Alessandro Codello

The New Science of Long Data

ARCHiPub. On Cultural and Digital Matters
Published by Fondazione Giorgio Cini
01/004

Alessandro Codello is PI of Venice Long Data & adjunct professor at UDELAR Montevideo (Uruguay) & visiting scholar at SUSTECH Shenzhen (China). Before, his post-doctoral research activity was carried manly at CP3-Origins SDU Odense (Denmark) and at SISSA Trieste (Italy), while its PhD is from THEP Mainz (Germany).

WHAT IS LONG DATA?

Our common European history is largely shaped around an infinite variety of documents contained in historical archives scattered all around the continent. These state, municipal, institutional, industrial and private archives constitute the core of our cultural heritage and are the main source of information about our Past. But this common memory is still mostly out of reach, difficult to access and almost completely not interlinked. It is our duty as Italian and European scientist and humanists, not only to keep track and preserve this heritage for future generations, but most importantly to make it accessible and explorable by all citizens. To achieve this goal we need to develop a 'new science of Long Data', a new approach to historical sources and archival information which allows a revolutionary kind of cultural heritage experience, and at the same time, a more integrated ability to perform scientific studies on this immense, mostly uncharted, database.

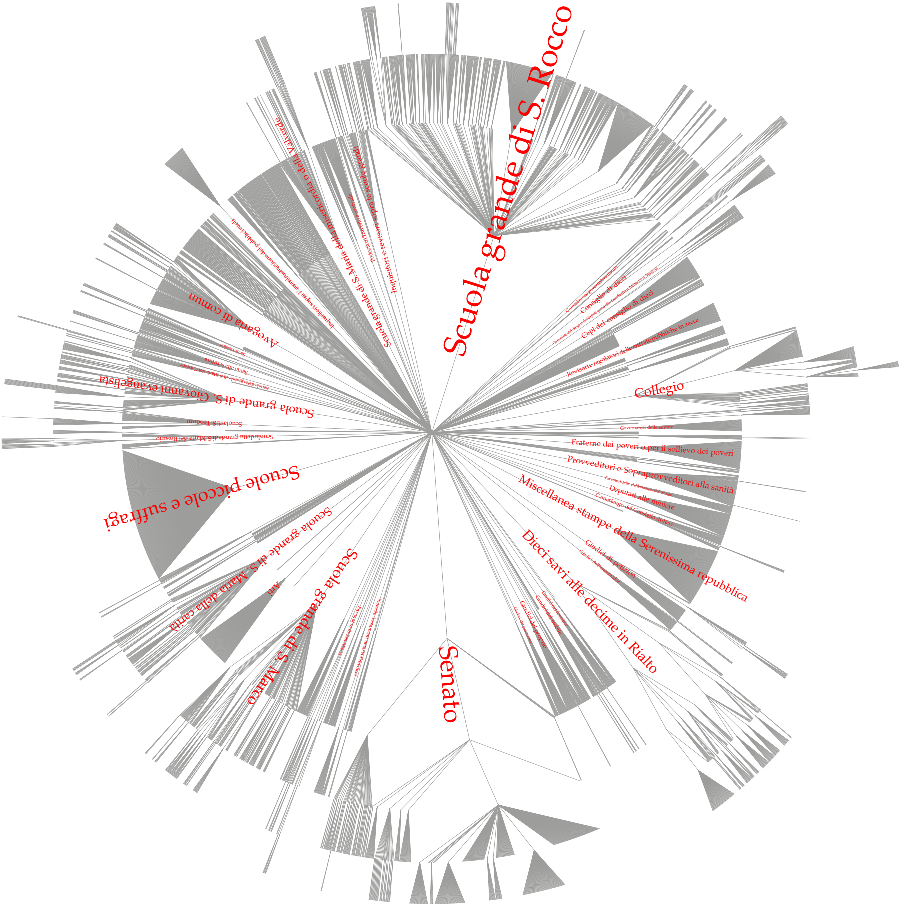
In the middle of a deluge of Big Data[1] production from all over the globe, our capacity, as Europeans and Italians, of having coherently organised Long Data spanning thousands years of history is unique. Moreover, historical archives contain extremely well organised and verified data (as opposed to current Big Data) which encodes the true multifaceted history and cultural heritage of Italy, Europe

[1] Big Data refers to a set of algorithms and techniques used in the analysis of large databases such as those that have become available since the advent of the internet, mobile phones, social media, real time financial markets, etc.

