. TEI encoding was used to:

- Structure document divisions and paragraphs;
- Annotate the relevant concepts;
- Identify the textual segments in which these concepts appear and are qualified.

In this context, TEI played a dual role. On the one hand, it enabled the preservation of the documentary structure of the sources while maintaining an explicit link to the original text. On the other hand, it allowed interpretive observations to be traced by recording both the relevant concepts and the textual segments associated with them.

However, although TEI makes it possible to structure textual content and preserve the philological connection with the source, the conceptual relations identified remain implicit within the documentary structure.

To make them explicit, a subsequent phase of semantic modeling was therefore introduced.

3.6 Conceptual Distillation and Graph Construction Criteria

Starting from the textual segments identified in the previous phase, conceptual assertions were distilled through a process of interpretive abstraction. In this phase, the discursive formulations present in the texts were progressively translated into explicit conceptual relations, with the aim of making the semantic relationships implicit in the author's definitions and qualifications formally representable.

The transformation from textual formulations to relational structures followed several recurring operational criteria:

- Selection of the main concepts expressed within the definitional segments;
- Individuation of related or qualifying concepts that contribute to specifying their meaning;
- Explicitation of relational predicates present in the language of the sources, transforming discursive assertions into relational structures representable in the form of RDF triples;
- Classification of the identified relations to detect possible recurring patterns useful for the subsequent selection of an ontology suitable for their formalization.

The selected textual segments were then first transformed into RDF assertions, serialized in Turtle format (TTL), while preserving as much as possible the relational formulations present in the sources and without referring to existing ontologies. This approach made it possible to preserve the link between the semantic structure of the graph and the author's discursive language, avoiding premature interpretive simplifications that could alter the theoretical meaning of the relations expressed in the text. This choice allowed a preliminary exploration of the relational structure emerging from the texts before any semantic normalization operations. The conceptual relations were therefore modeled cautiously, preserving the interpretive nuances present in the sources. The resulting graph should thus be understood as a first-level semantic representation based directly on the textual formulations derived directly from the author's original formulations.

4. *Results*

The results presented in this section focus less on the technical aspects of the computational procedures and more on the interpretive outcomes of the workflow described in the methodological section. Attention is therefore placed on how the quantitative exploration of the corpus, interpretive analysis through close reading, selective TEI encoding, and subsequent semantic modeling progressively made explicit conceptual relations that were initially dispersed and implicit in the texts. From this perspective, the main contribution of the analysis lies in the interpretive process of conceptual distillation that leads from the identification of lexical signals within the corpus to the formalization of theoretical relations between concepts.

4.1 Quantitative Signal Detection and Extraction

The analyzed corpus consists of 42 documents, including 142 typewritten pages and 11 manuscripts, for a total of 23,868 tokens. The initial extraction phase identified 454 lexical tokens distributed across the corpus both in capitalized and lowercase form (Tab. 1).

Tab. 1: Corpus statistics description.

Corpus statistics	Value
Number of documents	42
Total word tokens	23868
Capitalized tokens	454

To assess whether capitalization was linked to a conceptual meaning rather than to grammatical conventions, only those terms for which a corresponding lowercase form was also present in the corpus were selected for further analysis.
The quantitative phase produced:

- A preliminary list of candidate conceptual terms;
- Frequency ratios between capitalized and non-capitalized forms;
- Distribution patterns across different documents.

Capitalization was not considered an automatic indicator of conceptual relevance but functioned as an interpretive clue enabling the rapid identification of areas of the corpus characterized by high semantic density.

4.2 Interpretive Analysis and Conceptual Filtering

The following phase concerned the interpretive analysis of the 277 candidate terms selected after the preliminary quantitative filtering, with the aim of distinguishing between capitalization resulting from grammatical or syntactic reasons and capitalization used by the author with a conceptual function. Through an iterative process of close reading of textual contexts, the following operations were carried out:

- False positives were eliminated (for example capitalizations due to the syntactic position at the beginning of a sentence or occurring in titles);

- Terms showing effective conceptual relevance were retained.

This process led to the reduction of the 277 candidate terms to 58 conceptually relevant terms.

