Archeologia e Calcolatori
36.2, 2025, 15-22
doi 10.19282/ac.36.2.2025.03

THE NEW SCIENCE OF LONG DATA: PRESENTATION OF THE VENICE LONG DATA PROJECT

1. What is Long Data?

Our shared European history is preserved in an infinite variety of documents stored in historical archives scattered across the continent. These state, municipal, institutional, industrial, and private archives form the backbone of our cultural heritage and serve as the primary source of knowledge about our past. Yet, this collective memory remains largely out of reach-difficult to access and scarcely interconnected. As Italian and European scientists and humanists, our duty is not only to document and safeguard this heritage for future generations but, more importantly, to make it accessible and explorable for all. To achieve this, we must develop a new science of Long Data – a groundbreaking approach to historical sources and archival information that enables both an unprecedented cultural heritage experience and a more integrated capacity for scientific research on this vast, largely uncharted database.

In an era overwhelmed by the rapid production of digital data, our ability to coherently organize Long Data spanning thousands of years is a unique asset. Moreover, historical archives contain highly structured and verified information that encodes the true, multifaceted history and cultural heritage of Italy, Europe, and the Mediterranean. Long Data represents the intangible record of our past, complementing the tangible reality of archaeological sites, monuments, museums, and cities. In this respect, Italy stands out as one of the most – if not the most – Long Data-rich countries in the world. The Long Data approach seeks to establish a foundation for studying history from a macroscopic perspective, complementing the traditional microscopic approach. It always begins with original sources while leveraging the most advanced techniques available for analysing and transcribing historical records.

This emerging science integrates Artificial Intelligence Network Theory and Big Data-while respecting and incorporating traditional archival and historical methodologies-to accomplish tasks once thought impossible. These include transcribing and analysing entire archival series or even interlinking and modelling all the information contained within one or more historical archives. By combining methods from Digital Humanities and quantitative sciences, such as Complex Networks and Economics, Long Data enables a deeper understanding of the historical development of societies, as well as the evolution of language and culture over centuries. There are two key distinctions between Big Data and Long Data that must be kept in mind.

