Alma Mater Studiorum · Università di Bologna

DOTTORATO DI RICERCA IN
SCIENCE, COGNITION AND TECHNOLOGY
Ciclo 29

Settore Concorsuale di afferenza: SC 11/C2
Settore Scientifico disciplinare: SSD M-STO/05

THE WEB AS A HISTORICAL CORPUS

Collecting, Analysing and Selecting Sources On the Recent Past Of Academic Institutions

Presentata da: Federico Nanni

Coordinatore Dottorato
Prof. Marco Beretta

Relatore
Prof. Maurizio Matteuzzi

Correlatore
Prof. Simone Paolo Ponzetto

Esame finale anno 2017

A Giulia,
compagna di caffè la mattina,
di birre la sera, di vita in between.

Contents

List of Figures

List of Tables

Introduction: A New Kind of Primary Source

XXI Century is a digital book.[1]

The advent, adoption and widespread diffusion of the Internet[2] are having a global impact on our society of a kind that we have never experienced before. In the last thirty years, Internet and then the World Wide Web have caused radical transformations to several aspects of our everyday life: from communication to work, from politics to economy, from civil rights to the way we interact with media contents. Researchers have already compared what happened during the first decades of the "digital age" to the consequences of the invention of the printing press (Assmann, 2006), they have described it as the third industrial revolution (Rifkin, 2011) and studied its influence on a number of topics, from the changing in higher education (Pittinsky, 2003) to the improvement of world-wide commercial strategies (Pitt et al., 1999).

As a young historian who has conducted his entire studies across the first two decades of the XXI Century, many times I[3] have wondered how the way we acquire knowledge about the past will change when born-digital resources

[1] Each chapter of this thesis is introduced by an epigraph related to its main theme. The author has been intentionally omitted, leaving the text to speak for itself.

[2] All URLs last checked on 10th of March 2017. If a resource is offline, verify its availability on the Internet Archive: `https://archive.org/`

[3] In this thesis, I use the first person singular when describing its main argumentation and the way it is structured across the different chapters, i.e. in the introduction, in the

(i.e. materials that have been created and – in the overall majority of the case only exist – in digital format) will effectively become our primary "historical evidence"; however, I noticed how such an in-depth analysis on the future of the historical method is currently missing in the historiographic literature.

In fact, while in recent years a long discussion on the potential of using computational methods for interpreting sources in a "new" way has been the focus of innumerable conferences, panels, workshops and seminars, the historical community in general, and digital humanities (DH) scholars in particular, still have not dedicated enough attention to the impact that the transition from analogue and digitised sources to born-digital evidence will have on historian's craft.

The web today represents the largest collection of human testimonies that we have ever had at our disposal. While in recent years it has been remarked several times (often quite superficially[4]) that the overall majority of these materials will not be "useful" sources for historical studies, my argumentation starts in strong opposition to this shared attitude, by recalling a teaching from Marc Bloch:

> The variety of historical evidence is nearly infinite. Everything that man says or writes, everything that he makes, everything he touches can or ought to teach us about him. (Bloch, 1949)

Bloch, is his "total" approach to history, was highlighting the relevance of non-traditional sources (such as memoirs) to study the lives of non-traditional historical subjects, such as common people instead of kings and political leaders. However, these words also stress that it is the duty of the historian to respect and understand how to deal with any new kind of primary source. In the near future websites, digital library collections, blogs, tweets, emails, discussion-threads in forums, online news, web archive materials, edit histories of Wikipedia articles, large-scale knowledge bases will become the huge

conclusion and in the short overviews at the beginning of each main chapter. The body of the thesis is presented using the first person plural.

[4] See for example Chaitin (2016).

majority of the traces left by our recent past, while the same time their analogue counterparts, from printed news articles to personal diaries, from letter correspondences to scientific publications on paper, will be less and less produced.

Five hundred years ago, scholars have already experienced a revolutionary methodological change influenced by the advent of a new type of textual primary source. Eisenstein (1980) remarked in her book that without the invention of the printing press the ideas of the Renaissance would not have taken hold. As Sharfman (2015) recently pointed out, regarding its impact on source criticism:

> The technological shift to print culture created "typographical fixity" while enabled classical texts to penetrate cultural thought. The power of the printing press to make abundant copies of classical texts ensured that these texts would become more available, and never lost, as had been one of the limitations of manuscripts. Additionally, scholars could now compare printed copies of these texts, correct the errors and create a permanent and correct version of the classical text. These texts could be abundantly distributed because of the printing press, and scholars could study these texts and use them to further develop and spread humanist ideas. Thus, without the printing press, the shift to humanist thinking would have flickered out just like earlier classical revivals. (Sharfman, 2015).

At the beginning of the XXI Century, historians are, once again, dealing with a new type of document, which is rapidly overcoming so-far traditional primary sources and - I argue in this thesis - will completely change the way these testimonies of the past are approached and studied.

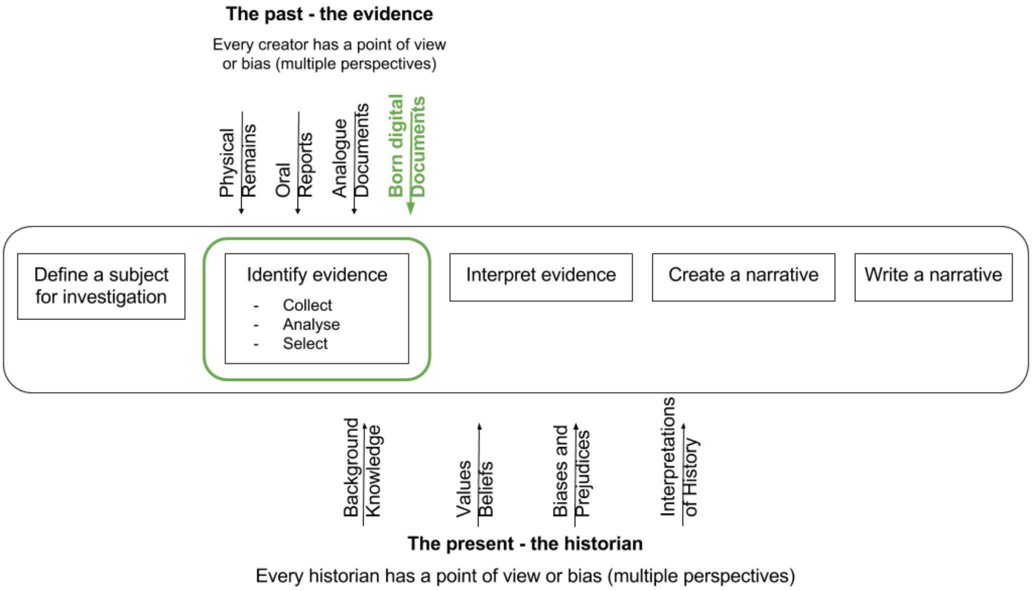


Figure 1: Graphical representation of the historical method.

The Effects on the Historical Method

The way historians conduct their work is generally described with the umbrella-concept of the "historical method" (Shafer, 1974). This approach – necessary starting point of any university course on history – could be exemplified by dividing it in five major steps (see Figure 1), which are highly interconnected and always influence each other.

Defining a Subject. In the first step, the scholar defines a subject of investigation and - together with it - an initial research question. This question could be "Is there a connection between Calvinism and Capitalism?" as well as "What has been the role of the Mediterranean Sea in shaping societies?" or even "Is the Orient a construct of the West?". The research question, initially presented at a coarse-grained level, will be sharpened and sharpened through the recursive repeating of the methodology here presented.

Identifying the Evidence. In order to dig deeper into the subject, the historian identifies sources through a complex process of collection, analysis and selection of the remains of the past. These traces could be physical remains (e.g. buildings, statues), oral memories, printed documents (e.g. chronicles, diaries, articles, census data) and - as I will remark in this thesis - will soon become born-digital documents, such as websites, forums, databases, etc.

The process of identifying primary sources has been shaped by decades and decades of discussions in historiography both on how to establish their reliability through source criticism (e.g. by conducting philological analyses) and on how much "true knowledge" we can derive from them (there are as many interpretations of the same text as many readers, Barthes (1967) taught us).

Interpreting the Evidence. This step represents the core of any historical research and (together with creating a narrative) the central focus of XXth Century debates in historiography. Interpreting sources helps sharpening the research question and could be done in many different ways: as I will present in the first chapter of this thesis, traditional historiographic approaches have been strongly based on close reading interpretation of textual documents, while other solutions - which emerged during the XXth Century - have been based on statistical analyses of census data or economic reports. The advent of post-modern and deconstructionist theories (Barthes, 1967; Derrida, 1967; Lyotard, 1984) has posed a major critique against the underlying assumption supporting the historical method, namely the possibility of discovering a unique "truth" about the past via the careful interpretation of sources. The consequences of this critique over historiography, which have been remarked by many cultural historians (Munslow, 2006; Burke, 2008), will be presented in the first chapter of this thesis.

In the last twenty years, while still dealing with the effect of the so-called cultural turn, the community has seen the advent of a new set of approaches for interpreting sources. These solutions come from the fields of natural language processing (NLP) and have been labeled with the digital humanities concept of *distant reading* (to know more see Chapter 2). These methods have been applied in historical research to analyse and extract quantitative evidence from large scale collections of digitised data.

Creating and Writing a Narrative. The final steps of the historical research sees the historian defining a narrative and finally writing a history. The creation of a narrative, which is highly connected with the initial definition of the research question, gives to the historian the possibility of placing the history he/she is writing as part of a larger contribution to the field. This is achieved in two interconnected ways: first of all, by offering a new/different perspective on the topic under study; in addition to this, by participating to the larger historiographic debate regarding how the past can be re-discovered, examined, described and - for certain authors (see for example the advent of Cliometrics (Greif, 1997)) - even modelled.

In this thesis, while I will argue that the advent of born-digital sources will affect all these steps of the historical method, I will primarily focus on what I consider to be the most critical change, namely the way new sources will be collected, analysed and selected.

Research Question

As it has been already presented (and will be expanded in the next chapters), historians will soon find themselves forced to deal with born-digital documents as these materials will become the overall majority of primary evidence left by our recent past. The main research question of this dissertation derives directly from this premise: what will change in the practices that we have adopted so far in order to identify, establish the reliability, analyse and select primary evidence, in a scenario where these materials will exist only in digital format?

Among the few authors that in recent years have begun to consider web materials as historical sources, in this work I start my argumentation by examining three specific publications, written by Rosenzweig (2003), Brügger (2009) and Milligan (2012). These different works highlight the consequences for the profession of the historian and envision necessary changes in his/her methodology, in particular for what concerns the processes of collecting, analysing and selecting evidence. More specifically, these articles put special emphasis on the two following key aspects, which are current obstacles for a researcher that intends to employ born-digital materials as primary sources.

Scarcity. On the one hand, the researchers remarked that born-digital materials have an extremely shorter life compared to traditional analogue sources as they are way more difficult to archive and preserve (LaFrance, 2015)[5]. This is due to a vast number of reasons (Brügger, 2005) and the consequence of it has been summarised by Rosenzweig with the concept of "scarcity" of digital primary sources.

While international web archive initiatives have a long tradition of preserving born-digital materials for future research (Gomes et al., 2011) (as described in the third chapter of this thesis), several issues still exist and new issues continue to emerge - due to constant innovations in web technologies. Therefore, researchers have to deal with the collected materials in a highly critical way, as Brügger (2012b) described when he introduced his definition of web archive documents as reborn-digital materials:

> One of the main characteristics of web archiving is that the process of archiving itself may change what is archived, thus creating something that is not necessarily identical to what was once online.
>
> [...] And, second, that a website may be updated during the process of archiving, just as technical problems may occur whereby web elements which were initially online are not archived. Thus, it can be argued that the process of archiving creates the archived web on the basis of what was once online: the born-digital web material is reborn in the archive. (Brügger, 2012c)

Abundance. The difficulties in the preservation of digital sources present a new series of issues for historians; however, they remain only part of the overall problem. As a matter of fact, already in 2003, Rosenzweig envisioned

[5] See also the extended discussion regarding the digital dark age (Kunny, 1997).

that future historians will not only deal with a consistent scarcity of primary sources, but they will be also challenged by a never experienced before abundance of records of our past[6]. The indispensable need of computational methods for processing and retrieving materials from these huge collections of primary sources has been a central topic of Milligan's publication. From his work it emerges that, while until today adopting or not computational approaches has always been a choice for the humanities researcher (in a constant comparison between quantitative and qualitative, hermeneutics and data analysis), now that the community is dealing with this abundance of sources that exists only in digital format, the use of these solutions (starting from information retrieval search tools) is not a choice anymore.

Due to this fact, it becomes essential that the researcher adopts these solutions critically, always knowing their potential and biases, and learns how to combine them fruitfully with the traditional historical method, as Milligan specifically remarked:

> Digital sources necessitate a rethinking of the historian's toolkit. Basic training for new historians requires familiarity with new methodological tools and making resources for the acquisition of these tools available to their mentors and teachers.
> [...] This form of history will not replace earlier methodologies, but instead play a meaningful collaborative and supportive role. It will be a form of distant reading, encompassing thousands, or tens of thousands, of sources that will complement the more traditional and critically important close reading of small batches of sources that characterizes so much of our work. (Milligan, 2012)

The argument of my Ph.D. work starts and strongly builds upon the pioneering theoretical works just presented. Through a series of case-studies and with a highly digital humanities approach, I intend to critically assess how methodologies from other fields of study (especially internet studies and

[6] See for example the event-collections offered by the Internet Archive: `https://archive-it.org/organizations/89`

natural language processing) can be properly combined with the traditional historical method in order to support researchers in collecting, analysing and selecting born-digital materials. These approaches will offer the possibility of effectively address established research questions and to start imagining many new ones.

Case Study

The Ph.D. research presented in this thesis has been conducted at the International Centre for the History of Science and Universities (University of Bologna). This research center, in its over 25 years of existence, has focused on the study of higher education as well as on science and technology research topics. In order to show how the advent of born-digital sources will have an impact on the way established history scholarships collect, examine the reliability and select primary sources, the case study of my work is on a research topic that my group has been studying since the beginning of the 90s, namely the past of academic institutions.

Studying the history of academic institutions and understanding the role and bidirectional influence of universities on our society has already attracted the attention of different research communities, which have addressed this topic with various approaches and final goals.

Historians of Universities. The massive four-volume book series, directed by the European University Association, edited by Hilde de Ridder-Symoens and Walter Rüegg and published between 1992 and 2011, offers an unprecedented overview of how European universities have changed during the last centuries. In these works, researchers adopt a large variety of primary and secondary sources, from close reading of university-archive materials to the interpretation of the results of large-scale analyses on – for example – admission statistics.

Historians of Science and Technology. History of academic institutions have also been studied by historians of science and technology, interested in understanding how scientific knowledge has moved back and forth between universities and the private sector and how political, economical and social actors have influenced research in academia (see Pancaldi (1993a)). Researchers often adopt a combination of methodologies: from close reading of primary sources to ethnography and anthropology practices.

Scientometricians. A third diachronic perspective on universities and the changes in their research output has been offered by the scientometrics community (Van Raan, 1997), whose goal is to establish quantitative measures to evaluate the research outputs of these institutions. In addition to traditional bibliometric measures, more recently, a series of publications have focused on the use of word-based and topic-based approaches (see Lu and Wolfram, 2012), in order to expand the type of materials that could be analysed.

As it will be highlighted in the final chapter of the first part of this thesis, while the scientometrics and the science and technologies communities are already dealing with the difficulties of collecting and analysing born-digital sources, this thesis foresees a more long term impact of these materials over the field of history of universities. In this thesis I will argue that born-digital sources such as syllabi (Cohen, 2011), bachelor, master and doctoral theses (Ramage, 2011), academic websites (Holzmann et al., 2016b) and their hyperlinked structure (Hale et al., 2014) will become relevant new materials that historians of academic institutions will employ as primary sources to continue study this topic. For this reason, today an analysis on the ways these sources could be collected, analysed and selected is strongly needed.

Contribution and Structure of the Thesis

This dissertation is divided in three parts. In **Part I**, I will highlight the new challenges that born-digital materials present to the historical method, especially for what concerns the processes of collecting, analysing and selecting them (see Figure 1). Based on this analysis, I will argue that historiography is about to face a drastic transition in its theories and methods, which is precisely due to the born-digital nature of contemporary primary sources, not only to the advent of computational methods to the craft (as other researchers such as Graham et al. (2016), Nelson (2016) and Scheinfeldt (2016) have recently remarked).

To support my thesis, in the two central parts of this dissertation I will present a series of case-studies showing many of the new challenges that emerge when studying the recent past of academic institutions adopting born-digital sources. In addition to this, I will describe how I extended the traditional historical method with approaches from the fields of internet studies and natural language processing, precisely for facing these issues.

In particular, in **Part II** of this thesis, I will start by addressing the so-called scarcity issue by considering university websites as primary sources for the study of the recent past of academic institutions. As a first step, I will examine the reliability of archived versions of these websites, preserved by international web archive institutions. Next, I will introduce an iconic example regarding the ephemerality of born-digital sources: the University of Bologna website had been excluded by the Internet Archive and has not been preserved by any of the teams that have managed it during the years. An issue like the one presented in this thesis could deprive future researchers of all digital documents published on the web so far by this institution. By combining traditional sources and methods (collecting archived materials, conducting interviews etc.) with solutions from the field of internet studies (web archive search, hyperlink analysis), I present how I reconstructed the digital past of this institution and recollected all these born-digital materials.

The collected resources allowed me to address the second issue presented above, namely the large abundance of born-digital sources. In **Part III**, I will specifically focus on collecting, analysing and selecting materials from large collections of academic publications. In particular, I will remark on the importance of adopting methods from the field of natural language processing in a highly critical way. I will argue on the fact that these approaches are the only solution at disposal when dealing with the vastness of collections that exist only in digital format; for this reason, it is essential that researchers will use them carefully in order to not misinterpret their outputs. I will stress this point by presenting a case-study focused on identifying interdisciplinary collaborations through the analysis of a corpus of Ph.D. dissertations and by showing how one of the most adopted computational methods in digital humanities yields to very low-quality results compared to other less-popular approaches.

Based on the case-studies here presented, this thesis intends to be a contribution to two different communities. First of all, this research places itself in the digital humanities environment, as it is focused on understanding how the research topics, concepts and methodologies of a community change when a digital element is involved. In addition to this, this thesis aims to be my initial contribution to the current debate on the present and future of historiography (Evans, 1997; Cohen et al., 2008; Graham et al., 2016). I aim to do so, by discussing how the process of identifying sources change in the digital age and by opening a discussion on how this fact will soon influence how we theorise, teach and practice historical research.

This dissertation has been conceived, planned and initially conducted at the International Centre for the History of Universities and Science, a research group whose origins are strongly connected to the last Centenary of the University of Bologna. These roots are also present in this dissertation, starting from the fact that the specific case studies will be Italian universities, with special attention on the recent past of the University of Bologna.
This work also benefits from the highly interdisciplinary perspective I matured during the visiting periods I conducted at the Centre for Internet Studies (Aarhus University), at the Human Language Technologies Group of the Foundation Bruno Kessler (Trento), at the Data and Web Science Group (University of Mannheim) and at the Department of Computer Science (University of New Hampshire). Spending three years learning, collaborating and

discussing everyday with internet studies scholars as well as computer scientists helped me shaping my curiosity on the future of the historical method like nothing else.

Main Publications

I list here the peer-reviewed contributions on which this thesis has been based. They are grouped by their main topic and – in case of multiple first authors – they have been marked with an asterisk (*).

Concerning How to Deal with the Scarcity of Sources

Nanni, F. (2015). Historical Method and Born-Digital Primary Sources: A Case Study of Italian University Websites. *Officina della Storia*, special issue on "From the History of the Media to the Media as Sources of History".

Chakraborty, A.* and **Nanni, F.*** (2017). The Changing Digital Faces of Science Museums: A Diachronic Analysis of Museum Websites, In Brügger N. (eds) *Web 25: Histories From the First 25 Years of the World Wide Web.* Peter Lang.

Nanni, F. (forthcoming). Reconstructing a Website's Lost Past - Methodological Issues Concerning the History of www.unibo.it. *Digital Humanities Quarterly.*

Concerning How to Deal with the Abundance of Sources

Nanni, F. (2014). Managing Educational Information on University Websites: A Proposal for Unibo.it. In *2nd Annual Conference of the Italian Association for Digital Humanities (Aiucd 2013).* CLEUP.