The comparison between capitalized forms and their corresponding lowercase occurrences showed that capitalized forms represent a relatively small percentage of the overall lexical frequency of the corpus (Tab. 2). This distribution is consistent with the discursive nature of theoretical texts, since if capitalization were dominant the text would assume the form of a continuous typographic manifesto. The relative rarity of capitalized forms instead makes these occurrences particularly interesting as potential indicators of semantic density.

Contextual reading also revealed that some capitalized terms lack conceptual function, while other relevant concepts emerge primarily in their lowercase form. Capitalization therefore does not always represent a reliable signal of potential conceptual relevance and may sometimes suggest an emphatic or rhetorical use within the architect's discourse.

Among the terms showing the highest conceptual density, *Architecture* emerges in particular, frequently appearing within textual segments with a definitional function. In these passages, the author explicitly introduces reflections on the theoretical meaning of the term, suggesting a process of explicit conceptualization. Similarly, terms such as *To build* or *To Construct* sometimes appear within definitional formulations (for example «*Architecture is [the act of] building*»), indicating a conceptual use of the term that does not necessarily depend on capitalization. Other terms appear relevant primarily because of their high frequency within the corpus, such as *Human Being* or *Life*, which, although occurring predominantly in lowercase form, reveal a strong presence within the author's theoretical discourse. In these cases, conceptual relevance emerges less from capitalization than from distribution and argumentative context.

Based on this interpretive analysis, the selected terms were divided into two categories:

- First-level concepts, identified because they are explicitly thematized or defined in the texts;
- Second-level concepts, which acquire meaning primarily in relation to other main concepts or as qualifications of the theoretical discourse.

Among the first-level concepts are terms such as *Architecture, to build, to construct, Art,* and *Nature*, while second-level concepts include terms such as *Education, Principle,* and *Modern*, which mainly perform a secondary or qualifying function.

A significant result of this phase is that several central concepts do not emerge directly from quantitative extraction but are identified only through interpretive analysis within their textual context. This result confirms that the computational detection of lexical patterns, although useful for identifying preliminary signals, is not sufficient to automatically identify ontologically meaningful concepts in theoretical texts. The terms selected in this phase therefore constitute the basis for the subsequent identification of textual segments with a definitional function, which are used for the semantic modeling phase.

4.3 Identification and Encoding of Definitional Segments

The transition from the identification of terms to the reconstruction of conceptual relations was not produced automatically by the computational system but emerged during the contextual reading of the textual segments in which the selected terms appeared.

Based on the list of concepts deemed relevant in the previous phase, the analysis focused on identifying textual segments with a definitional

Tab. 2: Example of first- and second-level concepts identified in the corpus.

Concept	Capitalized occurrences	Lowercase occurrences	Conceptual role
[Architecture]	18	91	first-level concept
Edificare [To construct]	4	5	first-level concept
Costruire [To build]	6	40	first-level concept
Uomo [Human Being]	2	83	first-level concept
Vita [Life]	4	241	first-level concept
Educazione [Education]	3	10	second-level concept
Principio [Principle]	1	24	second-level concept

Tab. 3: Examples of definitional segments.

Concept	Segment ID	Segment	Source ID
[Architecture]	0001	«*Architettura è l'arte di costruire per l'uomo*» [«*Architecture is the art of building for human beings*»]	man_-mio_0004_0006_0008_0004
[Architecture]	0002	«*Uomo, Natura, costruzione sono i termini fondamentali dell'Architettura*» [«*Human Beings, Nature, and construction are the fundamental terms of Architecture*»]	man_-mio_0004_0006_0008_0004
[Architecture]	0001	«*Architettura è costruire*» [«*Architecture is to build*»]	man_-mio_0004_0006_0007_0038
[To build]	0002	«[…] *costruire, sempre nel suo significato essenziale e cioè umano di questo atto, si identifica con edificare che vuol dire dar vita*» [«[…] *always in its essential, meaning human, sense of this act, identifies with to construct, which means to give life*»]	man_-mio_0004_0006_0007_0038
[Education]	0001	«*L'educazione è fattore basilare dell'Architettura*» [«*Education is a fundamental factor of Architecture*»]	man_-mio_0004_0006_0008_0045

function (Tab. 3), namely the passages in which the author explicitly attempts to clarify the meaning of a concept or to establish conceptual relations between terms. These segments were identified through the observation of recurring syntactic structures such as:

- Identification relations (e.g., *X is Y*);
- Explicit definitional formulations (e.g., *X means…*);
- Compositional relations (e.g., *X and Y are the fundamental terms of Z*);
- Qualified equivalences (e.g., *In its essential meaning…*).