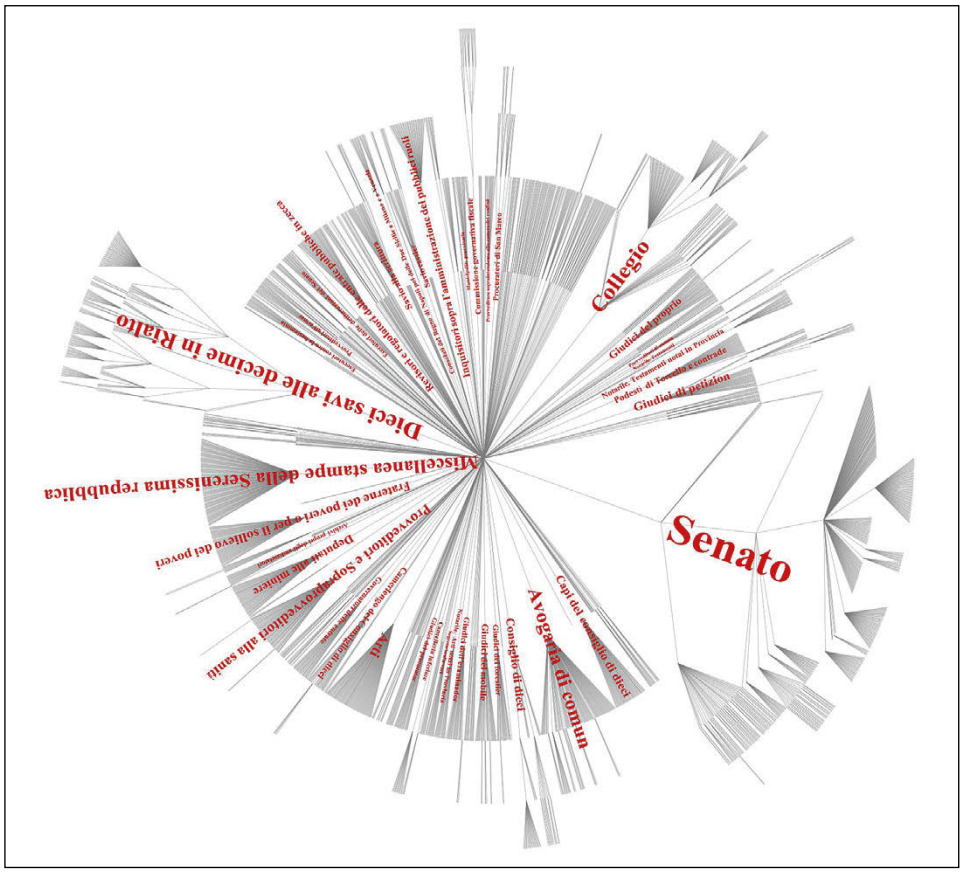

Fig. 1 – Hierarchical structure of Archives. Archives are generally structured in a hierarchical way that reflects the underlying operation and organisation of the institutions that created them (except for archives that have been, a posteriori, completely reorganised according to a different organising principle). Here we show a dendrogram map of the ASVe with over-impressed the names of the major archival fonds (the map shows only those fonds initiated under the rule of the Republic of Venice).

First, Big Data is wide-it consists of vast amounts of information generated daily-but thin, as it spans only a few decades since the advent of the Internet. In contrast, Long Data is relatively small in volume but extremely deep in the temporal dimension, containing records that stretch back over a thousand years and archival series that remain uninterrupted for centuries. Second, Long Data is far more valuable and rare than Big Data. Much of it is encoded in centuries-old manuscripts and concentrated in a limited number of archives. This rarity demands a significantly greater effort in its preservation, management, and analysis. Long Data is processed following rigorous

humanities standards for historical sources and represents a tangible form of cultural heritage in its digitised state. Due to the vast quantity of historical records in our archives, the complexity of transcription, and the constraints of traditional methods, only a small fraction of archival documents has been systematically integrated into our collective knowledge base. For example, in the case of the Senato collection at the Venice State Archive (ASVe), less than 1% of the documents have been transcribed (Fig. 1). This suggests that our understanding of the past may be far less complete than we assume, particularly regarding quantitative historical data, which remains largely unexplored.

The primary scientific goal of Long Data is to bridge this gap, providing historians and researchers across disciplines with comprehensive access to archival sources. By doing so, it enables the construction of a more complete and data-driven picture of our past.

2. What is complex networks theory?

The science of complex networks provides a powerful framework for analysing relationships and structures within vast and intricate datasets. Originally developed in fields such as physics, biology, and social sciences, network theory has proven to be an essential tool for uncovering hidden patterns and correlations in seemingly unstructured information. By representing data as a network of interconnected elements – whether individuals in a society, genes in a biological system, or historical figures in archival records – complex networks allow for a systematic and quantitative approach to studying connectivity and influence across different domains. One of the most significant contributions of complex network science is its ability to measure correlations and define a metrical space within a dataset. By quantifying the strength of connections between elements, it becomes possible to assess distances and similarities in a structured and mathematically rigorous way. In the humanities, this capability is particularly transformative, as it enables the mapping of relationships between historical entities, such as people, institutions, and events, based on archival sources (Fig. 2). This approach goes beyond traditional qualitative analysis, allowing for a quantifiable understanding of historical and cultural dynamics that were previously difficult to measure.

In the context of archival texts, network analysis can reveal latent structures and thematic clusters, providing a new way to interpret historical documents. By modelling texts as networks – where words, concepts, or entities are nodes connected by linguistic or contextual relationships – it is possible to identify patterns of influence, textual similarities, and the evolution of ideas over time. This application of complex networks in the humanities represents a paradigm shift, offering a mathematical foundation for exploring cultural heritage, much like how it has revolutionised other scientific fields.