Nanni, F., Dietz, L., Faralli, S., Glavaš, G., & Ponzetto, S. P. (2016). Capturing Interdisciplinarity in Academic Abstracts. *D-Lib Magazine*, 22(9/10).

Lauscher, A.*, **Nanni, F.***, Fabo, P. R., & Ponzetto, S. P. (2016). Entities as Topic Labels: Combining Entity Linking and Labeled LDA to Improve Topic Interpretability and Evaluability. *Italian Journal of Computational Linguistics*, special issue on "Digital Humanities and Computational Linguistics".

Concerning the Future of the Historical Method

Bonfiglioli, R.*, & **Nanni, F.*** (2016). From Close to Distant and Back: How to Read with the Help of Machines. In: Gadducci F., Tavosanis M. (eds) *History and Philosophy of Computing (HaPoC) 2015.* IFIP Advances in Information and Communication Technology, vol 487. Springer, Cham.

Nanni, F., Kümper, H., & Ponzetto, S. P. (2016). Semi-Supervised Textual analysis and Historical Research Helping Each Other: Some Thoughts and Observations. *International Journal of Humanities and Arts Computing*, 10(1), special issue on 'The Future of Digital Methods for Complex Datasets".

Part I

The Web as a Source

Chapter 1

Historiography in the 20th Century

The man who would cross an
ocean to verify a comma.

This first chapter presents a general overview of the central debate in XXth century historiography, namely to what extent history, as a discipline, can accurately acquire knowledge on the past. This will give me the possibility of introducing, in the following chapter, how the use of computational methods in historical scholarship has been depicted by many as the triggering factor of a new turning point in historiography.

1.1 Introduction

Examining the ways historians acquire knowledge about the past (in which way do they define and shape their research questions? how do they collect and establish the reliability of their sources? how do they interpret the selected materials?) is as challenging as studying the past itself. In the last fifty years several books have been published by relevant historians with the intention of describing, understanding and discussing different ways scholars have practiced the craft. In the next sections, we intend to start our research by offering a general overview of the most relevant debate of XXth Century historiography, which is focused on establishing (or denying) the possibility of discovering a single objective true narrative about the past with the adoption of precise methodological practices (often based on the use of social science theories and approaches). This introduction will allow us to present - in the next chapter - how the advent of computational methods has been perceived by part of the historical community as the agent of a new turning point in this debate (for example, it has been depicted as the begining of a new wave of quantitative history (Gibbs and Owens, 2012)). Next, we will argue - instead - that this "revolutionary" agent of change in historical scholarship will be the advent of a new type of document, as it will require that source and method criticism will regain centrality in the craft.

In order to offer an overview of this discussion, in this chapter we consider the positions expressed by the following authors. First of all, we start with Edward Hallett Carr's "What is History?" (Carr, 1961) and his argument on the fact that historical narrative cannot simply be reduced as an objective collection of facts. Sir Geoffrey Elton's reply to Carr's opinion, "The Practice of History" (Elton, 1967), which remarks that history's goal is searching for the objective truth, is the second text we examined. Following these authors, other important contributions to the discussion have been taken into consideration, namely "That Noble Dream" by Novick (1988), Iggers's " Historiography in the Twentieth Century" (Iggers, 2005) and Evans' "In Defence of History" (Evans, 1997). On the other side of the spectrum, critiques to the scientific validity of historical interpretation have been examined in authors such as Gaddis (2002) with "The Landscape of History: How Historians Map the Past", Burke (2008) with "What is Cultural History?" and Munslow (2006) with "Deconstructing History". Finally, we consider the work by Bloch (1949), "The Historian's Craft", as the representative example of a previous turning point in our discipline (i.e. the advent of "total history"), while the writings of David Harvey in "The Condition of Postmodernity" (Harvey, 1989) and Latour in "We Have Never Been Modern" (Latour, 1991) offer us a wider perspective on the debate.

1.2 History as a Social Science

While the debate on whether history can be considered a social science has attracted the interest of many historians and philosophers, the adoption of social science methodologies in historical research can be precisely tracked back to the professionalisation of history as an academic discipline taught at university level (Iggers, 2005). The work of Leopold von Ranke (1795 - 1886) in Germany, focused on the critical examination of sources through a combination of practices from philology and guided by the purpose of understanding the past through the study of precise facts in order to present to the reader "what really happened" (*wie es eigentlich gewesen*), helped history to become an independent discipline, different from philosophy and literature. In the following decades, the revolutionary work that Max Weber (1864 - 1920) conducted contributed to give precise identity to the field of sociology and had a huge impact on historical practices. In particular his book, "The Protestant Ethic and the Spirit of Capitalism" (Weber, 1905), showed how the "understanding" of the past could be represented as a highly rational process. In the same decades Marxist historiography grew as well, strongly inspired by the works of Friedrich Engels (Engels, 1850) and Karl Marx (Marx, 1867). This school of historiography combined social science practices from economics and sociology with a ideological narrative focused on the centrality of social class and economic constraints in determining historical outcomes.

While the two Wold Wars questioned the highly shared idea of human history as a progressive process, where our society learns from the mistakes made in the past precisely thanks to the objective and methodological way the past is studied and understood, in the following decades a second strong wave of belief in historical objectivity driven by social science practices raised. Its traces could be noticed by reading a famous issue of the Times Literary Supplement (1966) (as remarked by (Evans, 1997)) where young historians combined their work with practices from anthropology, social theory and statistics. A strong adoption of social science practices could also be noticed when examining the works conducted in France by the Annales school, where researchers placed history at the centre of an interdisciplinary network together with geography, economics and sociology (to know more read Bloch (1949), or examine the approaches adopted by scholars such as Braudel (1972) in works like "The Mediterranean and the Mediterranean world in the age of Philip II").

The apex of this period of historiography was reached in the United States between the 60s and the 70s with the rise of social science history thanks to economic historians such as Robert Fogel, who received (together with Douglass North) the Nobel prize in Economics for his studies[1]. By combining the use of complex statistical models, the computational power of early computers and with the support of the large interdisciplinary teams of researchers that these scholars brought together, these studies aimed at describing the past through the identification of general laws, which could allow modelling of relations between events, structures and processes. These works were often sustained by a rhetoric which remarked the advantages of social science history (or "cliometrics") as opposed to traditional historical scholarships.
The publication of "Time of the Cross" (1974), a two-volume study on slavery in the old American south by Fogel and Engerman condenses all the pros and cons of cliometrics. It is first of all a huge study of a relevant historical topic; it is sustained by a large variety of evidence conveyed through graphs, tables and equations; it is enriched by an highly technical narration and supported by an (often remarked) scientific rigor of the authors; finally, it has become also a relevant contribution to the historical debate on slavery by focusing on trying to disprove one of the most established opinions, namely the fact that slavery was unprofitable, economically inefficient and overall bad for the slave.

However, this work is often presented as the beginning of the end of social science history as - through a solid counter-analysis - many of its findings have been disproved by the research community (as also described in Thomas

[1] Because they have "renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change."

(2004)). Since then, history has moved far and far away from social science practices and from the overall goal of pursuing scientific objectivity, leaving quantitative analyses only to small groups of researchers, mainly working on economic and social history.

1.3 History as a Narrative

A critical attitude towards the idea that studying the past could be conducted as a scientific practice in search of a single objective truth is not a recent concept in the historical community. In the works of George Macaulay Trevelyan (1876-1962) he remarks many times on the importance of distinguishing between the value of a rigorous source criticism and the need of historical imagination to "fill the gaps" and put together a narrative that goes beyond a simple recomposition of the collected sources. Regarding it, Macaulay Trevelyan famously remarked: "Let the science and research of the historian find the fact and let his imagination and art make clear its significance" (Trevelyan, 1927).
While, at the beginning of the XXth Century his opinions were shared by other practitioners, Evans (1997) highlights that the initial crisis of social science history emerged only during the two World Wars, where historians like Benedetto Croce (1866-1952) presented how history has been always written from (and conditioned by) the perspective of the present.

The observations of Croce had a first important impact on historiography; however it is not until the second part of the 20th Century that this critique spread across history departments, journals and conferences. The turning point could be traced as a consequence of a series of inter connected events: firstly by the clear manifestations during the early 70s of the limits and flaws of the use of social science theories and methods in supporting an empirical approach to history (as described above); secondly by the advent of a new phase in philosophy of science, which - together with the growth of post-modern theories - had a huge impact over historiography.
As a matter of fact, Kuhn (1962)'s work on the structure of scientific revo-

lutions showed how science does not procede progressively, but the shifts in its paradigms are triggered also by social factors. For these reasons, science should be studied as a historically and culturally conditioned discourse. In the following years, Kuhn's argument supported the positions of many researchers who intended to broadly criticise the assumed independence and objectivity of scientific practices (Latour and Woolgar, 1979; Latour, 1991; Pickering, 1992) as well as scholars who intended to precisely address the limits of social science methods for discovering the unique "truth" (i.e. the real causes) about what happened in the past (Gaddis, 2002; Munslow, 2006). Almost contemporary to Kuhn, Carr (1961) published its most famous essay, "What is History?", which today is widely recognised as a pillar of contemporary historiography. In the book, while Carr defends the possibility of reaching precise knowledge on the causes of past events (in chapter 3 of his book he even considers history as a social science), he strongly emphasises how the point of view of the historian will be always flawed, given the fact that he lives and he is influenced by the present ("Study the historian before you begin to study the facts", he remarks). Regarding this topic, he concludes his first chapter by saying:

> The historian and the facts of history are necessary to one another. The historian without his facts is rootless and futile; the facts without their historian are dead and meaningless. My first answer therefore to the question 'What is history?' is that it is a continuous process of interaction between the historian and his facts, an unending dialogue between the present and the past. (Carr, 1961).

In addition to the critiques previously presented, another attack against the possibility of finding the truth while conducting historical research came from the field of linguistic theory. In fact, the advent of new theories, which in same years supported the idea that the text has no reference to a external reality (it is independent from the external world as it is independent from its author (Barthes, 1967; Foucault, 1972)), had the consequence of making the historical text not different from fiction, as highlighted by Iggers (2005).

During the last thirty years, all these different critiques have had a revolutionary impact over historiography, guiding - as Evans (1997) remarks - the field towards focusing more and more on subjective discourses and narrations of the past and far away from the rigorous analysis of primary sources in search for objective evidence sustaining a theory. This turn can be noticed by considering the importance that themes from anthropology on the relevance of culture and symbols have now in historical publications as well as the current widespread focus on how cultural practices relate to wider systems of power associated with social phenomena such as ideology, class structures, ethnicity, sexual orientation, gender, and generation.

1.4 History and the Digital

Digital History, a central component of the so-called Digital Humanities (Schreibman et al., 2004), has risen through the second part of the last Century and has interacted with both of the above described aspects of historiography. As it will be described in detail in the next chapter, the initial advent of digital history between the 60s and the 80s is strongly connected with social science practices and needs, from the development of solutions for the analysis of large collections of economic and census data (North and Thomas, 1973) to the organisation of primary sources in historical databases (Thaller, 1991).

The turning point from adopting digital solutions in support of social science analyses to employ them for enhancing the collection, presentation, and dissemination of material (Cohen and Rosenzweig, 2006; Seefeldt and Thomas, 2009), could be traced to the work conducted in the 90s at the University of Virginia by scholars such as William G. Thomas III and Edward L. Ayers (see for example the pioneering project "Valley of the Shadow"[2] by Ayers and Rubin (2000)).

[2] http://valley.lib.virginia.edu/VoS/choosepart.html

As it will be presented in the next chapter, large digitisation projects such as the ones conducted at the University of Virginia have rapidly brought to the web vast collections of primary sources. This gigantic number of materials is currently forcing historians to confront once again with the methodological discussion on how we acquire knowledge about the past and to what extend we can write an accurate narrative about it. In order to efficiently study these collections, historians started employing methodologies from other fields of study, first of all natural language processing, in what has been depicted as a new turning point of our profession (Scheinfeldt, 2016; Graham et al., 2016).

Chapter 2

Digital History and Natural Language Processing

Too many thoughts
and memories
crammed into my mind.

This chapter offers an overview of the field of digital humanities and digital history, with a specific focus on the advent of computational methods to the historian's craft. Parts of it rely on the paper "From Close to Distant and Back: How to Read with the Help of Machines" that I wrote together with Rudi Bonfiglioli (developer and researcher at Imagination Technologies). For what concerns the writing of that paper, I described the current trends in the use of distant reading in the humanities (focus of this chapter), while Bonfiglioli envisioned the future impact of deep learning approaches over the hermeneutical process of analysing sources. The chapter also relies on Nanni (2015) and Nanni et al. (2016b).

2.1 Humanities Computing

Digital humanities, originally known as humanities computing (Hockey, 2004), is a variegate field of study that combines a humongous number of interactions between humanities disciplines and the use of digital technologies. From the edition of manuscripts in digital form to the use of geographical information system in historical research, from the man-computer interactions in media studies to the development of digital libraries, this field of study has gradually attracted the attention of the entire humanities community (Svensson, 2010).

While the adoption of the computer in humanities research may sound as a new research trend when presented to a general audience (Kirschenbaum, 2012), this field of study has a long history (Hockey, 2004). Father Roberto Busa's Index Thomisticus (Busa, 1980), a complete lemmatisation of the works of Saint Thomas Aquinas developed in collaboration with IBM since the late Fourties, is generally considered the starting point of the field previously called humanities computing (Dalbello, 2011). In the following decades, different humanities disciplines have approached computational methods for different purposes: from conducting stylistic analyses in literary studies, such as authorship attribution (Stamatatos, 2009) to the realisation of geographical information systems such as representation of events on digital maps (Knowles, 2008), from the digitisation (Boschetti et al., 2009) and encoding (Ide and Véronis, 1995) of analogue analogue manuscripts for the creation of digital editions to the dissemination of them through digital libraries (Rydberg-Cox, 2005).
Moreover, since 1987 the community has been intensively communicating world-wide through the Humanist discussion group[1].

In the same years, the field of natural language processing (NLP) was also establishing its position in the academic environment (Manning and Schütze, 1999; Mitkov, 2005), by focusing on developing computer programs for automatising the analysis of human language in order to perform various tasks, from machine translation to text summarisation and topic extraction.
The application of NLP methods to analyse textual documents has become

[1] http://dhhumanist.org/

a contradistinctive trait of a specific sub-group of digital humanities[2] researches too, for example stylometric tasks such as authorship attributions (Juola, 2006). However, it was not until the last decade, that these solutions have attracted the attention of the majority of the research community (Kirschenbaum, 2007). This change happened primarily as a direct consequence of the success of several digitisation projects, which were started in the 90s and now offer direct access to large collections of documents in digital format to the humanities scholars[3].

The spread in the adoption of natural language processing solutions for quantifying cultural traits from text has been depicted by many as a major turning point in literary studies (Moretti, 2005, 2013; Jockers, 2013; Jänicke et al., 2015) and historical research (Graham et al., 2016). Before diving into this debate, in the next pages we offer an overview of the field on NLP, highlighting a few relevant solutions commonly adopted in humanities research.

2.2 Natural Language Processing

Being able of identifying specific morphological, syntactic and semantic aspects in texts via computer programs and employing this information to solve precise tasks (e.g. information retrieval or authorship attribution) is a research area with a fascinating history. Natural language processing has grown during the second half of the 20th Century at the intersection of computational linguistics, artificial intelligence and philosophy of language; it has been sustained, especially in the United States, with funds related to

[2] The digital humanities community derives its name from the book “A Companion to Digital Humanities” (Schreibman et al., 2004). The change of name in the community, which also involved the renaming of the partner associations (`http://adho.org/about`) and the historical journal (`http://llc.oxfordjournals.org/`) has generated the so-called “big tent” (Svensson, 2012). This term defines the current trend of expanding the borders of the digital humanities community far away from its original commitment (namely humanities computing research tasks) to any possible interaction between humanities topics and the use of technology.

[3] Among many examples, see the online Proceedings of the Old Bailey, 1674-1913: `https://www.oldbaileyonline.org/`

socio-political topics such as the Cold War and the recent war on terrorism, given its potential for automatic machine translation. Since the second part of the 80s, thanks to the unstoppable advancement in computational capabilities which has driven an empiric turn in computational linguistic, NLP researchers have been employing more and more statistical models for describing and processing linguistic phenomena.

Nowadays, the majority of the approaches developed in NLP rely on what is called machine learning (Murphy, 2012). Machine learning is an area of research in artificial intelligence focused on the development of algorithms that can *a)* learn from and make predictions on data and *b)* build a model from sample inputs. An example could be an algorithm that learns how to group newspaper articles by topic (e.g. foreign policy, economy, sport).

During the recent decades private companies (e.g. IBM, Google, Facebook) and research centres in computational linguistics and natural language processing have been incessantly presenting new NLP methods and tools based on machine learning[4]. These innovative approaches, strongly based on statistical models and highly expensive computational power[5], are often presented to the research community as approaches able of *outperforming* all previous solutions for a specific task, which give them the possibility of enthroning themselves as "the new state of the art"[6].

While this way of openly conceiving scientific research almost as a sport com-

[4] To get an initial idea of the amounts of publications, consider the growth in number of papers submitted to NLP conference during the last ten years, with specific attention to strong method-oriented venues, such as Empirical Methods in Natural Language Processing (EMNLP): `https://www.aclweb.org/aclwiki/index.php?title=Conference_acceptance_rates`

[5] Consider for example the massive amounts of computational power needed to use deep learning methods, to know more: Hof (2013); Manning (2016).

[6] It is important to remark that all papers introducing new methods always present an experimental quantitative evaluation in support of their claim. In the third part of this thesis, we will remark on the not-so-often discussed limits of standard evaluation and evaluation metrics.

petition, with races[7], ranking[8] prizes[9] and the consequent notoriety in the field, could (and should) be discussed and criticise in a larger perspective (an initial attempt could be found in Manning (2016) regarding the advent of deep learning (LeCun et al., 2015) in NLP[10]), it is undeniable that some of the approaches recently developed in NLP directly address the need of humanities researchers. For this reason, in the next paragraphs we will introduce the most commonly NLP solutions currently adopted in DH; next we will offer a short overview on three broad NLP tasks: classification, clustering and information retrieval.

2.2.1 Text Processing Techniques

In the next paragraphs we offer an overview of common NLP techniques for processing textual documents.

Tokenization. It is the process of dividing a stream of text (which could be for example a single document) in words, phrases or other meaningful elements. When a text is broken up in words, the obtained objects are called tokens.

Part-of-Speech Tagging. This is the process of identifying the correspondent part-of-speech (e.g. noun, adjective, verb) for each word in a document. A common statistical approach to disambiguate parts of speech relies on the use of Hidden Markov Models (Kupiec, 1992).

Lemmatisation and Stemming. The common goal of stemming and lem-

[7] `http://alt.qcri.org/semeval2016/`

[8] `http://trec.nist.gov/tracks.html`

[9] `http://www.loebner.net/Prizef/loebner-prize.html`

[10] See for example the passage: "I encourage people to not get into the rut of doing no more than using word vectors to make performance go up a couple of percent. Even more strongly, I would like to suggest that we might return instead to some of the interesting linguistic and cognitive issues that motivated noncategorical representations and neural network approaches."

matisation is to reduce inflectional forms (for example “plays”, “played”) of the same word to a common base (e.g. “play”). However the two approaches are different. While stemming applies heuristics such as “chopping” the end of words in order to achieve the goal, lemmatisation follows a linguistic approach based on vocabulary looking and morphological analyses. Choosing between one or the other approach depends on the specific task and on other factors, such as the fact that a lemmatiser needs linguistic resources to be build.

Named Entity Recognition. Named Entity Recognition (NER) is the task of detecting and classifying entities, which have been named in text, into pre-defined categories such as the names of person, organisations, locations, etc. NER systems can use linguistic grammar-based techniques as well as statistical machine learning solutions. Statistical NER systems (based for example on conditional random fields (McCallum and Li, 2003)) typically require a large amount of manually annotated training data, but show higher performances.