The analysis showed that such formulations characterized by high semantic density are relatively rare within the corpus and are often embedded in broader passages of a narrative or polemical nature. The identification of these segments therefore required a targeted filtering process aimed at isolating the declarative core of the definitions.

Attention was devoted to segments in which the author explicitly reflects on the meaning of *Architecture*, a term that emerges as one of the conceptual nodes of the corpus. In these passages, he attempts to clarify what architecture should be and what role it plays in the relationship between construction, human beings, and society.

The analysis therefore focused on those segments in which such theoretical reflections or definitions are made explicit, using them as the primary textual units for the subsequent phases of encoding and semantic modeling.

From a methodological perspective, the identification of definitional segments was conducted iteratively. The results of the quantitative analysis oriented the qualitative reading of the sources, while the observation of the context progressively refined the selection of relevant terms and segments.

In this phase, TEI encoding played an important role in maintaining the link between the selected concepts and their corresponding textual segments. By preserving the document's structure, TEI made it possible to track interpretive observations by precisely marking the passages where concepts are defined or related (Fig. 5). The selected segments in this way thus constitute the basis for the subsequent phase of semantic modeling of conceptual relations.

4.4 Interpretive Distillation of Conceptual Structures and Semantic Modeling

As a first step in semantic modeling, the textual segments identified as defining the concept of *Architecture* were transformed into intermediate conceptual structures, distilling the concepts and semantic relations embedded in the original discourse. These conceptual structures were then translated into a series of RDF assertions, formalized as triples (subject-predicate-object) and serialized in Turtle format (TTL) (Fig. 6).

At this preliminary stage, no specific ontology was adopted. The triples were expressed both in Italian, to preserve the link with the original formulations of the sources, and in English, to ensure accessibility for an international scholarly audience[7]. The representation relied exclusively on the RDF triple model, preserving the author's original predicates and relational formulations.

The analysis of the selected triples highlighted several aspects concerning both the concepts represented and the relations connecting them.

First, considering the concepts (subjects and objects of the triples), the graphical representation made it possible to identify different structural roles within the corpus. Some concepts appear central within the definitional structure and act as generators of conceptual clusters (first-level concepts). Others instead function as qualifying concepts or as connecting elements between multiple definitional chains (second-level concepts). For example, in the textual segment *«Architecture is the art of building. Its seat, the fundamental element of everything, is Nature»*, concepts such as *Architecture* and *Nature* assume a structurally central role, generating further definitional relations with other concepts in the corpus. Considering instead the relations, the analysis of the triples revealed the presence of relations characterized by different levels of explicitness:

- Explicit definitional relations, directly formulated in the text;

```
48
49      <text>
50        <body>
51          <p>
52            <seg xml:id="seg_man_mio_0004_0006_0008_0004-0001" type="definition">
53              L'<term ref="architettura">architettura</term> è l'<term ref=
                "arte_di_costruire">
54              arte di costruire</term>. Sede, elemento, base di tutto ciò è la
55              <term ref="natura">Natura</term>.
56            </seg>
57
58            <seg xml:id="seg_man_mio_0004_0006_0008_0004-0002" type="definition">
59              <term ref="uomo">Uomo</term>, <term ref="natura">Natura</term>,
60              <term ref="costruzione">costruzione</term> dell'
61              <term ref="architettura">Architettura</term>.
62            </seg>
63          </p>
64        </body>
65      </text>
66    </TEI>
```

Fig. 5: Example of XML-TEI annotation of definitional segments. Concepts selected during interpretive analysis are marked with the <term> element, while definitional passages are encoded as <seg type="definition"> to preserve the explicit link between concepts and textual evidence.

```
97    # --- Segment 3 ---
98    # [IT] "Architettura è costruire, e costruire, sempre nel suo significato essenziale
99    # e cioè umano di questo atto, si identifica con edificare che vuol dire dar vita"
100   # [EN] "Architecture is to build, and to build, always in its essential, meaning
      human,
101   # sense of this act, identifies with to construct, which means to give life"
102
103   :architecture at:is :toBuild .
104   :toBuild at:identifiesWith :toConstruct .
105   :toConstruct at:means :toGiveLife .
```

Fig. 6: Example of Turtle serialization of RDF assertions extracted from a definitional segment. The syntax preserves the relational formulations of the source text while transforming the discursive definition into explicit subject-predicate-object assertions.