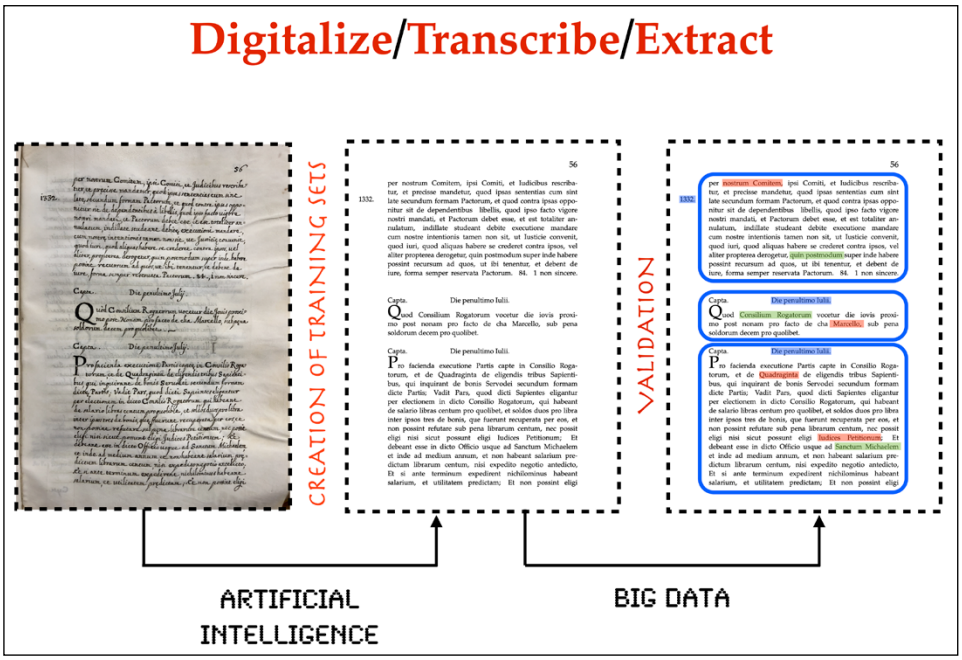

Fig. 2 – Long Data in action. After digitalisation, Artificial Intelligence is first trained on sample pages prepared by expert palaeographers and then used to automatically transcribe all compatible documents, which are successively validated by historians; feature extraction (people and locations in the example) follows document segmentation and dating (Big Data techniques); the resulting data is organised in a temporal multi-layer network that fully represents the information contained in the archival documents considered (Science of Networks approach).

3. Venice Long Data

On a broader scale, Italy is one of the most Long Data-rich countries in the world, with more than fifty state archives holding an invaluable wealth of historical records that remain largely untapped. The Long Data revolution presents a unique opportunity to transform the study of history and cultural heritage, breathing new life into archival research and elevating it to a cutting-edge, data-driven discipline for the 21st century. In particular, Venice, with its vast historical archives, libraries, museums, and private collections, along with its unique role in European and Mediterranean history, is the ideal starting point for the systematic development of the new science of Long Data. Among its many treasures, the Venice State Archive (ASVe) stands out as one of the richest Long Data repositories in the world. Spanning over 80 km of shelves and preserving more than a thousand years of uninterrupted records, ASVe offers an unparalleled depth and continuity of historical data, meticulously