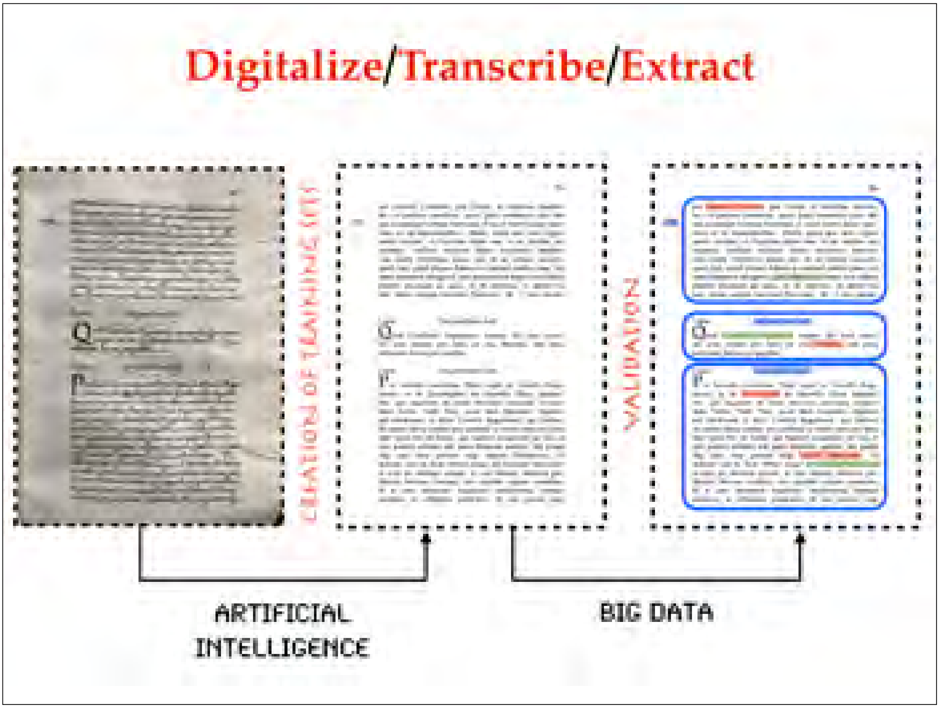
and the Mediterranean. In fact, Long Data is the intangible record of our Past that complements the tangible reality of our archeological sites, monuments, museum and cities. In this respect, Italy stands out as one of the most (if not the most) Long Data rich country in the world.

The Long Data[2] approach aims to create the basis for the study of history from a macroscopic perspective, complementary to the microscopic perspective usually adopted, always starting from the original sources and using the most modern techniques available to deal with the analysis and transcription of historical records. The new science of Long Data utilises Artificial Intelligence and Big Data —but respecting and integrating traditional archival and historical practices— to perform tasks that where before considered impossible, like the transcription and analysis of a whole archival series or even the interlinking and modelling of all the information contained in one or more historical archives. Long Data combines methods from Digital Humanities and quantitative sciences, like Complex Networks and Economics, to model and understand the historical development of our Societies, as well as the evolution of Language and Culture across the centuries.

[2] With Long Data we refer to the study, using adaptations of Big Data techniques, of databases that have a particular temporal depth, such as state archives, and are therefore of historical relevance. Another crucial characteristic of Long Data is the particular attention to the curation and validation of the datasets representing historical sources.


Scuola grande di S. Rocco
Inquisitori e revisori sopra le scuole grandi
Podestà di Torcello e contrade
Scuola grande di S. Maria della misericordia o della Valverde
Inquisitori sopra l'amministrazione dei pubblici ruoli
Avogaria di comun
Savio cassier
Savio alla scrittura
Scuola della grande di S. Maria del Carmine
Scuola grande di S. Giovanni evangelista
Scuola di S. Teodoro
Scuola detta grande di S. Maria del Rosario
Scuole piccole e suffragi
Scuola grande di S. Maria della carità
Arti
Scuola grande di S. Marco
Procuratori di San Marco
Notarile. Testamenti notai in Provincia
Senato
Commissione governativa fiscale
Consiglio di dieci
Consolati del Regno di Napoli poi delle Due Sicilie a Milano e a Venezia
Capi del consiglio di dieci
Revisori e regolatori delle entrate pubbliche in zecca
Collegio
Governatori delle entrate
Fraterne dei poveri o per il sollievo dei poveri
Provveditori e Sopraprovveditori alla sanità
Esecutori delle deliberazioni del Senato
Deputati alle miniere
Camerlengo del Consiglio di dieci
Miscellanea stampe della Serenissima repubblica
Giudici di petizion
Giudici dell'esaminador
Dieci savi alle decime in Rialto
Giudici del forestier
Giudici del mobile
Giudici del proprio
Giudici del procurator