[7] The author manually translated the concepts and predicates into English, as this step was part of the interpretive process itself and required the evaluation of context-dependent meanings.

- Implicit interpretive relations, emerging from the comparison of multiple segments or from the broader discursive context.

For example, in the passage «*To construct – The construction [industry], when it is truly architecture, is construction in its deepest and most complete sense*», the relation between *construction industry*, *architecture*, and *construction* appears strong, whereas the connection between these and *To construct* appears more indirect.

Beyond the relations between concepts, modeling also revealed relations between the textual segments themselves. Some segments appear to expand or refine previously introduced definitions, generating relations of enrichment between different passages of the corpus. At the same time, some definitions themselves contain further definitions of related concepts, generating nested conceptual definitional dependencies within the discourse.

For example, in one document (Man.-Mio.4.6.6.9) the segment «*Architecture is the most complete expression of the society of a given historical period*» appears, whose conceptual structure can be expressed through assertions such as:

Architecture → isExpressionOf → Society AND
Society → existsIn → Historical Period

In another document (Man.-Mio.4.6.6.10), the segment «*The architecture of a people, in a given era, is the result of civilization and, more specifically, of the degree and level of humanity of that people in competition with the environment and with the technical and economic factors of that time*» introduces additional factors that extend and refine the previous conceptual structure. These two passages are not independent, but form a nested structure that can be decomposed as follows:

Architecture → isExpressionOf → Society AND
Society → existsIn → Historical Period AND
Architecture → isResultOf → Civilization AND
Civilization → includes → Level of Humanity AND
Civilization → includes → Environment AND
Civilization → includes → Technical Economic Factors

These can then be translated into RDF assertions (Fig. 7). The definitions, distributed across different documents, contribute to the progressive construction of a shared conceptual configuration. When translated into RDF, this configuration is atomized into a network of interconnected assertions. Consequently, the resulting graph does not simply represent a set of isolated triples, but a semantic network in which groups of isolated assertions become organized around central concepts (Fig. 8).

5. Discussion

5.1 Quantitative and Qualitative Analysis as an Interpretive Loop

The results show that quantitative and qualitative analysis do not constitute alternative approaches but rather complementary phases of a single interpretive process. Computational analysis allows large corpora to be navigated rapidly, identifying signals distributed across documents. In this way, an interpretive hypothesis that emerges during an initial reading of the sources can be investigated across the corpus through the large-scale exploration made possible by the computer.

At the same time, the research experience shows that quantitative analysis cannot replace qualitative interpretation. Without a clear research question or an interpretive clue guiding

```
119   :architecture at:isExpressionOf :society .
120   :society at:existsIn :historicalPeriod .

130   :architecture at:isResultOf :civilization .
131   :architecture at:isResultOf :humanLevel .
132   :architecture at:inInteractionWith :environment .
133   :architecture at:inInteractionWith :technicalEconomicFactors .
```

Fig. 7: Turtle serialization of RDF assertions composing the two passages.