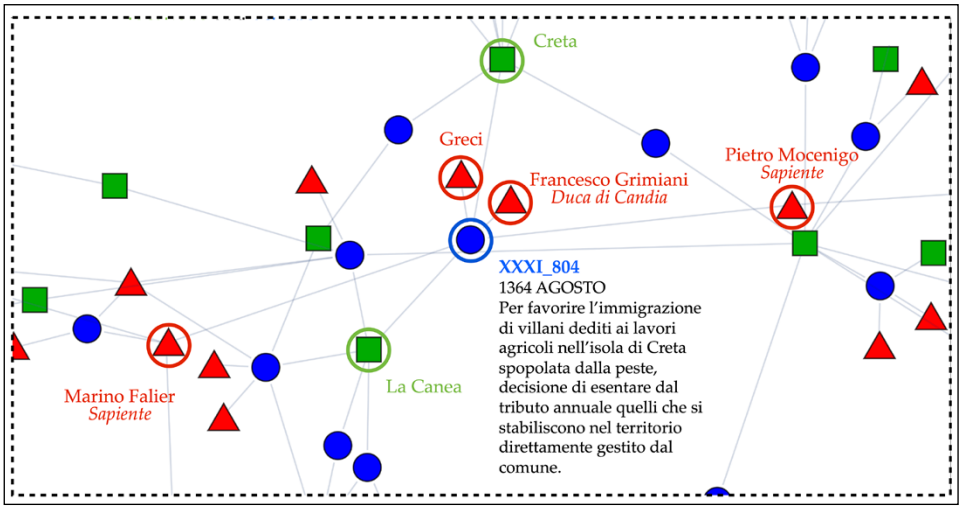

Fig. 3 – Temporal multi-layer network. The information contained in archival documents is naturally described by a temporal multi-layer network linking documents, people, locations and any other relevant quantitative or semantic data; this network structure is the basis for new integrated instruments for scientific studies and cultural experiences. In the Fig. deliberations from the Senato-Misti series (represented as blue dots) are connected to the people (red triangles) and locations (green squares) they (implicitly or explicitly) mention. The 'regesto' of deliberation n. 804 of register XXXI (August 1364) is shown as an example of additional data characterising a node.

conserved over centuries. This unprecedented richness allows us to propose an interdisciplinary research project that approaches historical inquiry from a quantitative perspective.

Rooted in the wealth of historical data provided by ASVe and other archives in the Venetian lagoon, the project integrates methodologies from Artificial Intelligence, Big Data, and Network Science. Its dual objective is to develop a historical meta-database, a fully transcribed, interconnected, and searchable digital version of the Venetian archives, and to establish the foundation for a quantitative study of the history of Venice, Europe, and the Mediterranean (Fig. 3).

4. Times are mature for Long Data

We can now realise the promises of the Long Data revolution because the time is ripe to integrate several cutting-edge technologies in the Digital Humanities, which have been evolving over the past two decades. First, digitisation has become standard practice in most archives, driven by the advancements of the Fourth Industrial Revolution. Second, automatic transcription of historical texts has significantly improved, thanks to the rapid progress in Artificial

Intelligence. Third, Natural Language Processing (NLP) and Semantic Web technologies have matured to the point where they can effectively handle the complexity of centuries-old archival sources. Beyond these advancements, we can also apply quantitative analytical methods from Network Science, which, for instance, propose the use of temporal multi-layer networks as a natural way to represent the rich, interconnected information encoded in historical archives. This approach enables a structural and dynamic understanding of archival data, revealing patterns and relationships that were previously difficult to analyse. The urgent task now is to coherently integrate these technologies to revolutionise how we study and preserve history. The time has come to adopt a quantitative, document-driven approach to our past, to systematically 'back up' our archives for better preservation, and to develop new ways of sharing and interpreting the invaluable stories they contain.

5. Multidisciplinary

Long Data has the potential to empower a new generation of humanists and scientists, fostering an interdisciplinary approach to history where network science, economics, social sciences, and historical studies converge to offer a new perspective on the past. By integrating diverse fields, Long Data provides a framework for collaboration between disciplines that traditionally operated in isolation, opening the door to innovative methodologies and a deeper, data-driven understanding of history. Several key domains contribute to the Long Data approach. Cultural Heritage Studies (Beni Culturali) play a crucial role in the management, preservation, and organisation of archives and artistic sources. Digital Humanities and Archival Sciences oversee digitisation, access, and storage while also setting the standards for digital preservation. Informatics provides the backbone for Artificial Intelligence applications in automatic transcription, as well as techniques to transform language into structured data and manage the databases encoding cultural heritage. History defines and oversees Long Data methodologies, ensuring the accuracy of transcriptions and the validation of metadata and extracted features. Network Science and Economics supply modelling tools and quantitative analysis techniques, while Linguistics enables the study of language and cultural evolution over time. Lastly, Communication Sciences and Information Design play a key role in dissemination and outreach, combining state-of-the-art infographics, compelling storytelling, and modern media to make historical insights widely accessible.