Entity Linking. The task of linking textual mentions to an entity in a knowledge base is called entity linking or entity resolution Rao et al. (2013). This is an information extraction task that involves being able to recognise named entities in text (such as people, locations, organisations), resolving coreference between a set of named entities that could refer to the same entity (e.g. “Barack Obama” and “Mr. President”) and disambiguating the entity, by linking it to a specific entry in a knowledge base such as DBpedia (Bizer et al., 2009), Yago (Suchanek et al., 2007) or Freebase (Bollacker et al., 2008). Some authors (Chang et al., 2016) distinguish between two tasks: first, Entity Linking, where mentions corresponding to named entities are considered. Second, Wikification, where mentions to any term present in Wikipedia (even if they are common nouns) are considered. In this thesis we refer of Entity Linking for both cases. The disambiguation process is especially challenging, as mentions of an entity in text can be ambiguous. For this reason, entity linking systems such as TagMe! (Ferragina and

Scaiella, 2010), TagMe 2 (Cornolti et al., 2013), DBpedia Spotlight (Mendes et al., 2011) or Babelfy (Moro et al., 2014) examine the mention in context in order to precisely disambiguate it. For instance, in the expression "Clinton Sanders debate", "Clinton" is more likely to refer to the DBpedia entity *Hillary_Clinton* than to *Bill_Clinton*. However, in the expression "Clinton vs. Bush debate", the mention "Clinton" is more likely to refer to *Bill_Clinton*.

Word Embedding. Word embedding identifies a group of language modeling and feature learning approaches, which map words (or phrases) to vectors of real numbers. Basically, these solutions connect (embed) a space where each word represents one dimension in a continuous vector space. These solutions for representing words have shown to improve performances in many NLP tasks (Manning, 2016) and have been recently adopted in digital humanities research as well[11]. The most emploied toolkits for producing word embeddings are word2vec, developed by Google[12] and GloVe, by Stanford University[13].

2.2.2 Text Mining and Information Retrieval

Humanities scholars have often a few key requirements when dealing with large textual collections: *a)* to identify semantic similarities; *b)* to quantify recurrent lexical patterns; *c)* to retrieve specific pieces of information from it. The first two requirements are generally described with the term of text mining, while the last task is commonly defined as information retrieval. Below we give an overview of two major approaches of text mining, namely classification and clustering, together with an overview of information retrieval. These different methods will constitue central pieces of the third part of this thesis.

[11] See for example: `http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html`

[12] `https://code.google.com/archive/p/word2vec/`

[13] `http://nlp.stanford.edu/projects/glove/`

Classification

In machine learning there are two main approaches that allow to conduct any text mining analysis. The first one consists of supervised learning methods, which focuses on classification tasks. In classification, humans identify a specific property of a subset of elements in the dataset (for example the property of being an article about foreign policy in a newspaper archive) and then guide the computer, by means of an algorithm, to learn how to find other elements with that characteristic. This is done by providing the machine with a dataset of labeled examples ('this is an article about foreign policy', 'this is not'), called a 'gold standard', which are described by a set of other properties[14] (for instance, the frequency of each word in each document).

The potentialities of a good classifier are immense, in that it offers a model that generalises from labeled to (a potentially very large set of) unlabeled data. However, building such models can also be extremely time consuming. In fact, researchers not only need a dataset with specific annotated examples to train the classifier but, perhaps even more fundamental, they need to have a clear goal in mind, since this influences the annotation guidelines and the learning task itself. For this reason, it is evident that classification methods are arguably not the most convenient approaches for conducting data exploration, i.e., in those situations where a researcher sets to investigate a large dataset with no clear goal in mind other than searching for any phenomenon they deem interesting a posteriori.

Common classification algorithm are decision trees, naive-Bayes and support vector machines (SVM)[15]. In particular, SVM has been adopted by us in the third part of this thesis to perform different classification tasks.

Clustering

The second class of methods consists of unsupervised ones, and address the problem of clustering. In a nutshell, clustering methods aim at grouping ele-

[14] In machine learning these properties are called 'features'.

[15] For applications in the humanities see for example: Yu (2008).

ments from a dataset on the basis of their similarity, as computed from their set of features (for example by looking at patterns in the frequency of words in different documents). This is achieved by merely computing likenesses across features without relying on labeled examples, hence the lack of supervision from the human side.[16] Crucially for the digital humanities scholars, researchers can then study the obtained clusters in order to understand what the (latent) semantic meaning of the similarities between the elements is.
Clustering techniques are extremely useful for analysing large corpora of unlabelled data (i.e., consisting of 'just text'), since they rapidly offer to researchers a tool to get a first idea of their content in a structured way (i.e., as clusters of similar elements, which can be optionally hierarchically arranged by using so-called hierarchical clustering methods). This is primarily because, as they do not require labeled data, they can be applied without having in mind a specific phenomenon or characteristics of the dataset to mine (i.e., learn).

Information Retrieval

Information retrieval (IR) is the prototypical task that every historian performs when consulting an archive, as well as when each one of us uses a commercial web search tool such as Google.
Given a collection of documents (e.g. a newspaper archive) and a specific user need defined as a query (e.g. "articles on the relation USA - Iran"), the IR system retrieves and ranks a sub-set of the collection which is - supposedly - related to the query. IR systems can be based on language models such as query-likelihood (Zhai and Lafferty, 2001), they can use metrics such as the Page rank (Page et al., 1999) to give a relevance score to the retrieved documents and can also adopt human supervision to improve their performance through machine learning (see for example Learning to Rank (Liu, 2009)).
The importance of IR systems for historical research was initially remarked

[16] In recent years the use of deep learning methods for the analysis of textual documents has grown in the fields of natural language processing and machine learning. In the near future these so-called techniques could become a new unsupervised approach for data exploration in digital humanities.

by Vannervar Bush in his article “As We May Think” (Bush, 1945), showing possible application of his theoretical system, the so-called “Memex”; since then an innumerable number of articles in digital history have remarked the usefulness of IR for historical research, together with the biases and flaws that these systems could have.

2.2.3 Enthusiasm and Critiques in the Digital Humanities

Even if, in the last twenty years, the NLP community has developed many resources and tools to deal with large-scale collections and has been involved in many interdisciplinary research projects, the initial unrest of the humanities community to the advent of big cultural data could be perfectly summarised by the title of a famous paper written by Gregory Crane: “What do you do with a million book?” (Crane, 2006).

Franco Moretti is often identified as the scholar that has offered the most convincing answer to this question, presenting an innovative perspective to humanities research and additionally bringing the discussion on the potential of using computational methods for studying cultural phenomena to the main public (Rothman, 2014). On the one side, his publications in the field of literary studies on the use of computational techniques in order to extract quantifiable information from large amount of texts (Moretti, 2005, 2013) attracted the attention of traditional humanities scholars (Liu, 2012) and of mainstream newspapers (Schulz, 2011). On the other side, his “scientification” of literary studies practices (Merriman, 2015), from the definition of “distant reading” to the creation of a “Stanford Literary Lab”, suggested a completely different way of conceiving research in the humanities.
In his works, Moretti addresses in particular the weaknesses of traditional close-reading approaches used in literary criticism, which are characterised by a careful interpretation of brief passages for the identification of different layers of meanings. In his vision, literature could and should be understood “not by studying particular texts, but by aggregating and analysing massive

amounts of data" (Schulz, 2011); examples of this could be seen in Moretti (2013) "Reflections on 7,000 Titles", where the author initially examines different aspects of novel titles, such as the presence of a woman's proper name or the mention of ethical content (e.g. discipline, correction, moderation), then quantifies their presence over time and explains the variation in the obtained diachronic distribution. Moretti then shows how certain thematic aspects could be captured into a specific formula, such as "the x of y" to precisely model a characteristic of novel titles in the XIX Century gothic genre.
Several digital humanities scholars follow Moretti's quantitative approach and agree with him in considering the advent of distant reading a drastic turning point for humanities studies (Jockers, 2013; Graham et al., 2016)[17]. They point out how computational methods could be a solid alternative to tradition hermeneutic approaches, both in literary studies and in historical research, not only in order to deal with huge amount of sources in digital form, but especially to re-discuss work practices, methodologies and long-standing established theories.

Distant reading approaches have attracted great enthusiasm in digital humanities so far (together with solid attention from outside the academia and large support in research founding), but they have also received specific critiques. First of all, it has been pointed out that these methods try to automatise an acquisition process of knowledge (Fish, 2012a,b). This may make the humanist scholar, his/her background knowledge and hermeneutic approach irrelevant to the production of insights and transforms every aspect of these studies in the identification of quantitative features, aspects and evidence (Posner, 2015). Secondly, its has been also remarked how, for the moment, distant reading studies have developed an immense number of new tools, methods and techniques but produced relatively little in terms of

[17] See for example of the new journal of Cultural Analytics is presented: "Cultural Analytics is a new open-access journal dedicated to the computational study of culture. Its aim is to promote high quality scholarship that intervenes in contemporary debates about the study of culture using computational and quantitative methods."

new humanities knowledge (Blevins, 2015).
As others (Owens, 2012) have already pointed out, nowadays digital humanists seem to be often too easily seduced by the "big data" rhetoric of "making the data speak for itself" (Owens, 2015) which, as it will be extensively remark later, is first of all very simplistic (no way of approaching a text is theory-neutral, even an algorithm relies on a long list of assumptions) and second of all absolutely not in line with the critical/skeptical approach that has always characterised a solid humanities investigation.

Let us consider a specific example, that precisely highlights this current issue. It involves the development and presentation of a computational tool for the extraction of sentiment and sentiment-based plot arcs from text.[18] This method, called Syuzhet,[19] has been presented by its author, Matthew Jockers, with a post on his blog (Jockers, 2015), anticipating the publication of an academic paper dedicated to it. On the following weeks, his work has attracted great interest in the digital humanities community, generating a series of discussions in the Humanist mailing list, on Twitter and on Jockers' personal blog.[20] In particular, Swafford (2015)'s critique of Jockers' work echoes a few recurrent issues in current stages of digital humanities: first of all little consideration on the statistical assumption on which the adopted method is based, secondly an almost-blind trust in the developed solution as it intuitively seems to produce reasonable results and finally the complete absence of solid data science evaluation practices while developing and adopting computational methods for the quantification of specific properties in text, which are absolutely necessary to back-up the interpretation of the obtained outputs (as also remarked in Traub and van Ossenbruggen (2015)). See for example how Jockers simplifies the reasons why he adopted the Fourier transformation in his system, around which Swafford built her cri-

[18] `https://github.com/mjockers/syuzhet`

[19] From Vladimir Propp's definition of the organization of the narrative as opposed to the fabula.

[20] As precisely summarised here: `\url{https://storify.com/clancynewyork/contretemps-a-syuzhet}`

tique (to know more, see Schmidt (2016)):

> So, I needed a way to deal with length. I needed a way to compare the shapes of the stories regardless of the length of the novels. Luckily, since coming to UNL, I've become acquainted with a physicist who is one of the team of scientists who recently discovered the Higgs Boson at CERN. Over coffee one afternoon, this physicist, Aaron Dominguez, helped me figure out how to travel through narrative time.
>
> Aaron introduced me to a mathematical formula from signal processing called the Fourier transformation. The Fourier transformation provides a way of decomposing a time based signal and reconstituting it in the frequency domain. A complex signal (such as the one seen above in the first figure in this post) can be decomposed into series of symmetrical waves of varying frequencies. And one of the magical things about the Fourier equation is that these decomposed component sine waves can be added back together (summed) in order to reproduce the original wave form–this is called a backward or reverse transformation. (Jockers, 2015)

While distant reading solutions have huge potential, this example serves the purpose of summarising the issues that emerge when adopting them uncritically in humanities research. In the next section, we will go deeper into the application of NLP solutions in historical research, in order to understand how the community has been dealing with a never experienced before abundance of primary sources.

2.3 History and Computing

Literary studies, philology and historical research have been long-term key components of humanities computing. In particular, during the second part of the XX Century the potential of computational methods for the analysis

of primary sources has been a recurrent topic in historiography. As Thomas (2004) remarked, already in 1945 Vannevar Bush, in his famous essay 'As We May Think', (Bush, 1945) pointed out that technology could be the solution that will enable us to manage the abundance of scientific and humanistic data; in his vision the Memex could become an extremely useful instrument for historians (Thaller, 1991).
The use of the computer in historical research, which consolidated between the 60s and the 70s with his application to the analysis of economic (North and Thomas, 1973) and census data (Wrigley, 1973), has been initially strongly related to the adoption of social science practices in historical studies. A pioneering work on the use of database technologies for historical research was conducted by Manfred Thaller during the 80s (Thaller, 1991). The rise of Cliometrics (McCloskey, 1978) supported by the intent of bringing scientific objectivity to the craft (Iggers, 2005) gave birth to a long discussion on the use of the results of quantitative analysis as evidence in the study of the past. Since the 80s, while specific sub-fields, such as economic history continued to rely on these methodologies (and strongly differentiated themselves and their research practices from the rest), the overall majority of the field had turned and embraced the so-called cultural turn (Burke, 2008), which is based on the adoption of methodologies from anthropology and literary studies and a major attention to discourses rather than on modelling causes and consequences (Iggers, 2005; Evans, 1997).

Due in part to this long dispute on the application of quantitative methods for interpreting the past in a "more objective way", in part to the cultural turn and in part to the new potentialities of the Web as a platform for the collection, presentation, and dissemination of material (Cohen and Rosenzweig, 2006; Seefeldt and Thomas, 2009), during the 90s a different research focus emerged in what was identified as digital history.[21] As Robertson (2014b)

[21] "So far few historians have tried to define 'digital history.' We were probably the first to use the term when Ed Ayers and I founded and named the Virginia Center for Digital History (VCDH) in 1997–1998" (William G. Thomas III in Interchange: the promise of digital history, 2008).

recently pointed out, this specific attention on the more "communicative aspects" of doing research in the humanities could be recognised as the main difference between how, during the last twenty years, historians have been interpreting the digital turn compared to their colleagues in literary studies. This could be also noticed by observing the importance given to digital public history topics (Noiret, 2015), the relevance of teaching in digital history (Dougherty and Nawrotzki, 2013; Robertson, 2014a) and the solid tradition of digital history mapping (Knowles and Hillier, 2008).[22]

However, the attention to the potentialities of computing techniques for exploring the past has not been confined only to Geographic Information System (GIS) research projects. In fact, in the last fifteen years, thanks in particular to the prompt availability of digitised historical primary sources, out-of-the-box natural language processing solutions and the potentialities of web technologies, a few interdisciplinary teams have been developing tools (Rockwell, 2006; Buechler et al., 2008; Sinclair et al., 2012; Moretti et al., 2014)[23] in order to help other traditionally trained historians to employ natural standard language based computational methods in their work. Results of the adoption of these tools are presented in Cohen (2010b); Houston (2014); Büchler et al. (2014); Sprugnoli et al. (2015); Moritz et al. (2016).

2.3.1 Developing Text Mining Platforms

In December 2010 Google presented a service, called the 'Google Ngram Viewer'[24]. This tool, based on natural language processing (NLP) techniques, allows us to look at the occurrence of single words or sentences in specific

[22] To know more about it, see for example the pioneering project "Valley of the Shadow" (`http://valley.lib.virginia.edu/VoS/choosepart.html`) (Ayers and Rubin, 2000) or the more recent "Visualizing Emancipation" (`http://dsl.richmond.edu/emancipation/`) (Nesbit and Ayers, 2013) and "Mapping the Republic of Letters" (`http://republicofletters.stanford.edu/`) (Findlen et al., 2011).

[23] As it will be described later, while the Google Books project is a tool for conducting digital historical analyses, for a series of specific reasons it has not been conceived and developed inside the digital history community.

[24] `https://books.google.com/ngrams`

subsets of the immense unstructured textual corpus digitised by the Google Book project.

With an article on Science (Michel et al., 2011) and the brand "culturomics" announcing their work, a few weeks later Erez Lieberman Aiden and Jean-Baptiste Michel, team leaders of the project, offered a demonstration of the tool at the annual meeting of the American Historical Association in Boston (Grafton, 2011). In front of 25 curious historians, they remarked on the enormous potential of conducting historical research by extracting information from large corpora. In particular, they revealed a way to deal with one of the biggest issues for historians that are exploring large collections of texts (Crane, 2006), namely rapidly quantify the distribution of specific words (representative of concepts) in the corpus.

The Google Ngram Viewer is a tool that enhances what Moretti (2000) defined as distant reading, namely analysing and identifying patterns in large collections of unstructured text. Interestingly, the development and the functionalities of this tool highlight some of the most relevant characteristics of the current interactions between the practice of historical research and the use of computational methods:

- First of all, as it has been already remarked (Cohen, 2010a; Grafton, 2011; Milligan, 2012), no historian has been directly involved in any step of the development of this project. This is particularly significant, given that they would be likely to be the primary targets of a tool able to process information from a corpus spanning five hundred years. As Aiden and Michel remarked themselves[25], this is due to two well known reasons: historians traditionally do not have solid computational skills in their academic background and they are often skeptical regarding the development of computational / quantitative approaches for the analysis of sources (Gibbs and Owens, 2012).

- Secondly, even if this solution adopts advanced NLP techniques with the goal of extracting meaningful information from large collections

[25] See: `http://www.culturomics.org/Resources/faq/thoughts-clarifications-on-grafton-s-loneliness-and-freedom`

of text (i.e. text mining), what the Ngram Viewer offers to the end-user is an over-simplified research tool which usually leads to general coarse-grained explorative analyses and only to few simple historical discoveries (Cohen, 2010a).

- Additionally, the Ngram Viewer remains a so-called black-box[26] and researchers do not know exactly how the algorithm works and cannot adapt it to perform a different analysis[27].
- Finally, the way in which the Ngram Viewer has been presented[28] and identified outside academia (Fischer, 2013) as a representative tool of the digital humanities also reveals the characteristic enthusiasm for methodology studies and big-data driven researches that is consistently growing in this community (Blevins, 2015). However, as also remarked before, researchers in this field of study need to bear in mind their long-term purpose, that is to use the computer in order to answer specific and relevant humanities research questions and not simply to build tools (Thaller, 2012)[29].

The Ngram Viewer represents a current widespread way of developing and employing computational methods, and in particular natural language processing solutions, in historical research, namely for coarse-grained exploration

[26] A black box could be defined as a tool or a system that can be viewed in terms of its inputs and outputs, but that does not provide any knowledge of its internal workings.

[27] The importance of always being critical towards computational methods and adopt them only after a careful examination will be a central aspect of the third part of the thesis, with precise attention to the use of topic modeling in digital humanities studies.

[28] 'Culturomics is part of what's known as 'humanities computing' or the 'digital humanities'. The digital humanities are a very broad field, comprising a vast array of ways in which computation can help humanists. It includes such things as tools that aid in teaching, citation, and collaboration as well as digital collections of various types. Culturomics is much more narrowly defined: its goal is to digitize and analyze data about culture on extremely large scales: all books, all newspapers, all manuscripts, etc.' (from Culturomics FAQ section: `http://www.culturomics.org/Resources/faq`).

[29] In particular Thaller (2012) remarks: "It was rather clear, however, that historical analysis improved by methodologically designed tools was the goal, and what had to be done to achieve that, like preparing analog sources digitally in such a way, that afterwards they would be accessible for the most diverse type of analysis possible, was (only) the way."

of datasets. The tools recently developed by digital history groups, try to tackle some of these issues, by offering a wide range of natural language processing implementations, such as part-of-speech tagging and named entity recognition (see for example Alcide (Moretti et al., 2014)). While these solutions guarantee different ways of conducting text explorations, their potentialities remain, of course, limited to the availability of a specific tool.[30].

This reason, combined with the natural difficulties of everyday collaboration between humanities scholars and computer scientists while conducting a joint interdisciplinary project and the economic cost of receiving technical support from IT experts (as remarked by Crymble (2015b)) brought a small but strongly connected community of traditional historians to focus its efforts in teaching themselves the basic of programming languages and independently exploring the potentialities of different textual analyses techniques for conducting exploratory studies of their datasets. As Turkel (Cohen et al., 2008) highlighted, re-echoing the already remarked importance of "teaching" in digital history: "My priority is to help train a generation of programming historians. I acknowledge the wonderful work that my colleagues are doing by presenting history on the Web and by building digital tools for people who can't build their own. I know that the investment of time and energy that programming requires will make sense only for one historian in a hundred".[31] In the final part of this thesis we will discuss whether this technical focus on the importance of programming is actually the key aspect that will allow historians to fully adopt computational methods in their research.

[30] For example no one of the solutions presented before permits to adopt word-embedding vectors (Turian et al., 2010) in the analysis, which are becoming an essential solution in NLP.

[31] Turkel (2008) also remarked the impact of the computational turn in the fields of geography and biology. Another important example is the interplay in language studies between theoretical linguistic on one side, and corpus linguistics and computational linguistics on the other side.