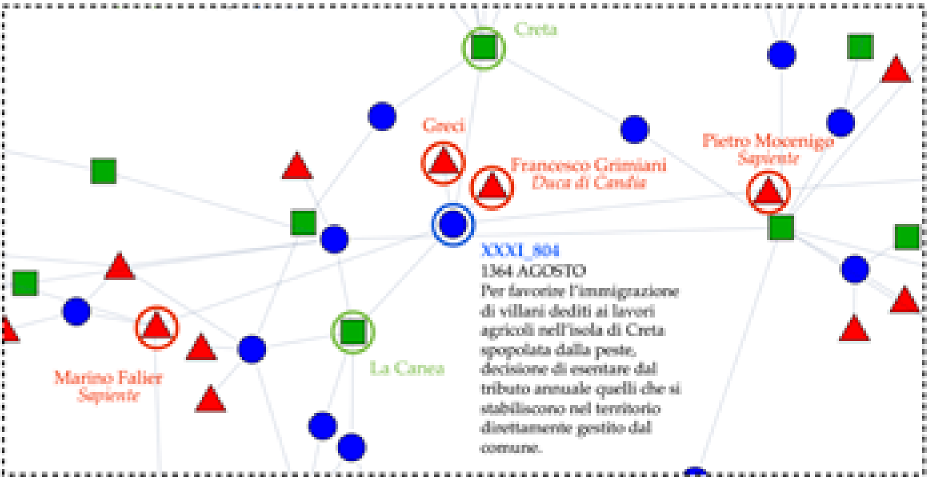
Mapping the structure of Archives
Archives are generally structured in a hierarchical way that reflects the underlying operation and organisation of the institutions that created them (apart those archives that have been, a posteriori, completely restructured according to a different organising principle). Here we show a dendrogram map of the ASVe with over-impressed the names of the major archival fonds.



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 represent the information contained in the archival documents considered (Science of Networks approach).

There are two fundamental distinctions between Big Data and Long Data to keep in mind: firstly, the former is wide (a lot of data collected daily) but thin (only few decades of records available since the advent of the Internet) while the latter is relatively small in terms of quantity but extremely deep and long in the time dimension (oldest data is more than a thousand years old and archival series stretch uninterrupted for centuries). Secondly, Long Data is much more precious than Big Data (it's mostly encoded in century old manuscripts) and rare (in greater part it is concentrated in very few archives); for this reason much more effort is devoted to its management. Long Data is processed following the standards of the humanities for historical sources and represents concrete cultural heritage in digitised form.

Due to the sheer size of the information present in our archives, and to the complexity of the transcription process as well as to the limitations imposed by traditional methods, only a small fraction of archival documents has been systematically integrated in our knowledge base (for example in the case of the Senato fond at the Venice State Archive (ASVe) only less than 1% of the documents have been transcribed). These order-of-magnitude considerations indicate that probably 'we do not know the past as well as we think' since the vast majority of historical data (in particular quantitative data) has yet to be studied. The main scientific goal of Long Data is exactly to fill this gap and furnish to historians and researchers from all fields a comprehensive



XXXI_804
1364 AGOSTO
Per favorire l'immigrazione di villani dediti ai lavori agricoli nell'isola di Creta spopolata dalla peste, decisione di esentare dal tributo annuale quelli che si stabiliscono nel territorio direttamente gestito dal comune.

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 figure 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 (dating August 1364) is shown as an example of additional data characterising a node.

access to archival sources upon which construct a complete and quantitative picture of our Past.

VENICE LONG DATA

Venice, with its many historical archives, libraries, museums and private collections, and due to its unique place in European and Mediterranean history, is the natural place where to start the systematic development of the new science of Long Data. In particular, the Venice State Archive (ASVe) stands out, for the depth and length of the data it contains (over 80 kms of shelves and over one thousand years of uninterrupted collections) and for the quality of the conservation, as one of the most Long Data rich places in the entire World.

Venice Long Data is an interdisciplinary research project that aims to address historical research from a quantitative point of view, relying, in the Venetian context, on the historical databases offered by the ASVe and the other archives situated in the lagoon, using techniques from Artificial Intelligence, Big Data and the Science of Networks. The dual scope of Venice Long Data is to create a historical meta-database representing a fully transcribed, interconnected and searchable digital version of the venetian archives; and to create the basis for a quantitative study of the history of Venice, Europe and the Mediterranean.

On a greater scale, 'Italy is a Long Data rich country'. Its complex of more than fifty state archives is an inestimable source of Long Data that has, to date, only been superficially scratched. The Long Data revolution is an opportunity to ignite a renaissances in the way we approach this immense cultural heritage and a means to rejuvenate archival and historical research into a full fledged 21st century science.