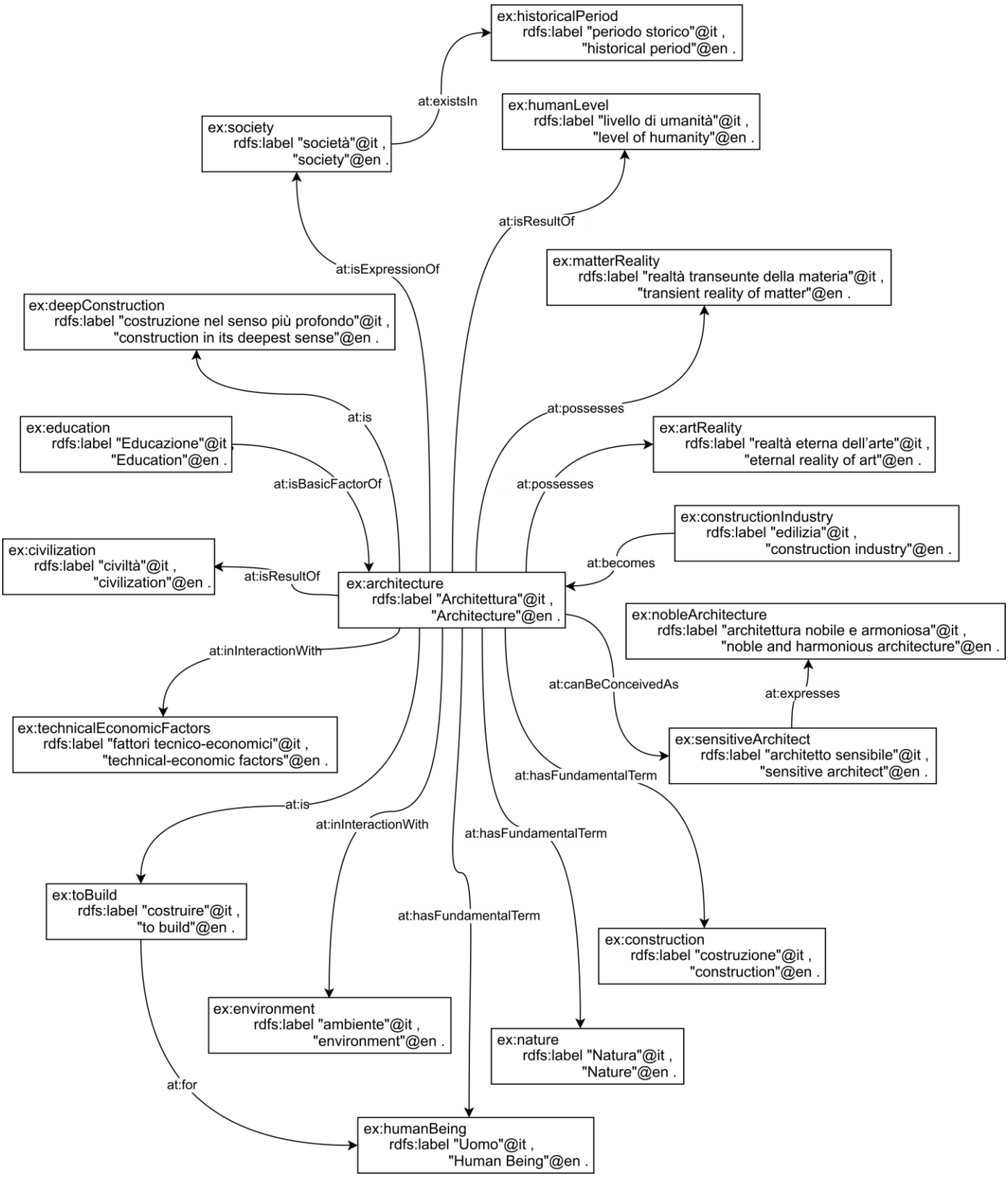


Fig. 8: RDF semantic graph aggregating all RDF assertions related to the concept of *Architecture*, extracted from passages in Mansutti's writings. Nodes represent RDF resources of concepts, while edges represent the semantic relations extracted from the sources.

the investigation, computational exploration risks becoming a blind practice, that is, a sequence of technical operations, such as data preparation, corpus cleaning, or text annotation, that absorb time and resources without producing interpretive advancement. This highlights the need to maintain the historical interpretive question as the cornerstone of the process and to use computational tools for analytical support rather than as an end in themselves.

Computational tools are useful in identifying patterns and anomalies distributed across the corpus that orient the interpretive investigation. However, such signals do not in themselves constitute proofs. Their meaning is contextual and emerges only through close reading of the sources and the subsequent interpretive reconstruction of conceptual relations through historical interpretation.

5.2 Data Structures, Computational Representations, and Cognitive Organization of Information

A second implication concerns the role of different data models within the workflow. In the proposed process, different tools and formats support distinct forms of analysis and representation.

Quantitative corpus analysis makes it possible to identify lexical patterns and large-scale statistical distributions, facilitating the preliminary exploration of signals present in the sources. However, this type of analysis operates primarily at the lexical level, producing lists of terms and statistical data without representing the conceptual relationships linking them.

TEI encoding, instead, operates at the level of textual structure. Through markup, it makes it possible to trace the concepts and textual passages deemed relevant, preserving the link between interpretation and source. However, since it organizes information through a hierarchical and linear text structure, TEI is not designed to directly represent networks of semantic relationships and concepts distributed across multiple documents. Within the proposed workflow, TEI thus plays the role of an intermediate layer, preserving the documentary structure of the sources and preparing the selected concepts for subsequent preliminary relational modeling.