2.3.2 Computational History

The works conducted by Willam J. Turkel at the University of Western Ontario, with particular attention to his blog[32] and his project “The programming historian”,[33] could be identified as the starting point of these digital interactions. Following Turkel’s approaches and advice, a group of historians has begun experimenting directly these different computational methods to explore large historical corpora (Gibbs and Cohen, 2011; Milligan, 2012; Crymble, 2015a). The use of natural language processing and information retrieval methods, combined with network analysis techniques[34] and a solid set of visualisation tools, are the points around which this new wave of quantitative methods in historiography has consolidated.

During the last decade several interesting examples of these interactions between historical research and computational approaches have been presented (Nelson, 2010; Wilkens, 2013; Blevins, 2014; Kaufman, 2015; Graham et al., 2016). For example, Nelson (2010) adopted a popular text mining method called LDA topic models (Blei et al., 2003)[35] to explore Confederate newspaper articles from the Richmond Daily Dispatch, while Kaufman (2015) combined text mining and network analysis to examine the Digital National Security Archive (DNSA) Kissinger Collections, which comprises approximately 17500 meeting memoranda (‘memcons’) and teleconference transcripts (‘telcons’) detailing Kissinger’s correspondence during the period 1969-1977.
In addition to this, thanks to the collaborations with other digital humanities colleagues (i.e. literary studies researchers and digital archivists), the words ‘text mining’ and ‘distant reading’ have become brands of this new trend in

[32] http://digitalhistoryhacks.blogspot.com/

[33] To know more see Crymble et al. (2012) and visit the website: http://programminghistorian.org/

[34] With the use of Gephi (Bastian et al., 2009) and Palladio for example. To know more, see: http://programminghistorian.org/lessons/creating-network-diagrams-from-historical-sources

[35] This approach will be described in details in the third part of this thesis.

digital history and started to appear at traditional historical conferences[36], journals (Ewing et al., 2014) and the American Historical Association has recently defined a series of guidelines for properly evaluating these contributions[37]. Moreover, a series of workshop on the adoption of computational and quantitative methods in historical research have been organised since 2013 at digital humanities conferences.[38]

The excitement regarding this growth in application of text mining methods, which is perceived by many in the community as a methodological turning point in the historical profession, is clearly described by Scheinfeldt (2016):

> My difficulty in answering the question "What's the big idea in history right now?" stems from the fact that, as a digital historian, I traffic much less in new theories than in new methods. The new technology of the Internet has shifted the work of a rapidly growing number of scholars away from thinking big thoughts to forging new tools, methods, materials, techniques, and modes or work which will enable us to harness the still unwieldy, but obviously game-changing, information technologies now sitting on our desktops and in our pockets.

However, if we go beyond this enthusiasm for new tools and methods, and we give a closer look at how more advanced computational techniques have been applied so far by historians (such as the works presented by Nelson (2010); Kaufman (2015), we could notice that the first goal of the researchers has been to show the explorative potentialities of these methods (as described by Brauer and Fridlund (2013)) and to test their accuracy by re-confirming already well-known historical facts (e.g. Au Yeung and Jatowt (2011) and

[36] See for example: `http://aha2013.thatcamp.org/`

[37] `https://www.historians.org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history`

[38] See for example the HistoInformatics workshops: `http://www.dl.kuis.kyoto-u.ac.jp/histoinformatics2013/` and the Computational History workshop: `http://kdeg.scss.tcd.ie/1st-international-workshop-computational-history`

Acerbi et al. (2013)). This is due to three main reasons. The first is the unsupervised nature of the specific textual analysis techniques most widely used in historical researches (e.g., topic modeling), which do not need (but at the same time cannot obtain benefit from) human supervision and in-domain knowledge during the computational process. The second reason is the *imprecision* of state-of-the-art solutions in natural language processing, which cannot be employed for generating quantitative evidence before having performed a solid task evaluation (as it will be extensively remarked in the third part of this thesis). The third, and most important reason, is the already mentioned widely diffused skepticism regarding the idea that understanding the past could be pursued through the quantitative examination of specific phenomena (such as the variations in frequency of words in a diachronic corpus as evidenced of specific cultural changes in society).

The combination of these three limitations has had a specific consequence on digital history, as is brilliantly pointed out by one of the most thoughtful articles on the topic:[39]

> [...] there is a fundamental imbalance between the proliferation of digital history workshops, courses, grants, institutes, centers, and labs over the past decade, and the impact this has had in terms of generating scholarly claims and interpretations. The digital wave has crashed headlong into many corners of the discipline. Argument-driven scholarship has largely not been one of them.
> There are many reasons for this imbalance, including the desire to reach a wider audience beyond the academy, the investment in collection and curation needed for electronic sources, or the open-ended nature of big digital projects. All of these are laudable. But there is another, more problematic, reason for the comparative inattention to scholarly arguments: digital historians have a love affair with methodology. We are infatuated with the power of

[39] Interestingly (but extremely common in the digital humanities community), this article has been published as a blog-post on the author's personal website, before becoming a book chapter.

> digital tools and techniques to do things that humans cannot, such as dynamically mapping thousands of geo-historical data points. (Blevins, 2015)

2.3.3 A Turning Point in Historiography?

Precisely addressing the above mentioned limitations of current digital history scholarships, Nelson (2016) recently presented the works conducted by Wilkens (2013) and Blevins (2014) as the first examples of researches in the field able of offering a contribution to the community in term of novel arguments and interesting quantitative proofs, instead of simple applications of computational methods for explorative research. Wilkens (2013), in his work, defines and assesses the "geographic imagination" of American fiction around the Civil War, based on an analysis of more than 1,000 novels by American authors published in the US between 1851 and 1875; Blevins (2014), in his study, examines how a middling regional newspaper such as the Johnston's Houston Daily Post constructed an imagined geography between 1894 and 1901 and produced space in relation to the large-scale forces reshaping late nineteenth-century America[40].

While highlighting the importance of these studies both for the digital and the traditional historical communities, three central pillars emerge:

- **Computational Methods.** The papers presented above adopt computational methods to quantify in different ways the "cultural construction of space in the United States", over large collections of primary sources in a way it could not have been possible with traditional approaches only.

- **Quantitative Evidence.** The argumentations of the two historical paper presented are based on solid quantitative evidence, which are extracted directly from the text. While Blevins clearly distinguishes

[40] In addition to this two studies presented by Nelson, we also consider Crymble (2015c) study on Irish immigration to Eighteen Century London as another example of a fruitful combination of computational methods and historical inquiry

himself from a cliometrician by saying: “I am not trying to revive the quantitative history movement of the 1960s and 1970s, and I do not elevate digital analysis as a form of scientific “proof.” Distant reading cannot and will not replace the close reading of historical texts and the interpretation of their meaning and context”, his work is a solid example of a historical scholarship that fruitfully adopts quantitative evidence in its argumentation.

- **Appendix.** The title of Nelson’s paper, “Digital History as Appendix”, summarises perfectly what DH would / could become in the near future. A technological and methodological appendix of a solid historical argumentation. This aspect re-evokes the importance of the statistical analyses conducted by Martin Offenbacher (Adair-Toteff, 2014), which are the basis upon “The Protestant Ethic and the Spirit of Capitalism” (Weber, 1905) has been built.

Wilkens (2013) and Blevins (2014) are solid initial examples of the beginning of a mature season in digital history, where computational methods are used to generate quantitative evidence in support of a historical narrative. However, while this step is important, we do not think it will have per-se a revolutionary impact on the theoretical assumptions of our profession as a whole. As a matter of fact, we already mentioned that the use of quantitative evidence in historical research has a long tradition and a solid group of practitioners (for example economic historians) that have continued to adopt these approaches during the entire cultural/linguistic turn (Rorty, 1992) of our discipline. The recent growth in publications and adoptions of computational approaches in humanities research (Manovich, 2011; Berry, 2011; Kitchin, 2014) is a methodological wave like others we have already experienced before[41], which during the last ten years has been especially strongly sustained and encouraged by public[42] and private[43] institutions as well as

[41] To know more see Howell and Prevenier (2001); Gaddis (2002); Iggers (2005)

[42] http://www.digitalmeetsculture.net/tag/horizon-2020/

[43] https://www.volkswagenstiftung.de/en/en/mixedmethodshumanities.html

by private companies[44] and (with some important exceptions (Marche, 2012; Fish, 2012b; Allington et al., 2016)) by academic and mainstream media sources[45]. For this reason, the final advent of historical scholarships that are based on the results produced by these methods cannot be considered as a sign per-se of a more theoretical turning point in historiography.

However, we do believe that historiography is about to experience a drastic turning point in its theories and methods. In the next chapters of this thesis, we in fact argue that what will change completely the way we will study the past in the next decades are not (simply) computational methods, but especially the inner nature of the new primary sources we are currently producing everyday. Born digital documents shared online, their preservation, availability and access, we argue, will be the turning point of our profession and this will have an impact both on methods and underlying theories and assumptions of the historical community. As it will be described from the next chapter on, it is not a matter of qualitative over quantitative, distant versus close, hermeneutics against statistical significance. It is a matter of learning a new way of acquiring knowledge on our past.

[44] https://googleblog.blogspot.de/2010/07/our-commitment-to-digital-humanities.html

[45] See for example how *The New York Times* covered the topic: http://topics.nytimes.com/top/features/books/series/humanities_20/index.html

Chapter 3

The Born-Digital Turn

And then one day you find,
ten years have got behind you.

This chapter introduces several relevant topics for this dissertation, such as the preservation of web documents, the relations between internet studies, contemporary history and digital humanities and how web archive materials could be adopted as primary sources. It relies on positions already expressed in Nanni (2015), Nanni (2017) and in Chakraborty and Nanni (2017).

3.1 The Web as a Source?

Digital History (with the capital "H"[1]) has been focused primarily on the exploration and study of digitised collections, such as the Proceedings of the Old Bailey[2], the Europeana Newspapers archive[3] and the HathiTrust collections[4] (see for example the works conducted by Cohen et al. (2011); Seefeldt

[1] Michael Frisch, distinguishes "digital [H]istory" and "digital [h]istory". By "Digital History" he intends research and analysis while by "Digital history" he refers to sources, which were digitised. (Cohen et al., 2008)

[2] `http://www.oldbaileyonline.org/`

[3] `http://www.europeana-newspapers.eu/`

[4] `https://www.hathitrust.org/`

and Thomas (2009); Moreux (2016)) and, until recent years, very few works have employed born-digital materials, namely documents that exist only in digital format. As Brügger remarked, so far "digital history largely equates a 'digital source' with a source which was previously analog but has now been digitised. In other words, digital sources have so far been limited to traditional analog sources in digital form" and "very little attention has been paid to the new digital media as historical sources" (Brügger, 2012c).
However, as part of a larger (and apparently unstoppable) transition to the digital that our society is currently experiencing (Mayer-Schönberger and Cukier, 2013), documents that have been traditionally adopted as primary evidence by historians (such as printed news articles, personal diaries, letter correspondences or scientific publications) are now created and shared only in digital format (to know more see Brügger (2016)). In addition to these sources, new materials such as websites, blogs, tweets, discussion-threads in forums, edit history of Wikipedia articles and large-scale knowledge bases are becoming evidence of our recent past.

While this transition to the digital could (and we argue will) have a revolutionary impact on the way we study the past, we can identify a few main reasons for the scarce attention that the historians community has so far given to this issue. The first reason is rather obvious, namely the fact that most of historians dedicate their work to pre-90s history, when the overall majority of the documents are analogues (or digitised), and this contemporary transition to born-digital sources will have basically no impact.
The second reason is slightly more complex. It relies on a diffused perception in the historians community that the 90s are still not "history"[5]. Ian Milligan has remarked on this issue several times, both in his publications (see for example Milligan (2016a)) and during his presentations (e.g. Fig. 3.1) at the most relevant international meetings of our community. Additionally, Milligan highlighted how much this attitude is in contrast with previous tendencies in contemporary history, where the past became object of study

[5] One of the first example of historical studies on the Nineties which uses born-digital sources is by Ben-David (2016) on the Yugoslav web sphere.

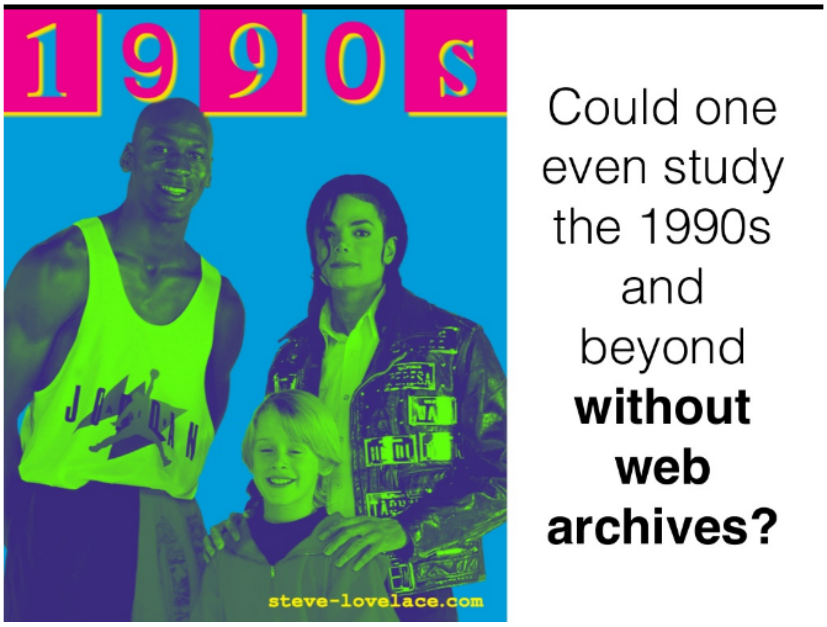
already after twenty, thirty years[6].

The third reason (which - we argue in this thesis - could give an explanation to the second one), is that, for a historian, studying the 90s means learning to deal for the first time with this huge variety of new primary sources that exist only in digital format. These materials, such as websites, forums, blogs, tweets, emails, are generally identified as born-digital sources (to know more see Cohen and Rosenzweig (2006); Noiret (2009); Brügger (2012c)) and, as it will be described in the next pages, differ completely from traditional analogue and traditional digitised primary sources. To give a toy example, a born-digital article on a news website is a completely different object of study compared to an article published in a printed version of a newspaper and to a printed article that has been digitised and re-presented online on a digital archive. Differences rely in the way a born-digital article is retreived (if the URL of the article changes or the article is removed from the website where it has been originally published that piece of news could be completely lost), its reliability is established (a digital article could be changed substantially over time without notifying it to the readers – to know more see Benvenuti and Morriello (2006)) and its existence is examined in a digital environment (an online article is strongly connected and interact with a series of other digital objects, such as the underlying comment area as well as social media channels).

Niels Brügger also recently remarked on this while discussing about archived web pages, which will be the topic of this chapter:

> A medieval manuscript is identical to itself over time. The manuscript ordered at the library desk is the exact same as was written 600 years ago. And asking for it at the library desk two persons would get the same manuscript: you get what you see. It is different with an archived web page where you do have different layers and different points of time in the same (archived) web page. (In Skouvig (2016))

[6] For example, already at the end of the 80s, events from the 60s such as the Cuba Crisis were examined as "historical" facts (see Paterson and Brophy (1986); Lebow (1990); Medland (1990))



Could one even study the 1990s and beyond **without web archives?**

Figure 3.1: A recurrent provocative slide from Milligan's presentations.

It is evident that, while "new media is that not new anymore" (Milligan, 2016a) for our society, it seems to remain a novelty for historians. In the next pages, we first present an overview on the advent of born-digital materials, we highlight an issue regarding their ephemerality and we present the international public and private efforts for preserving them. Then, we describe how these materials have been adopted as historical sources in a few experimental studies in the field of internet history.

3.2 The Ephemerality of Web Materials

In 1989 Tim Berners-Lee introduced its project at CERN, which later was identified as the "World Wide Web". In 1991 he created the first website, `http://info.cern.ch/`, and in the same year he publicly presented it in the Usenet newsgroup `alt.hypertext`.

After an initial slow start (Frana, 2004), by the end of 1995 the web had more than 16 million users,[7] which were already at that time the consumers

[7] `http://www.internetworldstats.com/emarketing.htm`

and the producers of a large amount of born-digital documents. In the following twenty years the widespread diffusion of internet connections (and consequently of web access) has kept growing at an unimaginable rate[8] (see Fig. 3.2[9]), and with that also the production of born-digital contents[10]. In 2013, IBM described this incredible growth as such: "90% of the data in the world today has been created in the last two years alone"[11].
Nowadays, the overall majority of the traces we are leaving of our everyday life are digital (Martini, 2012): we still write letters everyday (and almost certainly many more than ever before, as remarked by Cohen (2006b) regarding US administrative correspondences) but they are electronic mails, we still have diaries and memoirs in the form of personal blogs, we report news (both true and fake) through social media channels, we share scientific discoveries on open access all digital journals, we produce and consume music, books, movies which exist only in digital format, we update, modify, correct, re-modify, re-correct the largest encyclopaedia ever written[12].

In the last 25 years, thanks to all these efforts the web has kept growing at an unstoppable pace. However, given its digital nature, it has been also constantly changing, usually leaving only a few traces of its past. As a matter of fact, these documents are extremely ephemeral (as already remarked by Brügger (2005, 2009); Masanès (2006); Dougherty et al. (2010); Gomes et al. (2011)): web pages disappear constantly from the live web (because they are removed by the author or by the owner of the platform - e.g. for copyright issues), leaving a familiar trace of 404 status code messages (see Fig. A.7).

[8] Consider for example the current number of Facebook and Twitter users, two social networks created in the last decade. In June 2016 the number of Facebook monthly active users is over 1.65 billion, while on Twitter they are 305 millions.

[9] Figure from `http://www.internetlivestats.com/internet\-users/`

[10] See for example the growth in size of Wikipedia `https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia` and the amount of video time uploaded on Youtube every minute `http://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/`

[11] See `http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html`

[12] Which was partially printed (luckily) only once, as part of an art project (`http://printwikipedia.com/`).

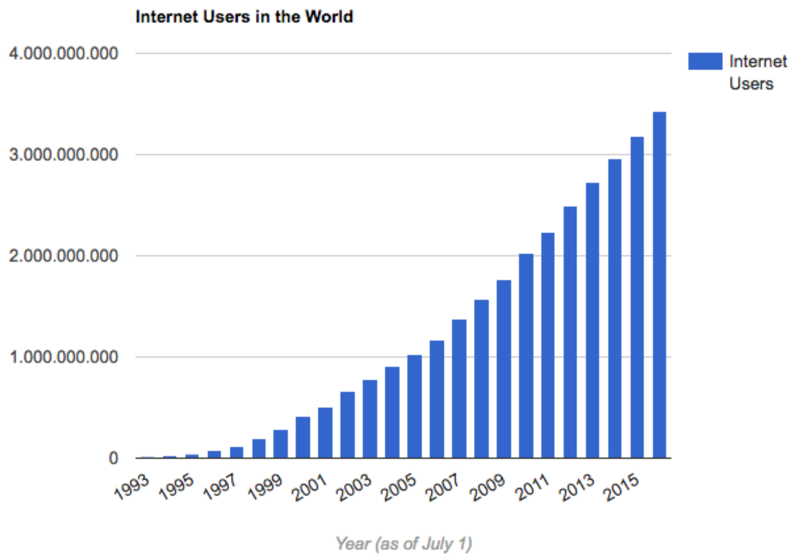


Figure 3.2: Growth of Internet users in the World in the last twenty years, from: http://www.internetworldstats.com/

Several research (Kunny (1997); Rosenzweig (2003); Brügger (2012c) among others) have already remarked on the great impact that the ephemerality of the web will have on the way knowledge produced in the digital age will be shared and accessed by next generations (see in particular Perkel (2015) on the impact of broken link in scientific publications on limiting the reproducibility of the results). As it has been already remarked, in opposition to the fact that "paper survives benign neglect for a long time" (Davis, 2014):

> The life cycle of most web pages runs its course in a matter of months. In 1997, the average lifespan of a web page was 44 days; in 2003, it was 100 days. Links go bad even faster. A 2008 analysis of links in 2,700 digital resources—the majority of which had no print counterpart—found that about 8 percent of links stopped working after one year. By 2011, when three years had passed, 30 percent of links in the collection were dead. (LaFrance, 2015)

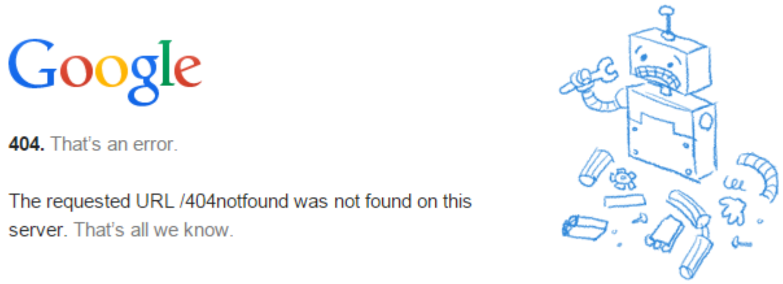
In addition to this, while some type of pages disappear more frequently than others (e.g. social media messages compared to official statements on administrative websites), those that do survive tend to change frequently (Dougherty et al., 2010). For example, articles on newspapers[13] as well as official administrative pages have been modified without a specific mention. The most famous example regards an update on the website of the White House in May 2003. The original title of a piece of news on the status of the War in Iraq was *President Bush Announces Combat Operations in Iraq Have Ended*, while in December 2003 it was modified - without a note reporting the change - to *President Bush Announces Major Combat Operations in Iraq Have Ended*[14].