This (necessarily incomplete) list of interdisciplinary synergies demonstrates that Long Data stands as a powerful bridge between the humanities and quantitative sciences. It has the potential to revolutionise the way we approach, analyse, and share our collective cultural heritage, marking a paradigm shift in historical research and public engagement.

6. Outcomes and benefits of Long Data

In conclusion, the benefits and outcomes of an extended Long Data treatment of our historical archives are manifold and affect positively the whole system of cultural heritage as well scientific research in the humanities. The most relevant points are summarised as follows:

– Long Data techniques allow the extraction of all qualitative and quantitative information contained in our archives, including non-numeric language data and implicit information encoded in the structure and interlinking of the archives;
– Long Data offers a way to make a digital backup of our unique sources (just imagine the irreparable cultural heritage loss if fire or flooding affects an archive like the ASVe) and a way to construct online infrastructures for research that go beyond mere photographic digitalisation;
– Long Data allows the reconstruction of missing information by exploiting the inherent redundancy built into archives and ways to overcome the language barrier in the humanities (for example by automatic translation of meta-data and document summaries);
– Long Data stands as a candidate to incarnate the convergence of humanities and quantitative sciences, which will unleash a revolution in how we approach, think and disseminate our past and common cultural heritage;
– Long Data is a solution to the limitations imposed by global pandemics like Covid-19 that hamper scientific and historical research allowing complete online access to archival sources;
– Long Data fosters touristic and creative industries by making accessible to the general public the richness of our cultural heritage through innovative storytelling and engagement;
– Long Data offers lessons from the past that are relevant to our decisions on topics like climate change, pandemics and migrations.

Guido Caldarelli
Istituto dei Sistemi Complessi - CNR
Università Ca' Foscari Venezia
guido.caldarelli@unive.it

Alessandro Codello
Università Ca' Foscari Venezia
alessandro.codello@unive.it

Acknowledgements

The publication was produced with co-funding from the European Union - Next Generation EU - Project ID Code PE_00000020 entitled 'CHANGES - Cultural Heritage Active Innovation for Sustainable Society', CUP H53C22000850006. This manuscript reflects only the authors' views and opinions, neither the European Union

nor the European Commission can be considered responsible for them. The authors would like to thank Raffaele Santoro (former Director of the ASVe) for the invaluable introduction to Archival Science and for his guide through Venetian historical sources and the inspiring environment of the European Center of Living Technology (ECLT), in particular to its Director Prof. Achille Giacometti and to Prof. Raffaella Burioni. This article is a revised version of a note by A. Codello, *The New Science of Long Data*, vol. 01, Venice Material, ARCHiPub, https://www.archive-venice.org/wp-content/uploads/2024/03/archipub_01_004-1.pdf

ABSTRACT

The article introduces the concept of 'Long Data' as an innovative approach to enhancing the cultural heritage preserved in historical archives. This concept distinguishes itself from Big Data by focusing on the deep historical context found in meticulously preserved archives, revealing valuable insights into cultural heritage. By using new Artificial Intelligence technologies in harmony with traditional archival methods, Long Data aims to analyze, transcribe, and model historical data on an unprecedented scale. This approach promises a more comprehensive understanding of history, improving studies on social and cultural evolution. A key example of Long Data's application is the Venice State Archive (ASVe), which holds documents dating back over a millennium. The initiative seeks multidisciplinary collaboration to make this vast archive accessible, thereby safeguarding cultural heritage and paving the way for a revolution in historical research.