RDF semantic modeling then introduces a different form of epistemic representation, making explicit the relationships between the modeled concepts. The graphical representation of triples (Fig. 8) provides an analytical tool to observe conceptual configurations and definitional structures that are otherwise distributed across multiple documents and would be difficult to grasp through a linear reading of the sources alone.

In this sense, the different data models give rise to different forms of computational representation that are not merely technical alternatives but support distinct cognitive ways of organizing and interpreting information.

5.3 Challenges of Modeling Theoretical Discourse

The analysis of the triples revealed the presence of semantically explicit relations such as conceptual identification relations, in which a concept is defined through another term or process (for example, Architecture → is → building), as well as conceptual conditioning relations, in which a concept is presented as a factor or foundation of another (for example, Education → conditions → Architecture), and so forth. Alongside these, qualified relations also emerged, in which the meaning of the relation is modulated by contextual expressions present in the text.

A significant example is the segment «*Architecture is to build, and to build, always in its essential, meaning human, sense of this act, identifies with to construct, which means to give life*». In this case, the modeling revealed the following conceptual chain:

Architecture → to build
to build (as a human act) → to construct
to construct → to give life

The expression *as a human act* does not introduce a new concept but qualifies the relation between *to build* and *to construct*, indicating that the identification between the two terms is valid only within a specific interpretation of the act of building. This type of element shows how, in theoretical texts, many relations are not simple terminological equivalences, but semantically rich conceptual relations conditioned by the discursive context. Their direct translation into strict logical identities would therefore entail a loss of meaning.

This raises questions regarding the choice of the most appropriate ontologies to support such data. At this early stage, however, the modeling remains at a preliminary level and does not yet involve a formal ontological structure, but it paves

the way for possible future developments. The use of ontologies is not motivated only by the need to visualize conceptual structures, but also by the possibility of explicitly representing relations between concepts and supporting semantic queries or inferences based on those relations.

In the case of theoretical texts, many relations appear semantically qualified, making them difficult to represent through existing ontologies, which require a certain conceptual stability and struggle to manage the ambiguity and varying degrees of interpretive explicitness present in the sources without a significant loss of semantic nuance. SKOS allows associative or hierarchical relations between concepts to be represented while maintaining a high degree of interpretive openness. However, this flexibility tends to flatten the semantic richness of theoretical formulations into generic relations such as *broader*, *narrower*, or *related*. Conversely, more formal declarative languages such as OWL allow more specific relations and explicit logical constraints to be defined, better preserving the relational structure emerging from the texts.

This highlights a challenge in the semantic modeling of theoretical discourse, namely, formalizing conceptual relationships without losing the contextual and interpretive nuances whereby they acquire meaning.

5.4 Graph Modeling as Cognitive Intensification

One of the most relevant results concerns the cognitive role of semantic modeling. Translating textual formulations into relational structures requires making explicit the conceptual dependencies, qualifications, and semantic conditions present in the text. In this sense, the construction of the graph does not merely represent an already formulated interpretation but contributes to making it more explicit and structured. The modeling process forces the researcher to clarify which concepts are central, which play a qualifying role, and which relations can be considered explicit or implicit.

A significant example concerns the nested conceptual structure that emerged from the comparison of definitions distributed across different documents. The attempt to represent these segments within a single relational structure made it necessary to clarify how the different concepts are articulated with one another in contributing to the definition of *Architecture* (Fig. 8). This configuration is not explicitly formulated in the sources but emerged during the modeling process.

In this sense, the act of modeling, although time-consuming, enables a form of high-intensity analysis of the conceptually dense passages of the corpus, intensifying interpretive cognitive activity. Modeling thus becomes an analytical extension of historical reasoning rather than a mere formalization of data aimed at automating analytical processes.

The resulting graph should not be understood as an objective and definitive reconstruction of the author's theory, but as an interpretive construction of the historian, emerging from an iterative process of analysis and synthesis.

6. Conclusion

This study proposed and tested a workflow for the analysis and formalization of theoretical concepts present in the writings of an architect through an iterative human-in-the-loop process. The bottom-up approach allowed conceptual relations to emerge directly from the sources, thus preserving the link between semantic modeling and textual evidence.