Another famous example has been presented in a recent article on the *New Yorker*:

> Malaysia Airlines Flight 17 took off from Amsterdam at 10:31 a.m. G.M.T. on July 17, 2014, for a twelve-hour flight to Kuala Lumpur. Not much more than three hours later, the plane, a Boeing 777, crashed in a field outside Donetsk, Ukraine. All two hundred and ninety-eight people on board were killed. The plane's last radio contact was at 1:20 p.m. G.M.T. At 2:50 p.m. G.M.T.: Igor Girkin, a Ukrainian separatist leader also known as Strelkov, or someone acting on his behalf, posted a message on VKontakte, a Russian social-media site: "We just downed a plane, an AN-26." (An Antonov 26 is a Soviet-built military cargo plane.) The post includes links to video of the wreckage of a plane; it appears to be a Boeing 777. [...]
> [Less than three hours later - A/N] Strelkov's VKontakte page had already been edited: the claim about shooting down a plane was deleted. (Lepore, 2015)

[13] This has been the topic of my master thesis, and it has been presented in an article on the journal *Diacronie. Studi di Storia Contemporanea*, Nanni (2013).

[14] As described in Benvenuti and Morriello (2006).



404. That’s an error.

The requested URL /404notfound was not found on this server. That’s all we know.

Figure 3.3: The 404 status code message received when using the Google Chrome browser.

3.2.1 Archiving the Web

Given all these issues, the first pioneering project focused on the preservation of web materials for future studies started already at the end of 1996 under the leadership of Brewster Kahle. His already-by-then utopian purpose was of archiving the web in its entirety (Lyman and Kahle, 1998).

During the last two decades, the project Kahle presented under the name of “Internet Archive” has become a fundamental work (and an example for many others) for the preservation of our digital past. Internet Archive’s crawlers[15] started acquiring and preserving snapshots of webpages in November 1996, conducting an endless fight with the never-ending growth and continuous change of the web.

In the following years, many other web archive projects have been developed, often with a more specific National focus, such as Pandora in Australia

[15] A Web crawler is an Internet software application (bot) that systematically browses the World Wide Web, typically for the purpose of Web indexing.

(1996)[16], the UK Web Archive (2004)[17], Netarkivet in Denmark (2005)[18] and the Portuguese Web Archive (2007)[19]. All these projects are public initiatives (while the Internet Archive remains a nonprofit private digital library (Momack, 2003)); to know more regarding web archive initiatives, the community is constantly updating a dedicated Wikipedia page[20].

In 2003, the Internet International Preservation Consortium (IIPC), was founded at the National Library of France[21]. During the last decade the consortium has coordinated national and international efforts to preserve internet contents for the future. Today, with a General Assembly meeting every year since 2011 and organisations joining from 25 different countries, the IIPC has become the leading guide of world wide born-digital preservation projects.

The Past of the Italian Web Sphere

Currently, the National Libraries of Florence and Rome are not part of the IIPC and no project with the specific purpose of preserving the Italian websphere exists. In 2006, thanks to the effort of the project "Crawler" (Tammaro, 2006; Bergamin, 2012), which was supported by the "Biblioteca Digitale Italiana" (Italian Digital Library), Italy cooperated with the European Archive Foundation (now called "Internet Memory Foundation") and conducted its first wide-spread crawling of the ".it" domain[22]. However, after this no other projects were conducted and the only part of its national websphere which has been constantly crawled and preserved are the Ph.D. theses repositories of Italian universities (Vignocchi et al., 2010), thanks to the ac-

[16] `http://pandora.nla.gov.au/`

[17] `http://www.webarchive.org.uk/ukwa/`

[18] `http://netarkivet.dk/`

[19] `http://arquivo.pt/`

[20] `https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives`

[21] `http://netpreserve.org/about-us`

[22] This was conducted between May and June 2006, the snapshots are available here: `http://collection.europarchive.org/bncf/`

tivities of the Magazzini Digitali project[23] (Bergamin and Messina, 2012). For these reasons, researchers interested in diachronically studying the Italian web sphere could currently rely only on the snapshots of websites preserved by international web archive initiatives, such as the Internet Archive.[24]

3.2.2 Retrieving Information in Web Archives

During the last twenty years, the Internet Archive has preserved almost 500 billion web pages, and currently has an extended collection of 25 petabytes of data (Srinivasan, 2016). Since 2001, this collection has become available for research on a platform called the Wayback Machine[25], which, until the end of 2016, has provided only a URL search tool. By using this tool it is possible only to retrieve web pages from a given URL and keyword search is not allowed. The results of a URL query (such as `http://www.nytimes.com/`) are displayed in chronological order on a calendar, with an associated preservation date.
Additionally, the Internet Archive offers a series of APIs to interact in advanced ways with the Wayback Machine[26]: for example it is possible to send extended URL-based queries, such as “only show at most one capture per hour of all preserved pages, given a specific domain”. While the Internet Archive has only now started supporting full-text search on their web archived collection in a demo introduced for its 20th anniversary[27], other initiatives are offering it, such as the UK Web Archive (which also offers N-gram search and a related visualisation tool[28]) and the Portuguese Web Archive.

[23] `http://www.depositolegale.it/`

[24] Or, as it will be described in the next part of this thesis, in some specific cases on snapshots archived in other national web archives.

[25] `https://archive.org/web/`

[26] `https://archive.org/help/wayback_api.php`

[27] `https://web-beta.archive.org/`

[28] `http://www.webarchive.org.uk/ukwa/ngram/`

3.3 History and the Web

As it has been remarked in the previous section, even if the web has kept changing, a huge amount of these born-digital sources have been archived during the last twenty years, thanks to the effort of digital archivists and academics interested in their preservation (Lyman and Kahle, 1998; Brügger, 2005; Gomes et al., 2011). Media studies scholars have been the first intensively exploring their potential for different research areas, such as political science analyses (Foot et al., 2003; Ben-David, 2011), journalism (Weber, 2014) and internet studies research (Dougherty et al., 2010; Ankerson, 2012; Hale et al., 2014; Ben-David and Huurdeman, 2014). In very recent years, historians started evaluating the reliability of web archive materials (Milligan, 2012) and their usefulness as primary sources for studying the development of online communities (Maemura et al., 2016), the Yugoslav conflict (Ben-David, 2016) and North African immigration in France (Gebeil, 2014).

We can identify two different ways of considering web archive materials as primary sources in historical studies. The first, described for example by Brügger (2012c), has its roots in the fields of media and Internet studies and aims at examining the web of the past by contextualising and understanding changes in layout, structure, content and use. In 2010, Brügger edited the first book on the topic (Brügger, 2010), and the title of the volume, *Web History*, clearly highlighted the research topic of the community, where the web and websites have become objects of study (Brügger, 2009). A similar focus emerges when reading the objectives of a new journal titled "Internet Histories"[29], recently launched by Niels Brügger and others.
The second way of considering these materials as primary sources has, instead, a wider spectrum of applications. As already remarked by Rosenzweig (2003), born-digital materials (with their scarcity and abundance) will have an impact on the entire historical profession; for this reason, web archive materials could soon become new sources for political as well as social, cultural and economic historians. Recent works, such as the already mentioned stud-

[29] `http://www.tandfonline.com/loi/rint20`

ies conducted by Anat Ben David (2016) and Sophie Gebeil (2016), already show how topics such as the Yugoslav conflict and North African immigration in France could be studied fruitfully from a web archive perspective.

In both areas, researchers have explored the potential of computational methods, such as text mining and network analyses, for extracting information from large web archive collections. Examples are presented in Milligan (2012), Hale et al. (2014) and Holzmann et al. (2016b).

3.3.1 Reliability of Web Archive Materials

A substantial number of articles focused on the reliability of web archives and web archival sources has been published in the last years. Thelwall and Vaughan (2004) studied country balance in the Internet Archive, Howell (2006) analyzed how to use the Internet Archive in research and Murphy et al. (2007) established the Wayback Machine as a valid tool for identifying, among other information, web page contents, "website age" and updates. In contrast to Murphy et al. (2007), Brügger (2008) highlighted a series of problems in web preservation and underlined the need of what he called a "web-philology" to deal with the reconstruction of partially archived websites. More recently he also defined the resources preserved in web archives as "reborn-digital materials" (Brügger, 2012c), which must be considered as different objects compared to the originals.
Along the same lines, Dougherty et al. (2010) summarized the state of the art of web archiving in relation to researchers needs. Ankerson (2012) remarked that web historians need to "consider broadcast[ing] historiography scholarship that grapples with questions of power, preservation, and the unique challenges of ephemeral media". Finally Ben-David and Huurdeman (2014) explored how to go beyond current limitations of search tools in web archives and, with others (Samar et al., 2014; Huurdeman et al., 2014), employed a new approach to analyse hyperlinks in web archives in order to deal with the reconstruction of the unarchived web.

3.3.2 Computational Web History

The adoption of distant reading solutions that has consistently grown in digital humanities has also reached the web archive community. General overviews of the application of these methods on web archives are presented in Dougherty et al. (2010), Hockx-Yu (2014) and Maemura et al. (2016). In particular, in order to deal with these huge collections, network analysis solutions have been adopted while conducting corpora exploration, for example by Milligan (2012), Brügger (2012a), Hale et al. (2014) and (Ben-David, 2016).

Text mining solutions are way less adopted, giving the difficulties of processing large web archives; in order to address this current issue, research groups are developing solutions for enhancing distant reading over web archive collections; some first examples are ArchiveSpark (Holzmann et al., 2016a) and Warcbase[30]. Results of these analyses could be found in Milligan (2012), where the author applied LDA topic modeling to explore Canada's Digital Collections of archived websites and in Gorsky (2015), where the author conducted word-frequency analyses for studying an argument about the contradictory rationale for public health policy under New Labour in the UK web archive (1996-2009).

3.4 Beyond Internet Studies

Both Foot and Schneider (2010) and Brügger (2012b) recently used the term "web historiography" to remark on the need of a theoretical and methodological discussion around the way researchers (and especially historians) are beginning to study the web and adopt it as a primary source. However, as this is an extremely new research topic in the digital humanities environment, practitioners are still trying to establish their position at international conferences and on academic journals[31]. As Milligan (2016b) recently described,

[30] https://github.com/lintool/warcbase

[31] The first journal on the topic, presented in 2016, is Internet Histories http://explore.tandfonline.com/page/ah/internet-histories. However, by looking at

researchers are currently meeting at a huge variety of venues, from digital libraries / web archives meetings (such as the Joint Conference on Digital Libraries (JCDL) and the International Internet Preservation Consortium (IIPC) annual meeting) to traditional history conferences (American Historical Association), from digital humanities venue (the main Digital Humanities conference)[32] to computer science meetings (the Web Science and the World Wide Web conferences). While this highly interdisciplinary situation has allowed the research community to build a large variety of collaborations, the absence of a venue that could be called "home"[33] does not guarantee the possibility of structuring a consistent methodological and theoretical discussion between its members[34].

Due to the above presented difficulties of establishing the identity of this new research area in the academic environment, researchers are currently trying to consolidate the theoretical and methodological debate around two pillars: the *history of the web* as the main research focus of the community and *web archives* as the most important primary source.

While these choices are understandable given the complex context in which this community has risen, the goal of this thesis is to move beyond this situation, in particular by remarking how the advent of born-digital primary sources will influence any possible research area in contemporary history, from political to social history to the history of science (and not only the studies focused on the web of the past) and how web archives will be only one of the new primary source collections at our disposal[35].

the suggested topics, the journal seems to maintain a strong focus on internet studies.

[32] I have been the main organiser of the first panel on Web Historiography at DH2016.

[33] And additionally the fact that the closest thing to a "home" for web historians currently is IIPC - traditionally a conference for digital archivists where history panels are always on a side.

[34] The meeting organised by Resaw (`http://resaw.eu/events/international-conference-aarhus-june-2015/`) has been the first conference completely dedicated to web historiography and an extremely enriching opportunity for its community.

[35] Additionally, web archive materials are not the easiest resources to obtain, for a researcher who is not affiliated with any national library and who is not involved in any web archive project.

In order to remark on this, from the next chapter on, we will highlight both the methodological impact and the usefulness of adopting born-digital sources in a historical research focused on a very traditional topic in the history of science, namely the history of academic institutions. Additionally, in the last part of the thesis, we will highlight how the current theoretical and methodological discussion on the use of born-digital materials as historical sources should go beyond the fields of internet studies and digital humanities, by embracing the larger current discussion on the present and future of historiography.

Chapter 4

Case Study: Web Materials as New Academic Sources

In via Petroni si svegliano,
preparano libri e caffè.

In this chapter I introduce the case-study of this thesis. First, I offer an overview of established methodological approaches and research topics related to the study of the past of universities. Next, I describe how born-digital sources will have an impact in this research area, by presenting new challenges but also the possibility of addressing new research questions. This chapter relies on Nanni (2015) and Nanni (2017) and connects Part I of this thesis to Parts II & III.

4.1 Introduction

Considering academic institutions as political, economic and social actors has attracted the interest of many research communities. In this chapter we start by offering an overview of three of these research areas, highlighting which are the established research topics and methods that the communities have addressed. Next, it is also described how this Ph.D. thesis work is related to previous research efforts conducted at the University of Bologna on studying the past of academic institutions. This overview will support us in presenting, in the second part of this chapter, the central argument of this thesis, namely the fact that born-digital sources will condition the research practices of all the examined communities, in particular for what concerns the processes of collecting, analysing and selecting evidence.

4.2 Studying the Past of Universities

Higher education has been a fundamental need of several advanced civilisations. Being able to train future generations of political, spiritual and military leaders has characterised the rise of athenaeums and lyceums of ancient Greece as well as the Roman empire and several dynasties in ancient China. However, only in medieval Europe we can assist to the full recognition of what we define now as "university", which Perkin describes as "a school of higher learning combining teaching and scholarship and characterised by its corporate autonomy and academic freedom" (Perkin, 2007).
Since the 12th century this corporate institution has emerged, evolved and survived through wars, revolutions, crises and societal dismantles. From their original status of schools of higher learning in the middle age to the cosmopolitan institutions communicating all around Europe in Latin, from the advent of the French *grandes écoles* to the combination of teaching and research that German universities have fostered through the Industrial Revolution, from instruments of National propaganda to the current status of central pillars of the post-industrial society[1], universities have shaped (and have been shaped by) our society during the last eight hundred years.

This specific type of institution has already attracted the interest of many historians, who wanted to understand how its power, role and influence changed over time, especially in relation to other actors, such as the city, the church, the national government (Brockliss, 1978; Macleod and Moseley, 1978). The massive four-volume book series "A History of the Universities in Europe",

[1] While at the same time always striving for maintaining autonomy and academic freedom.

commissioned by the European University Association, edited by Hilde de Ridder-Symoens and Walter Rüegg and published between 1992 and 2011, offers an unprecedented comprehensive overview of how universities have changed what they have taught and researched, how they have been institutionalised and how they have interacted with the society.

While universities have been examined largely as institutions that evolve in relation with each other, they are also the physical place were scientists and scholars conduct their work. For this reason, during the last century academic institutions have been also studied by historians, philosophers and anthropologists of science and technology, interested in understanding how STEM (Science, Technology, Engineering and Mathematics) were taught and studied in universities (Fox and Guagnini, 1993), how scientific knowledge moves to the private sector (Mahoney, 1988; Guagnini, 1988), how political, economical and social actors have influenced scientific research in academia (Pancaldi, 2006) and how scientists work in their laboratories (Latour and Woolgar, 1979; Pancaldi, 1993b; Worboys, 2011).

A third, completely different, perspective on how universities have been studied in recent years has been offered by the scientometrics community, which goal is to measure and analyse the impact of publications, journals and institutes, and to produce indicators that would be adopted in policy and management contexts. The use of metrics such as citation and co-citation measures (Van Raan, 1997) has attracted a huge attention from university administrations, politicians, sociologists and quantitative historians (for further discussion, see De Bellis (2009)); additionally, the quantification of the scientific output as a measure of evaluating and comparing academic institutions has had a huge impact on their recent past, influencing hiring strategies as well as the pursuit of certain research topics and practices.

4.2.1 Established Methodologies

Different researchers have been studying universities and their past with different methodologies. In the next few paragraphs, we intend to examine which are the sources and methods at current disposal and how they have been adopted to study the recent history of academic institutions. This will guide us towards remarking how born and reborn-digital sources will open new challenges and new possibilities for the different fields.

History of Universities. Historians of higher education generally adopt a large variety of primary and secondary sources in their works, from university-archive materials such as matriculation and graduation statistics (Brockliss, 1978; Macleod and Moseley, 1978) to academic dissertations (Richardson, 1999), from public reports (Committee, 1985; de Wied et al., 1991) to large scale statistical analyses previously conducted in closely-related areas (see for example Fielden et al. (1973); Friedberg and Musselin (1987)). Based on these data, researchers have described and drawn conclusions on the recent history of universities on a large variety of topics, such as the way universities have managed resources, the way the admission process has changed before and after 1970, and how sciences and humanities have been taught and studied.
While a large number of these researches relied on statistical analyses, their results often remain a piece of the puzzle in larger historical narratives. For example, if we consider the chapter "The Biological Sciences" (Macgregor, 2010), in the fourth volume of "A History of University in Europe" (Rüegg, 2011), we can notice how the author builds the narrative on quantitative studies such as "Milestones and rates of growth in the development of biology" (Glass, 1979).

Science and Technology Studies. Researchers of science and technology focus on the recent history of universities with a few broader focuses. First of all, by studying the way scientists work (Latour and Woolgar, 1979; Latour, 1987), this community also examined the ways academics interact with the institutions where they conduct their research. Secondly, by studying the relations and mutual influence between science and society, researchers have aimed at understanding the social role of academics (Weil, 2002), at highlighting the "myth of neutrality" in social science research (Scott et al., 1990) and at openly criticising the pre-assumed objectivity of research practices. Especially this last topic has provoked an intense reaction by the scientific community (Gross and Levitt, 1996), which gave the rise in the Nineties to the so-called science-wars (Parsons, 2003). In addition to this, the community also directly participated to science policy discussions (Hoppe, 2005). As Hoppe (2005) precisely remarked, "scientists, like everyone else, are motivated by self interest, pride, profit, power and the anticipation of glory and public heroism", which means that the way scientists interact with institutions and society are through social and political acts, that need to be studied and comprehended.

For what concerns methodologies adopted in these studies, a solid combination of qualitative approaches from critical sociology, symbolic interactionism and ethnomethodology have been the key components of the field. Examples are participant observation, interviews, focus groups and close readings of primary sources. As Wyatt et al. (2015) recently remarked "quantitative methods, based on numerical data and/or statistical analysis of large-scale surveys, experiments, national censuses and, more recently, data and information visualizations, are less visible within STS".

Scientometrics. The goal of scientometrics (Van Raan, 1997) is to establish in a quantitative way the quality of academic research and academic institutions (Dill and Soo, 2005). While the field was established in the 60s, from the 90s scientometrics measures have gained increasing importance in academia, especially when dealing with budget decisions such as funding applications, promotions, prizes and tenure. Today, scientometrics is actively influencing the way researchers plan, conduct and promote their research (Bornmann and Leydesdorff, 2014).