TIMES ARE MATURE FOR LONG DATA

We can achieve the goals promised by the Long Data revolution because times are mature to combine several key state of the art technologies in the Digital Humanities that have been developing in the last two decades. First of all, digitalisation is now fully employed thanks to the advancements made possible by fourth industrial revolution and is common practice in most archives; second, the advancements of automatic transcription of ancient texts has seen manifold improvement due to the world wide advancement of Artificial Intelligence technologies; third, Natural Language Processing and Semantic Web technologies have became mature enough to be applied to complex language data as those of century long archival sources. In addition, we can apply quantitative methods of analysis steaming from the Science of Networks, which, for example, suggest the use of temporal multi-layer networks as a natural way to represent the information encoded in an historical archive. The effort to coherently combine all these advancements has now become urgent: times have come to start studying our Past in a quantitative and “document driven way”; to systematically “back-up” our archives to better preserve them; and to create new ways to disseminate the stories they contain.

MULTIDISCIPLINARY

Long Data has the potential to empower a new generation of humanist and scientists. It can set the stage for an interdisciplinary study of history where Network, Economical, Social and Historical sciences join forces to give us a new perspective on our Past. Many fields are integrated by the Long Data approach: Beni Culturali for the management and preservation of cultural heritage as well as the organisation of archives; Digital Humanities and Archival Sciences cover the process of digitalisation, access and storage of archival and artistic sources, as well as the fundamental role of setting the standards of digital preservation; Informatics for the application of Artificial Intelligence to automatic transcription and for the extraction techniques needed to convert language data into structured data, as well as for the more standard role of creating and maintaining the databases encoding cultural heritage; History to define and overview the general methodologies of Long Data, to validate the transcription process and to certify feature and meta-data extraction; the Science of Networks and Economics to furnish modelling tools and to perform quantitative analysis; Linguistics for the study of language and cultural evolution; last but not least, the role of Communication Sciences and Information Design for dissemination and outreach, combining state-of-the-art infographics with an engaging storytelling and modern media. This (clearly incomplete) list of complimentary synergies shows that Long Data stands as a candidate to incarnate the convergence of humanities and quantitative sciences, with the potential to unleash a revolution in how we approach, think and disseminate our common cultural heritage.

OUTCOMES AND BENEFITS OF LONG DATA

In conclusion, the benefits and outcomes of an extended Long Data treatment of our historical archives is manifold and affects 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 tourist 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.

It’s definitely time to start the Long Data revolution and make our rich Past a driving force for an even more prosperous Future.

ACKNOWLEDGMENTS

The author 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 it's Director Prof. Achille Giacometti and to Prof. Guido Caldarelli and Prof. Raffaella Burioni.

ARCHiPub
On Cultural
and Digital Matters

EDITOR IN CHIEF
Chiara Casarin

MANAGING EDITOR
Claudio Piscopo

FOCUS RESEARCH 01
Venice Material

VOL. 004
The New Science of Long Data

CONTRIBUTOR
Alessandro Codello

Publisher
Fondazione Giorgio Cini Onlus
Island of San Giorgio Maggiore, Venice

ARCHiVe
Analysis and Recording
of Cultural Heritage in Venice

ARCHiVe
Analysis and Recording
of Cultural Heritage in Venice

is a project by
Fondazione Giorgio Cini
Factum Foundation
EPFL – Digital Humanities Laboratory

Supporting funder
The Helen Hamlyn Trust

ARCHiPub. On Cultural and Digital Matters is an interdisciplinary book series that gathers research on topics such as archival studies, digitisation projects, and cultural heritage conservation. Each volume focuses on a research theme, to be explored by authors from different academic backgrounds.

This contribution is part of the focus research *Venice Material*. Venice Material is the starting point. Venice as a city, as an environment in which history has formed the present civilization and as a fertile humus of ever-new cultural sap. Venice as a bridge between worlds that were sometimes created, sometimes destroyed and still a bridge between ways of producing culture as a primary good. Venice is made out of matter, stone, painting, poetry, a rich and sensitive work, layered materially and immaterially like no other city in the world. Venice as a launching pad for new experimental horizons, as a landing place for new generations of scientists and creative people. But Venice is also considered as a subject of study, a focus of scientific and humanistic research, endowed with the persuasive force of authentic insights that seem to multiply rather than run out. Venice as an object to be investigated, disassembled and reconstructed, digitized and disseminated, curated.

fondazione Giorgio Cini onlus