Since the process is intrinsically interpretive, tracing the analytical phases becomes essential to make explicit and assessable the path leading to the construction of the semantic graph, understood not as an objective and definitive representation of reality but as an interpretive formalization of historical reasoning.

The results suggest that the construction of semantic graphs should not be considered merely as a data representation technique but as an analytical extension of historical reasoning.

The case study, however, highlights several difficulties in the representation of theoretical texts: conceptual definitions often assume partial, conditional, or contextual forms that are difficult to reduce either to rigidly defined ontological relations or to excessively generic ones. This suggests the need for modeling strategies capable of combining more formal ontological relations with more flexible conceptual vocabularies able to preserve the ambiguity inherent in theoretical discourse.

Future research will further explore this approach to develop semantic models better suited to representing complex conceptual relations and the different levels of explicitness present in theoretical discourse.

The meaning of such concepts rarely coincides with a single term but rather emerges from formulations distributed across different segments of discourse. The possibility of identifying and linking such formulations makes it possible to reconstruct more systematically the conceptual structures present in the architect's writings.

In the analyzed case, the concept of *Architecture* represents one of the central concepts of the author's theoretical reflection, but it constitutes only one of the nodes within his broader conceptual network. The progressive extension of the analysis to other writings will contribute to enriching the understanding of the concept itself and to reconstructing more completely the author's theoretical structure.

Understanding the ideas and theoretical assumptions that guided an author's design choices is fundamental for a culturally more informed reading of buildings and their historical meaning.

In the long term, the progressive collective formalization of such conceptual structures could contribute to the construction of shared knowledge bases on architectural ideas. As open and interoperable data structures, these could be integrated into knowledge bases related to real buildings, enriching their informational dimension with historical and theoretical knowledge that is also relevant for their interpretation and transformation over time.

REFERENCES

Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., & Gleicher, M. (2014). Serendip: Topic model-driven visual exploration of text corpora. *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 173–182. https://doi.org/10.1109/VAST.2014.7042493

Bekiari, C., Bruseker, G., Doerr, M., Ore, C.-E., Stead, S., & Velios, A. (2021). *Definition of the CIDOC Conceptual Reference Model v7.1.1* (Version v7.1.1). The CIDOC Conceptual Reference Model Special Interest Group. https://doi.org/10.26225/FDZH-X261

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*. http://web.dfc.unibo.it/buzzetti/IUcorso2006-07/materiali/bl-engl.html

Biemann, C., Crane, G., Fellbaum, C., & Mehler, A. (2014). Computational Humanities—Bridging the gap between Computer Science and Digital Humanities (Dagstuhl Seminar 14301). *Dagstuhl Reports*, *4*, 80–111. https://doi.org/10.4230/DagRep.4.7.80

Borst, W. N. (1997). *Construction of engineering ontologies for knowledge sharing and reuse* [University of Twente]. https://doi.org/10.3990/1.9789036509886

Bromley, H. (2021, May 24). *Uploading images for text items (update on *_images.zip format)* [Blog]. Blog.Archive.Org. https://help.archive.org/help/files-formats-and-derivatives-a-basic-guide/

Burnard, L. (2014). *What is the Text Encoding Initiative?* OpenEdition Press. https://books.openedition.org/oep/426?lang=en

Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. In M. Martin, M. Cuquet, & E. Folmer (Eds.), *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems—SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016* (Vol. 1695). CEUR-WS.org. https://ceur-ws.org/Vol-1695/paper4.pdf

Fiormonte, D., Tomasi, F., & Numerico, T. (with Rockwell, G.). (2020). *The Digital Humanist: A Critical Inquiry* (C. Ferguson & D. Schmidt, Trans.). Project Muse.

Galeazzo, L., Grillo, R., & Spinaci, G. (2024). A Geospatial and Time-based Reconstruction of the Venetian Lagoon in a 3D Web Semantic Infrastructure. *Proceedings of the 20th Conference on Information and Research Science Connecting to Digital and Library Science (Formerly the Italian Research Conference on Digital Libraries)*, *3643*.

Ginzburg, C. (1986). Spie: Radici di un paradigma indiziario. In C. Ginzburg, *Miti, emblemi, spie: Morfologia e storia* (pp. 158–209). Giulio Einaudi

Ginzburg, C. (2019). *Il formaggio e i vermi: Il cosmo di un mugnaio del '500*. Adelphi.