In the field of scientometrics, the use of bibliometric approaches (such as direct citation and cocitation) and network analysis techniques is one of the most common solutions for the evaluation of scientific output (Lu and Wolfram, 2012). More recently, a series of publications have focused on the use of word-based and topic-based methods in order to conduct scientometric studies. Dietz et al. (2007) used LDA topic models to quantify the impact that research papers have on each other. Gerrish and Blei (2010a) showed that LDA is able to identify a qualitatively different set of relevant articles, when compared to traditional citation-count metrics. The use of text mining approaches for examining scientific publications have been applied in order to study academic fields (Anderson et al., 2012) and to quantify how ideas spread through publications (Hall et al., 2008).

With the goal of highlighting future advancements in the use of text mining methods for improving scientometrics research, since 2012 Petr Knoth has organised the International Workshop on Mining Scientific Publications (WOSP). The workshop brings together people who *a)* are interested in analysing and mining datasets of scientific publications, *b)* develop systems that enable such analyses and *c)* present novel measures for evaluating the way research has being conducted[2]. All papers presented at WOSP have been published on the digital library journal D-Lib, in a solid attempt of bridging digital libraries, text mining researchers and the scientometrics community.

As it has been described in the previous pages, the recent past of universities can be studied under different perspective, with different goals and combining completely different methodologies. In the next section, the focus will be dedicated to the University of Bologna, the most recurrent case study of this Ph.D. thesis.

[2] The more recent Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016) has a similar focus.

4.3 Studying the Past of the University of Bologna

The University of Bologna is considered to be the world's oldest university in terms of continuous operation. Its nine-hundred years old roots lead to the figure of Irnerius, a jurist and instructor of law of the XII Century (Brizzi et al., 1988)[3]. The university has been a fundamental element of Bologna for more than nine centuries, making the city one of the most important academic centres in Italy and Europe.

The university, in its first centuries, has been mainly focused on the study of law and the teaching of the Justinian's Code, while its political activities were dedicated to maintaining independence from the local political power. In the 14th century the university extended its schools to scholars interested in medicine, philosophy, arithmetic, astronomy, logic, rhetoric and, in the second part of the century, theology. In this period Dante Alighieri, Francesco Petrarca and Guido Guinizelli, among others, spent periods of study in Bologna. In the following centuries the university extended again its interest to the teaching of Greek, Hebrew and of natural sciences. Revolutionaries were the "Tractatus de immortalitate animae" by Pietro Pomponazzi (1516) and the writings of Ulisse Aldrovandi[4].

Between the XVI and the XVII century Bologna has risen also as a highly relevant centre for the study of medicine, thanks to Gaspare Tagliacozzi and, in particular, to the extraordinary work on the use of the microscope in medicine by Marcello Malpighi. In more recent centuries, Bologna contributed to the industrial revolutions thanks to the work of Luigi Galvani and has experienced another grown after the unification of Italy, with the academic contributions of Giosué Carducci, Giovanni Pascoli, Augusto Righi, Federigo Enriques, Giacomo Ciamician, Augusto Murri and, in more recent years, of Umberto Eco, among many others.

[3] See also: `http://www.unibo.it/en/university/who-we-are/our-history/university-from-12th-to-20th-century`

[4] Aldrovandi was behind the creation of the botanical garden of the University of Bologna, one of the first in Europe.

The study of specific aspects of the history of this institution has already offered to historians a unique possibility of digging deeper in the relationship between the university, its large students' community and the city of Bologna itself (Bellettini and Tassinari, 1984; Barbagli et al., 2009); moreover this kind of research has also guaranteed a better understanding of its key historical role in the Italian academic (Mazzetti, 1848) and political ecosystem (Baldissara, 1994; Salustri, 2010). Several sources have been used to trace its past, starting from textual documents preserved in its archives (Romano, 2007) to its collection of over seven hundred portraits (Gandolfi, 2011).
At the University of Bologna two different research groups have been constituted for studying the past of academic institutions, and in particular the history of the Alma Mater.

4.3.1 The Research at CIS and CISUI

Currently, the University of Bologna has two centres specifically dedicated to the history of academic institutions, namely the International Centre for the History of Universities and Science (CIS) and the Inter-university Centre for the History of Italian Universities (CISUI). These research groups, constituted in the Nineties, continue the tradition of the oldest centre in Italy for the history of universities, the Centre for the History of the University of Bologna, founded in 1906.

The reason behind the foundation of these centres has a long history. In 1888, the University of Bologna decided to celebrate its eighth-centenary anniversary, having as a reference date the year 1088, which was conventionally accepted as the beginning period of the activities of the university. The ceremony[5], was overviewed by Gousué Carducci and was a great event for the university and for Bologna, affirming the institution as the Alma Mater of all other universities (Brizzi et al., 1988). Almost a hundred years later, Fabio Roversi Monaco, candidate for the position of Rettore of the University, be-

[5] Inspired by the ceremony that the University of Heidelberg organised two years before

gan to look for a National support from the Italian Parliament in order to organise the 'Ninth Centenary Celebrations' of the University of Bologna. He was able to get a specific law approved (LEGGE 12 aprile 1989, n. 131) that - among other things - supported the creation of a research centre dedicated to the history of universities, science and institutions of high-culture. Walter Tega, professor at the University of Bologna, was the initial supporter of the organisation of this centre; however, in the following years another professor, Giuliano Pancaldi, took the lead.

Pancaldi, a historian of science, started his work by organising in 1988 a two-weeks international summer school on the history of science in Bologna, with more than 20 researchers. Then, with the initial support of Roversi Monaco, Pancaldi created in 1991 a small research centre at the department of Philosophy focused on the history of science and universities. The first volume published, "Le Universita e le Scienze. Prospettive storiche e attuali" (Pancaldi, 1993a) and the pioneering Online Iconographic Archives[6], clearly highlights the focus of the group, as a bridge between the history of higher education and the history of science. Among the board of directors there was Gian Paolo Brizzi, professor of modern history with a solid research focus on the history of Italian universities. After a few years, Brizzi started a second research center completely dedicated to the history of Italian universities, this time at the Department of History and in collaboration with other institutions such as the universities of Padova, Messina, Sassari and Torino.

In the two following decades CISUI and CIS highly differentiated their research topics, with the first becoming a focused research centres for the history of Italian academic institutions (Negrini, 1998; Dröscher, 2002) and the second moving more and more towards science, technology studies, as can be noticed by the doctoral dissertation of some of its researchers (Crocetti, 2011; Panajoli, 2012; Arnaudo, 2013; Iori, 2013; Banerjee, 2013). Nevertheless, thanks to a Ph.D. program at CIS focused on interdisciplinary interactions between science, technology and the humanities, doctoral students have

[6] `http://137.204.24.205/icon/home.html`

dealt with the relationships between universities and politics, economy and society (Serafini, 2011; Parolini, 2013; Piazza, 2013; Kleinveldt and Booysen, 2015), adopting science and society as well as scientometrics methodologies[7].

4.4 Academic Born Digital Documents

As we remarked earlier, the availability of digital materials is about to have an impact on the research topics and practices of all communities working on the recent past of academic institutions.
If we consider first the work conducted by the scientometrics community regarding the recent past and present of academic institutions, we will notice that so far this research area has mainly adopted bibliometric information for evaluating the research output of universities. However, in the recent years, the scientometrics community has also begun to deal with a huge new variety of born-digital documents such as bachelor, master and doctoral theses (Ramage, 2011) as well as the content of scientific publications (Dietz et al., 2007) and grant proposals (Nichols, 2014). All these resources are generally offered in unstructured texts directly on the web, thanks, for example, to the work of academic digital libraries[8]. While these materials are way more difficult to analyse than structured bibliometrics data (consider for example the natural ambiguity of the language adopted), their prompt availability is already offering new ways and more extensive perspectives for the evaluation of the outcome of academic institutions (as precisely remarked by the examples offered by Nichols (2014); Bromham et al. (2016)).

If we examine now the work conducted so far by STS and science and policy researchers and its interactions with the past of academic institutions (Guston and Keniston, 1994), we can imagine which new perspectives born-digital

[7] The same interdisciplinary environment has been the place where, in October 2013, I started working on my doctoral research, focused on understanding how to continue studying the history of academic institutions now that born-digital sources are increasingly replacing traditional materials.

[8] See for example the variety of materials available on the University of Bologna Digital Library: `http://www.sba.unibo.it/it/almadl`

materials are about to offer. The prompt availability of syllabi (Cohen, 2011) will support researchers in understanding how/whether specific topics have been taught to students (e.g. do students of neurobiology study Obama's BRAIN initiative?[9]; are the more-technical aspects of the project Manhattan presented to young historians?)[10] as well as examining the cultural and political reasons behind widespread adoption of specific textbooks.
Another perspective on academia as a social environment could be obtained by observing the dynamics in the discussions between academics on social media platforms. An example of this type is the debate generated after the publication of the tool Syuzhet[11], which we already mentioned in the previous chapters.
In addition to this, many other resources can be studied in order to address STS topics. Electronic lab notebook (Giles, 2012) will become a new fundamental source for studying the "laboratory life" (Latour and Woolgar, 1979); examining how the institution presented and advertise itself online could guide researchers in exploring which are the aspects that are highlighted and which are the ones that are hidden[12].

While the scientometrics and the STS communities are already directly benefiting from the availability of born-digital sources, this thesis imagines a more long term impact of these materials over the field of history of higher edu-

[9] `https://www.whitehouse.gov/share/brain\-initiative`

[10] A first attempt in this direction comes directly from the digital humanities community with the work of Cohen (2005), published on the Journal of American History and focused on understanding the role of textbooks in teaching U.S. history.

[11] `https://storify.com/clancynewyork/contretemps-a-syuzhet`; If we go beyond the technical debate, social dynamics emerge, for example, a (perceived by third readers) gendered diminution of the woman who raised the critique, Annie Swafford, by the author of the work Matthew Jockers: `https://twitter.com/clancynewyork/status/573177096362692609`

[12] For example, it is interesting to see how the University of Bologna presents itself on Youtube (`https://www.youtube.com/watch?v=JOJeWse5wKE`), which messages are conceived to the viewers and which images of the city are presented (such as San Giovanni in Monte, Piazza Santo Stefano) and which other are voluntarily excluded (not even a frame is dedicated to the highly controversial and at the same time extremely characteristic "zona universitaria")

cation. As it has been described in the previous sections, research questions in this area involve topics such as the role of academic institutions in nation building as well as on the mutual influence with political actors, on the experience and life of students as well as on the relationship of institutions and professional realities. Another topic that has attracted significant attention is the examination of universities as institutions of higher education, where the manner in which topics are taught and education is provided are also highly influenced by several political, economic and social factors. In this thesis we argue that born-digital sources such as syllabi (Cohen, 2011), bachelor, master and doctoral theses (Ramage, 2011), academic websites (Holzmann et al., 2016b) and their hyperlinked structure (Hale et al., 2014) will not only impact the scientometrics and the STS communities, but will become relevant new materials that historians of academic institutions will employ to continue addressing solidly established research questions.

4.4.1 University Websites as Objects of Study

In the previous paragraphs we remarked on the potential of born-digital sources for the different research communities that are studying the recent past of academic institutions. It is also important to note that, while the number of born-digital sources related to the recent history of universities has drastically icreased over the recent years, their traditional counterpart are usually not produced, archived or simply not promptly available for research anymore[13].

The starting point of Part II of this thesis relies on the fact that this sudden and rapid transition from analogue to born-digital materials combined with the ephemerality of this new type of documents (as described in previous

[13] An example could be the way researchers could currently consult materials such as syllabi and theses offered and written at the University of Bologna. Syllabi exist only in digital format while theses are currently archived both in digital and in analogue format, but only the digital copy is directly available for consultation (at: `http://www.sba.unibo.it/it/almadl`). Additionally, also the National libraries of Florence and Rome are involved in the preservation of doctoral theses, but only in digital form (to know more see: `http://www.depositolegale.it/`

chapters) threatens to leave future researchers of academic institutions without documental sources. To better understand the potentially huge impact of this issue, let us imagine to be for a moment in 2088, for the millennial anniversary of the University of Bologna. Which documents could we adopt to understand how this institution has changed since the last centenary?

In order to address this issue with specific solutions, Niels Brügger defined five levels of analyses for researcher interested in adopting web materials as a primary source:

> a) the web element (e.g. an image or a video); b) the web page (what is seen in a browser window, e.g. the front page); c) the website (interrelated web pages); d) the web sphere (web activity in relation to a theme or an event, e.g. political election); and e) the web as a whole (phenomena transcending the web, e.g. the web's content in its totality) (Brügger, 2012c)

In the next part of this thesis, we primarily focus on considering academic websites as objects of study. We have chosen the website as the main focus of this research, as it fully represents the digital alter ego of a university, the online "main door" for its entire national and international community. On academic websites a large variety of information could be collected, from datasets of scientific publications (such as doctoral dissertations) to overviews of research teams, from summaries of national and international collaborations to detailed descriptions of funds and project budgets. Additionally, academic websites highlight the role that online communication has played in the interactions between the academic institution and its large and variegated community.

In the next part of the thesis, we firstly explore how the born-digital nature of academic websites makes them very difficult to archive, preserve and collect in their entirety for historical research. Specific case studies will be four Italian university websites, including the University of Bologna. This issue forces historians that intend to use these materials in considering a more interdisciplinary approach, that combines the traditional historical method with solutions from the field of internet studies, where researchers have been dealing with born-digital documents for more than a decade. This will be the focus of the second chapter of the next part. Finally, through a series of case study, we will remark on the importance of approaching these new sources in a highly critical way, always examining their potential and limits; additionally we will show how the conducted work on reconstructing the University of Bologna website has brought to the surface new collections of documents on the recent past of this academic institution.

Part II

How to Deal with Scarcity

Chapter 5

The Status of Archived Snapshots of University Websites

There's a darkness
on the edge of town.

This chapter is based on the findings presented in the paper "Historical Method and Born-Digital Primary Sources: A Case Study of Italian University Websites", which has been published in a special issue of the journal Officina della Storia dedicated to the topic "From the history of the media to the media as sources of history". I wrote this paper during my visiting period at the Centre for Internet Studies of the University of Aarhus, under the host supervision of Prof. Niels Brügger. The goal of this chapter is to examine the reliability of the archived versions of university websites for the study of the recent histories of Italian academic institutions. Starting from a pool of twenty university websites, four case studies are discussed into detail, related to the websites of the Polytechnic University of Turin, the University of Rome 2, the University of Trento and the University of Bologna. In the next pages, a series of specific preservation issues are described and precise solutions are defined, discussed and adopted.

5.1 Introduction

Italian universities have a strong tradition of innovation in communication technologies. On the 30th of April 1986, the University of Pisa activated its connection with ARPANET, thereby making Italy the third country in Europe to be "online", after Norway and the United Kingdom (Chiapparini, 1995; Valentini, 2011) and in the same years Italy also became the first European country on BITNET (Chiapparini, 1995). In 1987 the National Research Council (CNR) registered the first ".it" domain[1] and in the early 90s the Center for Advanced Studies, Research and Development in Sardinia (CRS4) created what they declared to be the first Italian website (www.crs4.it), which was the second one in Europe[2]. Another fruitful digital environment bloomed in Bologna in the early 90s, thanks to the collaborations between the University, the Municipality, several small IT companies and CINECA (a non-profit consortium of academic institutions, currently composed of almost 70 Italian universities). Here in 1995 "Iperbole" was created, one of the first Civic Networks in the world (Romagnoli, 1994; Chiara, 1998).

With this early interest of academia in extending its presence on the Internet, a robust and active web community was also born. In fact, even if in 1994 the number of Internet users in Italy were just 15.000[3], already in those years a consistent number of people were curious about this new technology, as can be noticed from the articles[4] published by some of the most important newspapers, where pieces of advice regarding the Internet were given and common doubts were clarified.

[1] As described on its website: `http://www.iit.cnr.it/servizi/registro\_it`

[2] Pinna A., "Soru: un incontro con Rubbia, così nacque il web in Sardegna", Il Corriere della Sera, 28/12/1999, p.24.

[3] See Romagnoli (1994), this information is offered in a box.

[4] Miccoli, M. " Cosa serve per collegarsi ad Internet", *La Repubblica*, 03/10/1994; Merciai, S. A., "Scriveteci siamo su Internet", *La Stampa – Tutto Scienze*, 15/02/1995, p. 2; *Il Corriere della Sera*, "La finestra che cambierà il personal", 24/08/1995, p. 18.

Today, even though less than 65% of Italians use the Internet[5], the country nevertheless has more than 30 million active users, which are the target and also the producers of a vast amount of daily information.

As it has been already remarked, websites keep changing without leaving any trace (see for example Brügger (2005)); for this reasons web archives - in their attempt of preserving materials available on the live web - are one of the necessary starting points of any kind of historical research that intends to use websites as primary sources. In the next pages, we focus on critically assessing the reliability of the resources at our disposal in web archives for studying different aspects of the recent past of Italian universities, highlighting their potential and limits.

5.2 Collecting Snapshots

Despite an important collaboration between the National Libraries of Florence and Rome that started in 2006[6], currently an Italian Web Archive is still not available for research purposes. If we consider the forty most important web archival projects (Gomes et al., 2011; Niu, 2012), only three of them have an international spectrum; therefore only these three could have consistently preserved Italian university websites. They are the Internet Archive (created in 1996), the Internet Memory (founded in 2004 under the name of European Archive) and the California Digital Library Web Archiving Service (WAS, 2005). In this study we used mainly the Internet Archive, as it presents an enormous amount of snapshots compared to the Internet Memory, and because Italian academic institutions have not been preserved by WAS.

[5] `http://thenetmonitor.org/countries/ita/access`

[6] `http://www.bncf.firenze.sbn.it/pagina.php?id=212`

5.2.1 Using the Wayback Machine

Currently the Internet Archive offers the largest collection of preserved pages (Lepore, 2015), through its Wayback Machine. Moreover, in the last decade web archive materials have been already used by media studies scholars, historians and political scientists for a large variety of researches (Foot et al., 2003; Dougherty et al., 2010; Milligan, 2012; Hale et al., 2014; Ben-David, 2016). These different studies have revealed the great potential of web archives in offering new/different perspectives on our recent past.

During our diachronic analysis of twenty Italian university websites, we initially focused on studying their homepages over time, as they offer a general overview of the structural organisation of the website and highlight which subsections were considered highly relevant at specific points in time (for example, the ones listed in the sidebars). In this analysis we identified, in a coarse-grained fashion, all major changes in the structure and the layout of the homepages. Next, we examined the transitions between the layouts of each website, more specifically, how layouts changed over time and what were the main modifications. This helped us in recognising, for instance, the subsections which remained linked to the homepage after a layout change, those which were removed and those subsections which were introduced.
The analyses presented in the following chapters have been conducted manually, by browsing through the Internet Archive collections using the Wayback Machine [7].

[7] As the process of identifying transitions in the overall structure and layouts of a website has been found a useful starting point to understand the role of the website during time (see (Nanni, 2015, 2017)), in order to support future studies on similar topics we developed a simple script able to detect, given a website's homepage, the major layout changes of this page during time. The code is available here: `https://github.com/fedenanni/Structural-Changes-in-Archived-Pages`; an improvement of this solution could be very useful for web historians, as these modifications can be hard to track.

5.2.2 Results

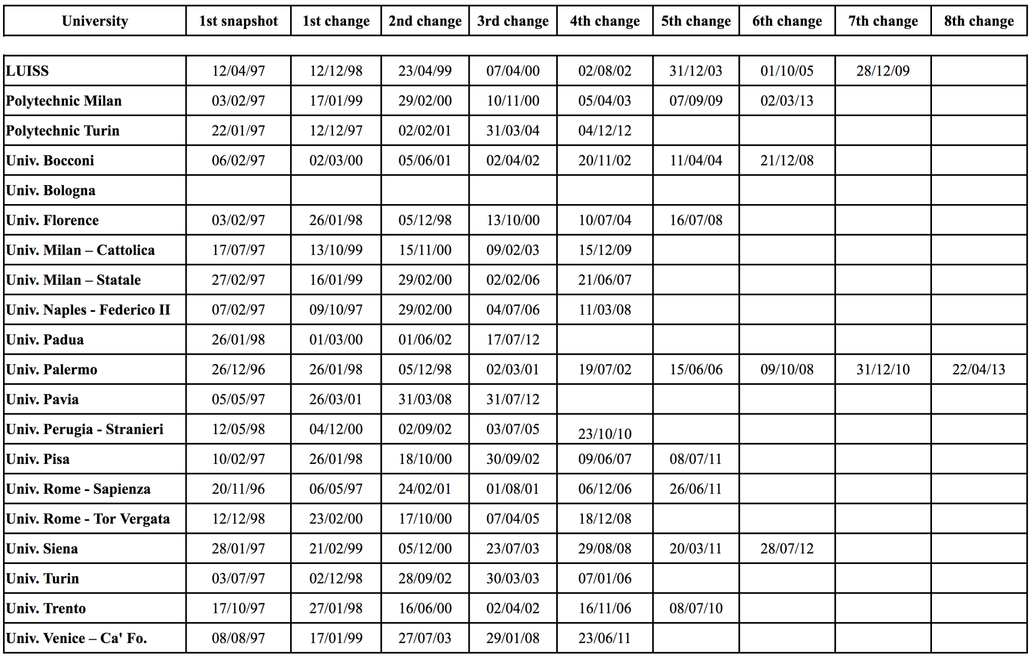
In order to establish the reliability of Internet Archive's snapshots for this research, for each university under study the first snapshot available on the Internet Archive and every major layout change in its homepage were identified (see Table 5.1)[8]. As it can be immediately noticed, universities present very different behaviours related to the changing of their websites; it is also evident that each university has re-designed their website at least four times in the last 20 years[9]. As it will be extensively presented in the next chapter, not all these changes were "revolutionary" for the the website itself, and rarely they imply the systematic removal of materials from the live web. However, any kind of in-depth close reading analysis that intends to study the reasons behind the changes or adopt the retrieved snapshots as primary sources, needs to establish the reliability of the collected digital materials (meaning, examining how well these snapshots have been preserved). Regarding this, in the following paragraphs our attention will be focused on studying four specific issues emerged during our study. They are representative of how challenging the retrieval, analysis and interpretation process of archive snapshots can be and how essential it is to critically examine these materials when the intention is to employ them as primary sources.

5.3 Issues with Archived Snapshots

In the next pages, we present four different issues with snapshots of Italian universities preserved by the Internet Archive. They concern the websites of the Polytechnic University of Turin, the University of Rome 2 – Tor Vergata, the University of Trento and the University of Bologna.

[8] It is important to remark that the dates are related to the first snapshot that presents a different layout and they can be different with the real date when the change happened.

[9] The issue regarding the University of Bologna website will be discussed in the next pages.



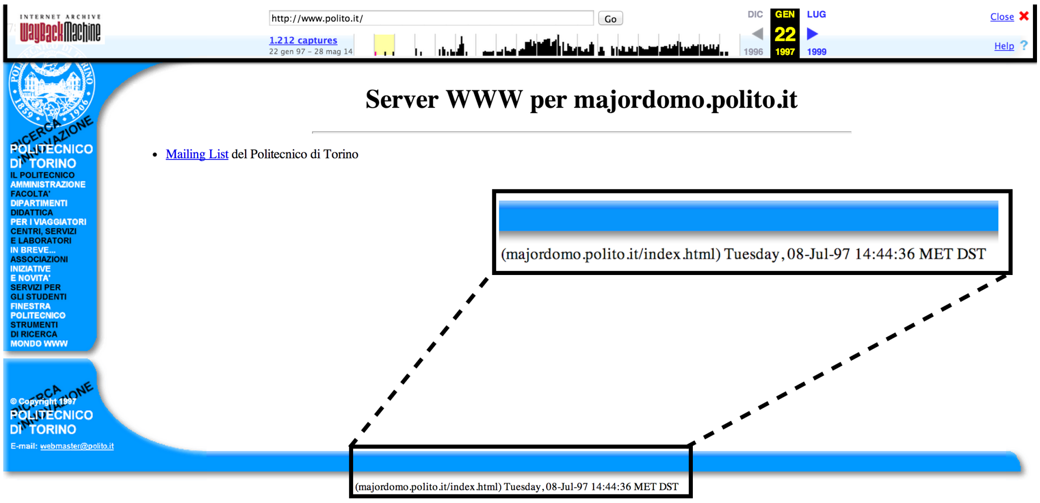
University	1st snapshot	1st change	2nd change	3rd change	4th change	5th change	6th change	7th change	8th change
LUISS	12/04/97	12/12/98	23/04/99	07/04/00	02/08/02	31/12/03	01/10/05	28/12/09	
Polytechnic Milan	03/02/97	17/01/99	29/02/00	10/11/00	05/04/03	07/09/09	02/03/13		
Polytechnic Turin	22/01/97	12/12/97	02/02/01	31/03/04	04/12/12				
Univ. Bocconi	06/02/97	02/03/00	05/06/01	02/04/02	20/11/02	11/04/04	21/12/08		
Univ. Bologna									
Univ. Florence	03/02/97	26/01/98	05/12/98	13/10/00	10/07/04	16/07/08			
Univ. Milan – Cattolica	17/07/97	13/10/99	15/11/00	09/02/03	15/12/09				
Univ. Milan – Statale	27/02/97	16/01/99	29/02/00	02/02/06	21/06/07				
Univ. Naples - Federico II	07/02/97	09/10/97	29/02/00	04/07/06	11/03/08				
Univ. Padua	26/01/98	01/03/00	01/06/02	17/07/12					
Univ. Palermo	26/12/96	26/01/98	05/12/98	02/03/01	19/07/02	15/06/06	09/10/08	31/12/10	22/04/13
Univ. Pavia	05/05/97	26/03/01	31/03/08	31/07/12					
Univ. Perugia - Stranieri	12/05/98	04/12/00	02/09/02	03/07/05	23/10/10				
Univ. Pisa	10/02/97	26/01/98	18/10/00	30/09/02	09/06/07	08/07/11			
Univ. Rome - Sapienza	20/11/96	06/05/97	24/02/01	01/08/01	06/12/06	26/06/11			
Univ. Rome - Tor Vergata	12/12/98	23/02/00	17/10/00	07/04/05	18/12/08				
Univ. Siena	28/01/97	21/02/99	05/12/00	23/07/03	29/08/08	20/03/11	28/07/12		
Univ. Turin	03/07/97	02/12/98	28/09/02	30/03/03	07/01/06				
Univ. Trento	17/10/97	27/01/98	16/06/00	02/04/02	16/11/06	08/07/10			
Univ. Venice – Ca' Fo.	08/08/97	17/01/99	27/07/03	29/01/08	23/06/11				

Figure 5.1: Major layout changes for each university website.

5.3.1 Conflicting Archival Dates

The first issue emerged when working in reconstructing the digital past of the Polytechnic University of Turin using snapshots from the Internet Archive. If we type its URL (`http://www.polito.it/`) in the search tool of the Wayback Machine, we will identify the first snapshot available of its homepage and the date on which it was taken. The Internet Archive states that it was archived on the 22nd of January 1997, but at the bottom of the page the "last modification date" indicates the 8th of July 1997 (see Fig. 5.2). Therefore this snapshot, or at least part of it, was not harvested in January but after the beginning of July.

Researchers have already pointed out how the reliability of the archived date presented by the Internet Archive is from time to time questionable (see for example Brügger (2012b); Ainsworth et al. (2015); Jackson (2015)). As a matter of fact, web archives often collect and re-compone different parts of pages in different moments while archiving a website (Ainsworth et al., 2015),



Server WWW per majordomo.polito.it

- Mailing List del Politecnico di Torino

(majordomo.polito.it/index.html) Tuesday, 08-Jul-97 14:44:36 MET DST

(majordomo.polito.it/index.html) Tuesday, 08-Jul-97 14:44:36 MET DST

Figure 5.2: The conflicting information regarding the preservation date of the Polytechnic University of Turin first snapshot.

and this could result in incoherences as the one described above. Researchers have also remarked that the re-composition of these snapshots by web archive institutions is not evidently explained to the final user and this could easily generate misinterpretations (and having impact on historical analyses). In fact, as the "last modification date" is generally not present on the archived snapshot of a specific page, it remains extremely difficult to establish the reliability of a retrieved source.

While conducting an extensive analysis over European and North American academic institutions, the same issue appeared in other cases (Duke University: 19/02/97 – 18/06/97; University of Edinburgh: 04/01/97 – 07/05/97). Nielsen in her work (Nielsen, 2014) underlined a similar problem while using Internet Archive snapshots from 1997. Moreover Brügger (2008, 2011b), while presenting several issues that emerge in dealing with archived websites, remarked the necessity of "methodological principles, rules and recommendations for a future critical textual philology of the website" (Brügger, 2011b).

5.3.2 URL Changes Over Time

The second issue emerged when we retrieved snapshots of the University of Rome 2 – Tor Vergata through the Internet Archive. The first snapshot available for the URL `http://www.uniroma2.it` is from December 1998, which is very unusual because almost all other university websites we analysed in our study (see Table 5.1) had been preserved at least since the end of 1997. Given that it seemed very unlikely that *a)* Internet Archive crawlers did not find Tor Vergata's website in almost two years (the Internet Archive started preserving the web in November 1996), and *b)* Tor Vergata created its website only in 1998[10]), we decided to investigate further.

Tor Vergata, as all other Italian universities analysed in this study, did not offer specific information about the "history of the website",[11] especially regarding who initially led the project and what changes had been made to the platform, we could not establish whether the website was already online in 1997 by looking at these snapshots alone.
Luckily, a link offered on Tor Vergata homepage: "Per i visitatori: Università Italiane" (Information for visitors: other Italian universities) turned out to be very helpful. It directed us to another website, realised by CILEA (Italian Universities Consortium)[12], which offered links to all Italian university websites online in those years. As the Internet Archive preserved a previous snapshot of this resource, obtained on the 25th October 1997, the information needed was identified. The University of Rome 2 - Tor Vergata appears in the list but the URL (`http://www.utovrm.it/`) is different from the current one (`http://www.uniroma2.it/`). This change could be due to a decision to standardise the addresses of Italian universities to a common form: "uni

[10] In our research Nanni (2017) we highlighted how Italian university websites where already active online at least in 1995 - which coincide with the findings of studies conducted on other national web spheres (see for UK Hale et al. (2014) and for Germany Holzmann et al. (2016b)).

[11] As opposed to American academic institutions, for instance: `https://web.archive.org/web/19970518021303/`; `http://www.utexas.edu/teamweb/history/`

[12] The importance of this specific resource for future historical studies on the Italian web sphere will be shown and remarked in the following chapter.

+ the initials of the city".

Even though the complete change of URL appeared only once in this entire study,[13] this could be identified as one of the most difficult issues that web historians will have to face[14], especially when the people who were working on the website at that time are difficult to identify[15]. Without the fundamental help of an external reliable source, such as the CILEA website, it could have been a really complicated issue to solve.

5.3.3 Robots Exclusions

The University of Trento, located in the cities of Trento and Rovereto, is the main node of a fruitful cultural ecosystem, which involves several other scientific institutions, such as the Bruno Kessler Foundation (originally founded in 1962 under the name of Istituto Trentino di Cultura), the Microsoft Bioinformatics Research Center COSBI (opened in 2005) and the Center of Integrative Biology CIBIO (2007).
For these reasons, it could be argued that, during the last twenty years, the university website has played an important advertisement role in order to attract international excellences from all around the world, to study and work in Trento. This aspect will guide us, while examining the reliability of its snapshots from the Internet Archive.

If we type the URL `http://www.unitn.it/` in the Internet Archive, we will receive a substantial number of snapshots after the middle of 2004 and a decent number of them going back until 2000. The situation is not as well documented between December 1999 and the 17th October 1997[16], which is

[13] While small modifications appeared during this work - which are managed by a redirection link offered automatically by the Internet Archive, e.g. `http://www.unipv.eu` was previously `http://www.unipv.it`

[14] To go deeper into the "changing of domain names problem" in historical researches we suggest Brügger (2011a).

[15] This issue will be described in detail in the next chapter.

[16] As the Italian version of the website has been preserved more times than the English

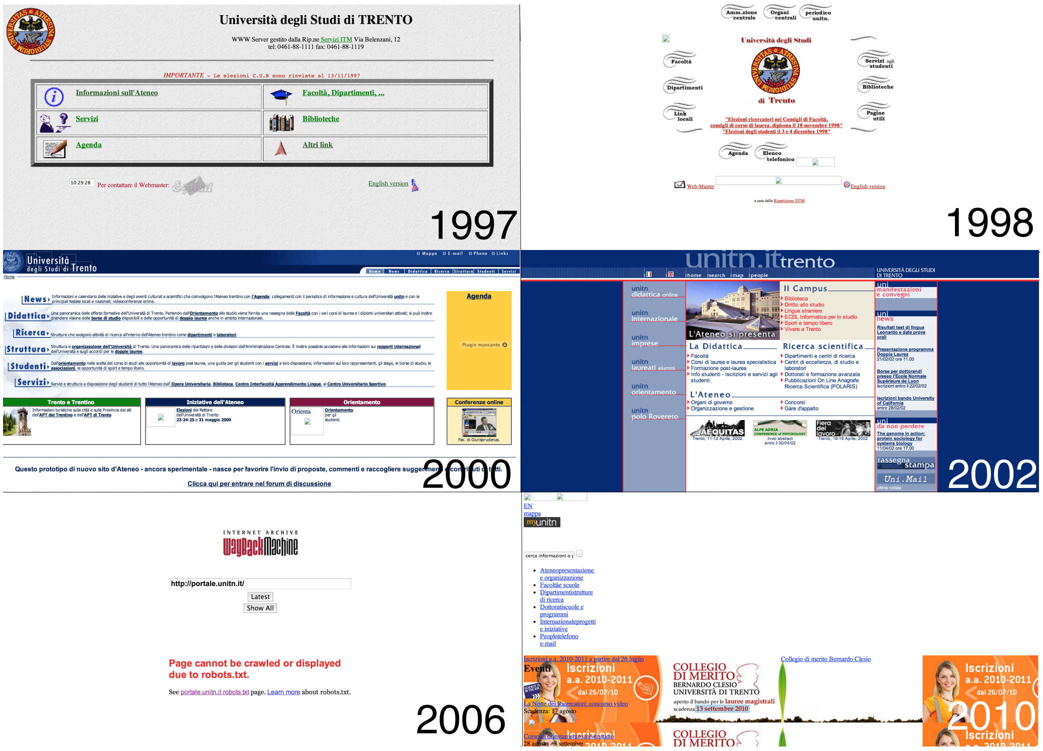


Figure 5.3: Different layout of the University of Trento website over time. Preservation issues are evident after 2006.

also the first available snapshot. During these first years the University of Trento has offered to users an almost identical English version of its website (even though the section “General Information” is not available in English). As our intent is to analyse how the University of Trento has emphasised its connections with other geographically closed research centres, in order to attract international students and researchers, it is important to notice that the link to the Istituto Trentino di Cultura was offered in the section “Other Links” already in the first archived snapshot of the website.
The first significant change in the layout of the homepage is on the 27th January 1998. On this date we can notice that the link to the Istituto Trentino di Cultura has been moved to the section “Pagine utili” (Useful pages). However, if we visit the English version, we will also notice that this specific link is not present in the “Useful pages” section.
On 16th May 2000 the website changed again. Now the link to the Istituto Trentino di Cultura is offered in a little box on the right of the section “Ricerca” (Research), however without a specific description. Astonishingly, from this date on, there is no link to the English version of the website.

Between the 2nd of July 2001 and the 2nd of April 2002 the website changed again. Studying both the Italian and the English versions of the website we noticed that the Istituto Trentino di Cultura is not mentioned anymore.
The layout of the website remained more or less the same until the 16th November 2006. From that date on, each snapshot taken from the homepage was automatically redirected to the page `http://portale.unitn.it/`, which has not been preserved by the Internet Archive (See Fig. 5.3). The reason of this exclusion from the Internet Archive is due to the fact that it follows the “robots.txt protocol”, which is a convention of advising web crawlers and other web robots to access only parts of a website which is otherwise publicly viewable. Other national web archives (such as Netarkivet and the British Web Archive) do not follow this protocol, preserving websites anyway.
Regarding the website of the University of Trento, the robots.txt page of the

version, the dates mentioned in this chapter will refer to it.

“Portale” [17] said:

```
User-agent: *
# Directories
Disallow: /
Disallow: /*
Allow: /calendariobv/*
```

”User-agent: *” means that this rule affects all crawlers. “Disallow: /” tells the crawler that it should not visit any page on the site.
The consequence of this is a total blackout in the preservation of the University of Trento’s website for almost four years; the website became available again on the Internet Archive on the 8th of July 2010. However, since that date, it has been often preserved poorly (as can be noticed in Fig. 5.3)), making the interpretation of the retrieved snapshots extremely difficult.

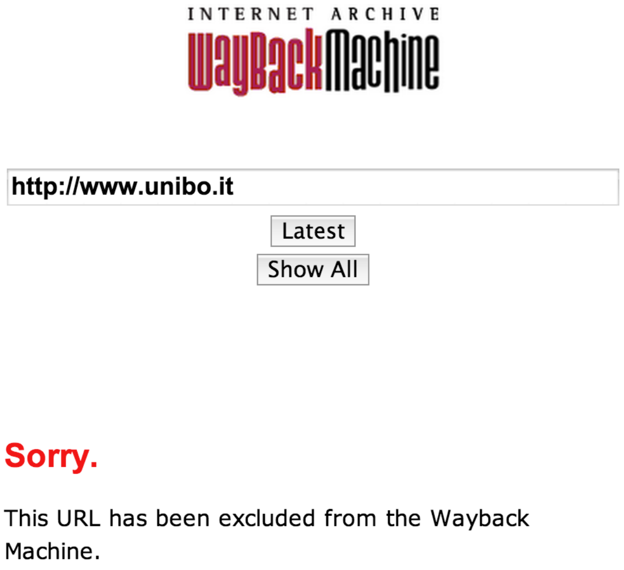
Similar preservation issues appeared in this study a few times while conducting a large-scale step of our analysis, for instance with the University of Manchester website after the 26th March 2004 and with the University of Pisa after the 8th July 2011.

5.3.4 A Different Kind of Exclusion Message

The University of Bologna is one of the largest universities in Italy, with more than 85.000 students.[18] As previously remarked, Italy does not have a National Web Archive, therefore the Internet Archive is the only platform for retrieving previous versions of its website.
However, when this research was initially conducted (October 2013) the University of Bologna’s website was excluded from the Wayback Machine

[17] This page is still available online at `http://portale.unitn.it/robots.txt`. You can see the differences with the page of the current homepage at: `http://www.unitn.it/robots.txt`.

[18] `http://www.unibo.it/it/ateneo/chi-siamo/luniversita-oggi-tra-numeri-e-innovazione/`



Sorry.

This URL has been excluded from the Wayback Machine.

Figure 5.4: The exclusion message regarding Unibo.it.

(see Fig. 5.4), presenting an issue that was never encountered before by researchers in the entire web archive community.

In particular, as the message displayed was different from the message of the snapshot of the University of Trento website in Fig. 5.3, we consequently argued that could be related to a different exclusion reason. In order to understand the reasons of this removal, the first step was to find, in the exclusion-policy of the Internet Archive, information related to the message "This URL has been excluded from the Wayback Machine", which appeared when searching: `http://www.unibo.it`.

In the FAQ section, the Internet Archive briefly describes the cases in which a website could be excluded by its platform.[19] The most common reason, as described before, is when a website explicitly requests to not be crawled

[19] `https://archive.org/about/faqs.php\#2`

by adding “User-agent: ia_archiver Disallow: /” to its robots.txt file. As specified in the FAQs: “Alexa Internet, the company that crawls the web for the Internet Archive, does respect robots.txt instructions, and even does so retroactively. If a web site owner decides he / she prefers not to have a web crawler visiting his her files and sets up robots.txt on the site, the Alexa crawlers will stop visiting those files and will make all files previously gathered from that site” unavailable.

However, it is also explained that “Sometimes a website owner will contact us directly and ask us to stop crawling or archiving a site, and we endeavour to comply with these requests. When you come across a “blocked site error” message, that means that a site owner has made such a request and it has been honoured. Currently there is no way to exclude only a portion of a site, or to exclude archiving a site for a particular time period only. When a URL has been excluded at direct owner request from being archived, that exclusion is retroactive and permanent”.

Following the information offered by the Internet Archive in their FAQ section and assuming that “This URL has been excluded” message is what they define as a “blocked site error” message, the only possible conclusion is that someone explicitly requested to remove the University of Bologna website from the Archive.