Ginzburg, C. (2022). *Rapporti di forza: Storia, retorica, prova*. Quodlibet

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. https://doi.org/10.1006/knac.1993.1008

Guidarelli, G., Cavicchioli, S., Borin, P., Rubbi, V., Savy, B., Frank, M., De Paoli, M., Zampieri, P., Brusori, G., Gottardi, S., Gaio, S., Placentino, P., Papa, I., Tonin, R., Pivetta, C., Giammetta, S., & Bernardello, R. (2025). *CoenoBIuM*. www.digitalcoenobium.eu. https://www.digitalcoenobium.eu/

Halpin, H. (2008). The Principle of Self-Description: Identity Through Linking. *IRSW*. https://api.semanticscholar.org/CorpusID:2606607

Indurkhya, N., & Damerau, F. J. (Eds.). (2010). *Handbook of natural language processing*. Taylor & Francis. https://doi.org/10.1201/9781420085938

Kaplan, F. (2015). The Venice Time Machine. *Proceedings of the 2015 ACM Symposium on Document Engineering*, 73–73. https://doi.org/10.1145/2682571.2797071

McGillivray, B., & Tóth, G. M. (2020). *Applying Language Technology in Humanities Research: Design, Application, and the Underlying Logic*. Springer International Publishing. https://doi.org/10.1007/978-3-030-46493-6

Miles, A., Matthews, B., Wilson, M., & Brickley, D. (2005). SKOS core: Simple knowledge organisation for the web. *Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications: Vocabularies in Practice, DCMI '05*.

Miller, T., Anderson, C., Zhuge, J., Zuo, L., & Campbell, A. (2023, September 7). *Using TEI for Chinese Architectural Data*. Https://Teimec2023.Uni-Paderborn.De/. https://teimec2023.uni-paderborn.de/contributions/169.html

Osadetz, S., Courtney, K., DeMarco, C., Crawford, C., & Eslao, C. F. (2018). Searching for Concepts in Large Text Corpora: The Case of Principles in the Enlightenment. In J. G. Palau & I. G. Russell (Eds.), *13th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2018, Mexico City, Mexico, June 26-29, 2018, Conference Abstracts* (pp. 254–256). Red de Humanidades Digitales A. C. https://dh2018.adho.org/en/searching-for-concepts-in-large-text-corpora-the-case-of-principles-in-the-enlightenment/

Papa, I. (2026). *The Architectural Heritage of the Benedictine Cassinese Congregation (15th-18th century): Digital and Spatial Analysis Strategies through BIM Models. The Monastery of San Paolo d'Argon (BG)* [Unpublished PhD Thesis]. Department of Civil, Environmental and Architectural Engineering (ICEA), University of Padua.

Piper, A. (2016). There Will Be Numbers. *Journal of Cultural Analytics*. https://doi.org/10.22148/16.006

Rockwell, G., & Sinclair, S. (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. The MIT Press. https://doi.org/10.7551/mitpress/9522.001.0001

Ruiz Fabo, P., & Poibeau, T. (2019). Mapping the Bentham Corpus: Concept-based Navigation. *Journal of Data Mining & Digital Humanities*, *Atelier Digit_Hum* (Data deluge: which skills for...), 5044. https://doi.org/10.46298/jdmdh.5044

Sowa, J. (1992). Semantic Networks. In *Encyclopedia of Artificial Intelligence* (2nd ed.). John Wiley & Sons.

Vanderbilt University. (2014). *Architectura Sinica. An Interactive Resource for the Study of China's Traditional Architecture*. Www.Architecturasinica.Org. https://architecturasinica.org/index.html

W3C. (2003, October 10). *Resource Description Framework (RDF): Concepts and Abstract Syntax*. Resource Description Framework (RDF): Concepts and Abstract Syntax. https://www.w3.org/TR/2003/WD-rdf-concepts-20031010/#section-simple-data-model

W3C. (2009, August 18). *SKOS Simple Knowledge Organization System Reference* [W3C Recommendation]. Www.W3.Org. https://www.w3.org/TR/skos-reference/

Wajer, M. (2023, February 23). *OCR at the Internet Archive with Tesseract and hOCR* [Developers Manual]. Archive.Org. https://archive.org/developers/ocr.html