This preservation issue, which directly impedes us from collecting previous versions of the University of Bologna website and iconically highlights the risks of depending only on the efforts of an international organisation such as the Internet Archive for the preservation of a national web sphere, also emphasises the need of a new and highly interdisciplinary approach for reconstructing the past of this academic website. In the next two chapters we will present an overview of the methodology we defined for reconstructing the University of Bologna website, and next the practical results of our study.

Chapter 6

Source and Source Criticism

No one dare disturb
the sound of silence.

This chapter presents an overview of the methodology I adopted in Nanni (2015, 2017) and that I extensively described in "The Changing Digital Faces of Science Museums: A diachronic analysis of museum websites", a book chapter that will appear in the volume Web 25, edited by Niels Brügger. I wrote this work together with Anwesha Chakraborty, a colleague of mine at the International Centre for the History of Universities and Science, University of Bologna. Specifically, we applied the methodology I have developed during my Ph.D. to her research on scientific heritage institutions.
The goal of this chapter is to discuss the types of sources employed for studying the digital past of the University of Bologna's website.

6.1 Introduction

In the previous chapter, we overviewed different preservation issues that emerged analysing archived snapshots of Italian university websites, and we remarked in particular on the gravity of the complete exclusion of the University of Bologna's website (from now on Unibo.it) from the Internet Archive. In this chapter, we present a methodology for reconstructing the past of the

Unibo.it. The approaches presented in this chapter and applied in the following chapter will be mainly related to two levels of analysis: the website[1] in its entirety, and specific webpages which are highly relevant (e.g. the homepage; department, research group, professors pages).

6.2 Sources at Disposal

By consulting CeSIA (Area Sistemi Informativi e Applicazioni),[2] the team which is currently managing the website, we discovered that Unibo.it has not been consistently archived by the teams who managed it during the last two decades. For this reason, different types of primary sources needs to be adopted, in order to reconstruct the website and obtain an overview of its role for the university and the student's community. In the next sections, we present the sources we used, discussing how we assessed their reliability. As a first step, information related to the website was collected from traditional sources, such as university yearbooks and through the analysis of the university's archived records. This provided an initial understanding of the administrative role of the website (through a top-down view) and indicated the people involved in its supervision. As a second step, interviews were conducted with those who have been managing the website during the last decades. This helped, especially, in discovering the motivations behind specific changes, in collecting private backups of the platform and in tracing down who created the website and for what reasons. The analysis was then consolidated by employing information retrieved from local and national newspaper archives, student forums and Usenet discussion groups. These materials facilitated a better understanding of the role that the website has played over the years as a "bridge" between the institution and its community.

The last step of the study aimed at restoring access to the previous versions

[1] In our research we adopt the term website as a synonym of web domain, and we consider part of a website any page that belongs to the domain of study.

[2] `http://www.unibo.it/it/ateneo/organizzazione/amministrazione-generale/1931/index.html`

of the website. In order to do so, information currently available on the live web was collected and its availability in National web archives was explored. The combination of these sources offers a comprehensive perspective on the changes of the website and the political (e.g. school reforms) and educational reasons behind specific choices and decisions.

6.2.1 Library and archive materials

In order to study Unibo.it, different materials from the university library and archives have been used as primary resources. One source that has been very useful in different steps of our work is the university yearbook. The yearbook offers a general overview of the main activities of the university during the year, highlighting its management and indicating innovative decisions as well as presenting several statistics. Professor Fausto Desalvo has been in charge of the publication of the yearbook since the early 90s. The yearbooks are accessible online (first edition available: 1994/95) and at the library of the Department of History.

Even if, especially during the 90s, only a few pieces of information regarding the website would be mentioned in the yearbook, this source has nevertheless been an essential starting point for obtaining a diachronic overview of the official teams that were managing Unibo.it. When the different teams were contacted, the goal was to conduct interviews and to collect materials related to the website, such as archived documents as well as backups.

The website has been managed by four different teams in the last twenty years. However, especially during the 90s, large parts of the website were directly modified by single departments and research groups. Very little analogue archival information has been preserved by the teams and researchers that have worked on the website and its sub-sections. Even more importantly, not even a single backup of the old versions of the website has been preserved. However, this initial research helped in identifying the key people to interview.

6.2.2 Interviews

Given the ephemerality of born-digital materials and the general lack of their preservation by the teams that worked on the website, oral memories have played a key role in our research. These direct sources have been helpful for capturing the rationale behind the changing architecture of the website[3]. For this work the different teams who managed the main website were interviewed, together with technicians and researchers who worked on the development of the pages of various departments in the past two decades.

So much has been already written about the reliability and the criticism of oral memories (see for example Hoffman and Hoffman, 1994). In this research – especially given the fact that primary materials (e.g. backups of the website) were not at our prompt disposal – assessing the validity of the collected pieces of information has not always been an easy task. Therefore we proceeded by comparing the outputs of different interviews and, when possible, validated them by using other sources, such as newspaper articles. It is important to remark here that public and private backups of emails have often been used by the interviewees in order to recollect memories of their experience in working on Unibo.it and to confirm passages of the historical reconstruction. While email backups are "waiting to become" a new primary source for historians[4], the social and ethical implications of collecting, consulting and sharing their content to sustain an argument still must be fully discussed.

6.2.3 Newspapers and forums

Another way of finding information related to previous versions of Unibo.it and its role for the University of Bologna has been to search in newspaper archives and retrieve articles that mention or describe it.

[3] At the 2015 International Internet Preservation Consortium General Meeting (IIPC2015), the importance of oral memories for web historical research has been emphasised both by Ahmed AlSum and by myself in two consequential presentations: `https://www.youtube.com/watch?v=AHrxvRWf9OM`

[4] Dan Cohen (2006) discussed it when considering the large abundance of sources that public administration will leave us in the next decades.

The practice of using printed media to retrieve information about the web of the past has been already described, for instance in Brügger (2011b). In this research the digital archive of the newspaper *La Stampa*[5] was used in order to retrieve specific articles published between 1996 and 1999 that described the general use of the web by Italian universities[6]. A great role in this study has been played by local and national newspapers (such as the digital archives of *La Repubblica* and *Il Resto del Carlino*), which especially during the 90s offered an overview on the new functionalities on the website (e.g. free email account for all students, online fee payments, etc.), together with university digital magazines (Alma2000, AlmaNews, Unibo Magazine). However, it is important to employ news articles critically and always consider how and why a specific piece of information regarding the website was selected and published in the daily edition of a general newspaper[7].

Other sources that have been employed in this study include student forums (e.g. UniversiBo) and, to go further back in time, Usenet discussions preserved by Google. While academic forums offer new materials for historians of universities interested in better understanding student life, they also present the perspective of a very small and specific subset of the academic community. In particular, in the early 90s, these online forums were mainly kept running by students (together with researchers and professors) in STEM fields, whose departments were often the first to offer access to the web.

6.2.4 Live web materials

The current version of a university website offers to the user a variety of primary sources. Live web materials reveal the current role of website in the university's organisation and management (e.g. attracting national and

[5] http://www.archiviolastampa.it/

[6] E.g. the article "Anche l'università via Internet", written by Giovanna Favro and published the 14th of May 1998.

[7] For example this short article on the possibility of creating university email accounts in 2002: http://ricerca.repubblica.it/repubblica/archivio/repubblica/2002/10/09/mail-gratuita-per-gli-studenti.html?ref=search

international students and researches, promoting collaborations with the private sector, etc). Additionally, by combining documents from the website and from social media pages of the institution (such as Facebook, Youtube and Twitter profiles), we can make reasonable assumptions about the digital interactions with the larger community. While materials from social networking websites will play a fundamental role in better understanding the multidirectional communication between academic institutions and their community, it is important to remember that their suitability for historical analysis is currently under scrutiny, as several issues have been raised (Webster, 2015; Zimmer, 2015).

6.2.5 Italian Websites in Other National Web Archives

Since the second part of the Nineties, initiatives have been taken by private (Lyman and Kahle, 1998) and public (Gomes et al., 2011) actors to preserve the web for future research. Aside from the Internet Archive, since 1996 several different national libraries have also begun preserving their national web past. PANDORA, started in 1996 by the National Library of Australia, the UK Web Archive (2004), the Netarkinvet (2005) in Denmark and the Portuguese Web Archive (2011) are just a few example. Given our focus on Italian academic websites and the fact that Italy does not have a National web archive, in this research we examined the potential of retrieving primary sources both from the Internet Archive and from other National web archives.

The practice of retrieving primary sources related to an Italian university website in foreign web archives could sound strange as the goal of a national web archive is precisely to preserve the web of its country. However, as this preservation process is highly complex (as described by Brügger (2009)), from time to time part of the non-national web will also end up to be unintentionally preserved. For example, to archive national web spheres in an automatic way, archivists could set up crawlers with a maximum number of hyperlinks they can follow, given a specific set of starting points. A crawler which is set to go at most 10 links away from one of these URLs, could also end up crawling non-national content, as it will systematically follow all the hyperlinks. For this reason, if the University of Bologna were to organise a Summer School and the University of Amsterdam had linked it from its website, the University of Bologna website (or at least part of it) would be unintentionally preserved in the Netherlands Web Archive.

6.3 A Critical Combination of Sources

The critical combination of the sources presented above provided the possibility of reconstructing the changes in Unibo.it, emphasising the different roles that the institution assigned to its website during the years and the way the student community interact with the website to establish a dialogue with the university. The results of our study is presented in the next chapter.

Chapter 7

Reconstructing a Website's Lost Past

Spring is here again.

This chapter is based on the findings presented in the paper "Reconstructing a website's lost past - Methodological issues concerning the history of www.unibo.it", which will appear in the journal Digital Humanities Quarterly *and was awarded as second best paper at the 2015 Göttingen Dialog in Digital Humanities. The goal of this chapter is to describe how to deal with the scarcity of born-digital primary sources while retrieving sources on the recent past of an academic institution. The case study is an analysis of the first 25 years online of the University of Bologna. The focus of this work is primarily methodological: several different issues are presented, starting with the fact that the University of Bologna website has been excluded for thirteen years from the Internet Archive's Wayback Machine, and possible solutions are proposed and applied.*

7.1 Introduction

This chapter presents a study focused on reconstructing the past of the University of Bologna website. It starts by considering the facts that (a) Unibo.it was not accessible through the Internet Archive's Wayback Machine when this research was conducted (as described in the previous chapters) and (b) Italy is one of the few countries in Western Europe that does not systematically preserve its National web sphere. By doing so, the goal of this chapter is to address the following research questions:

- Is it possible to reconstruct and study the past of a university website (namely the changes in its layout, structure and content, but more importantly the reasons that have caused them) without having at prompt disposal a collection of web archive snapshots?
- Could this study guide us to better understand the role the website has played in the interactions between the academic institution and its large and variegated community and, by that, could we obtain new insights in the recent past of the institution in itself?
- Will this research bring new materials to the surface, in ways that are useful for the research communities that, so far, have focused on the past and present of academic institutions?

7.1.1 Setting Up the Research

In order to reconstruct and study the digital past of the University of Bologna website, we have adopted a few different types of primary sources (which reliability has been already discussed in the previous chapter). As a first step, we collected information related to the website from the university yearbooks and through the analysis of university archived records. Additionally we conducted interviews with the people who have managed Unibo.it during the last two decades. This helped us especially in discovering who initially created the website and for which reasons.

As snapshots of the Internet Archive are not at our prompt disposal and a Italian web archive still does not exist, our second step was to examine the information currently available on the live web and, by projecting them back in time, we explored their availability in foreign web archives.

By combining these sources we obtained a comprehensive perspective on the evolution of the platform and we highlighted the political and educational reasons behind specific changes and decisions. We consolidated our study employing information retrieved from local and national newspaper archives, such as *La Repubblica* and *Il Resto del Carlino*, student forums and Usenet discussion group.

This research highlights that it is possible to reconstruct and study the past of a website, and consequently the history of the institution behind it, even without snapshots from the Internet Archive. Given the peculiarity of the exclusion issue happened to Unibo.it, in the last part of this chapter we presented how we intensively worked together with the Internet Archive, to discover the reasons behind it and to find a way to solve it.

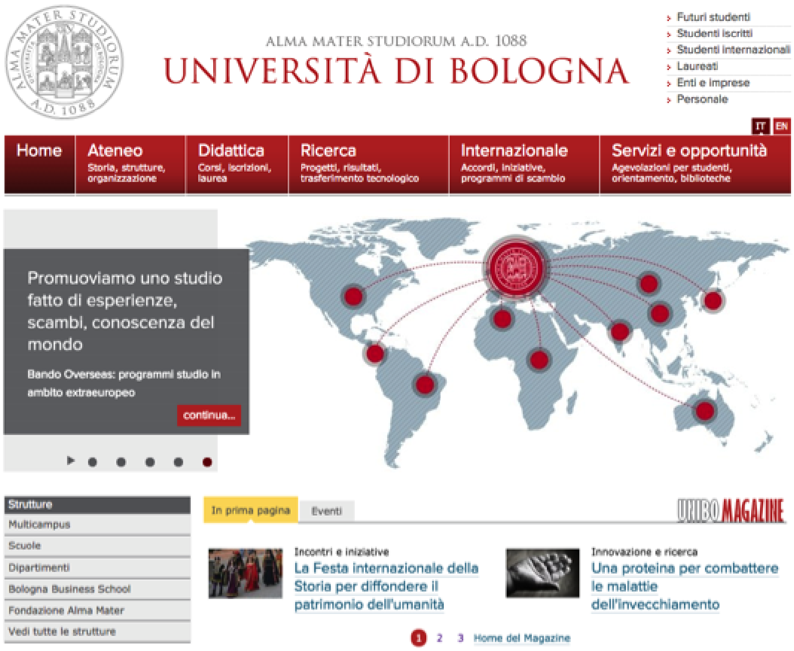
7.2 Going Back in Time: Unibo.it Between 2015 and 2002

As it will be explained in this section, the majority of the information published on the website between the early 2000s and 2015 are still available online (for example all the courses programs, the descriptions of research projects, the contracts and grants published by each School).[1] However, when this research was conducted, the Internet Archive was not offering any snapshot of the homepage of Unibo.it. For this reason, it was not possible to monitor how the layout changed during time or analyse the impact of Italian school reforms on the structure of the platform. In order to face this issue, we started by studying the materials available on the website today, and then we tried to reconstruct their past by employing alternative primary sources.

7.2.1 The Website as It is Structured Today

The website of the University of Bologna is currently offered in two different versions: Italian (which is available at the URL: `http://www.unibo.it/it`),

[1] Since the academic year 2012/13 the 23 faculties have been reorganised in 11 schools.



Promuoviamo uno studio fatto di esperienze, scambi, conoscenza del mondo

Bando Overseas: programmi studio in ambito extraeuropeo

Incontri e iniziative

La Festa internazionale della Storia per diffondere il patrimonio dell'umanità

Innovazione e ricerca

Una proteina per combattere le malattie dell'invecchiamento

Figure 7.1: Unibo.it homepage in 2016.

and English. Moreover, as the university is divided in five campuses, the website is consequently divided into five subsections (for example, `http://www.unibo.it/it/campus-forli`). As the English version offers the translation of only a part of the website, the focus of this research will be mainly on the Italian version.

This website (see Figure 7.1) is currently managed by two different offices: "CeSIA - Settore Tecnologie web" that takes care of the structure (called "Sistema Portale di Ateneo"), and "AAGG — Ufficio Portale Internet e Intranet di Ateneo"[2] that manages the content. If we consider its subsections, such as "Didattica" (educational information) and "Ricerca" (research"), its sub-domains, such as the "Unibo Magazine", and retrieve current and old abandoned department web pages[3], we can obtain an initial overview of the current status and structure of the website.

[2] `http://www.unibo.it/it/ateneo/organizzazione/amministrazione-generale/81380/817/index.html`

[3] `http://www2.stat.unibo.it/`; `http://www2.classics.unibo.it/`

This allows us to notice that large pieces of information published online by the university between the early 2000s and 2015 are still available online (for example all the course programs, the descriptions of research projects and the contracts and grants published by each School). However, in order to retrieve them, a very basic "string-matching" search tool is the only tool promptly available[4].

7.2.2 The Moment of Transformation

To understand why most of the materials from the early 2000s are still available online, while resources from the 90s seem way more difficult to retrieve, we contacted the people who have been involved with the management of the website during the last fifteen years.

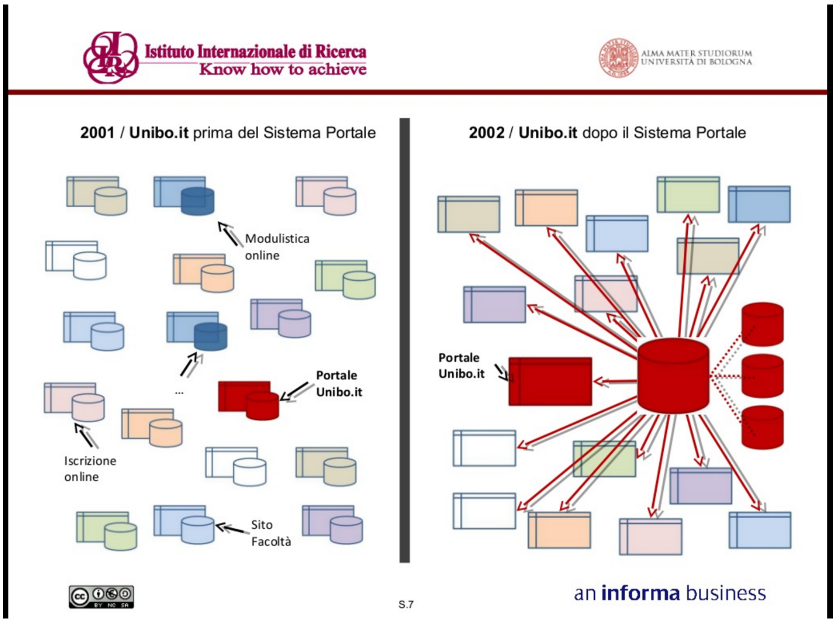
Luca Garlaschelli was the Chief of the Information/Innovation Office (CIO) at the University of Bologna between 2002 and 2012. Under his supervision the "Sistema Portale di Ateneo" was created. This is a general interface to a hierarchical organization of all the digital resources of the university that are available online (see Fig. 7.2), with a specific focus on enhancing the accessibility of the information at disposal[5]. As it will be presented in the next pages, this has led to a revolutionary transformation of the digital presence of the institution, which made Unibo.it a reference point for all other Italian university websites. As a matter of fact, for three consecutive years Unibo.it received the "Osc@r del web" prize as the best Italian public administration website[6].

Among several improvements of the website, this transformation required

[4] http://search.unibo.it/UniboWeb/UniboSearch/Default.aspx

[5] As described by Garlaschelli here: http://www.slideshare.net/lucagarlaschelli/private-cloud-computing-in-organizzazioni-complesse and also presented in the "Annuario degli anni accademici 2003-2004 e 2004-2005", pp. 747-750: http://www2.unibo.it/Annuari/Annu030405/Annuario0304-0405.pdf

[6] http://www.magazine.unibo.it/archivio/2007/oscar_del_2007. In 2007 Luigi Nicolais, the Italian Minister of Public Administration, was also present to confer the prize.



Istituto Internazionale di Ricerca
Know how to achieve

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

2001 / Unibo.it prima del Sistema Portale

2002 / Unibo.it dopo il Sistema Portale

Figure 7.2: A slide from Garlaschelli's presentation on the impact of the Portale d'Ateneo.

that all departments and webpages which provided information on the various degree programs change their structure and adopt a common layout and organization of their content. As an example, the Department of Classic and Medieval Philology and the Department of Computer Science had to change their URL addresses to standardized ones ("abbreviation of the name of the department" + "unibo.it"). Thus, the first one changed from: `http://www.classics.unibo.it/` to `http://www.ficlit.unibo.it/` and the second one from `http://www.cs.unibo.it/` to `http://www.informatica.unibo.it/`.

This transition started in 2004 and often required the creation of completely new department pages. A few departments decided to keep the older version of their sub-domains online by adding a "2" after the "www"[7] (as an example, the previously mentioned `http://www2.classics.unibo.it/`),

[7] This could be due to a personal choice of the person who was managing each department page at that time and not to a decision of the CIO. On the Department of Classic and Medieval Philology homepage it is explicitly written that "the pages will continue to be available, but will be no longer updated".

while the majority simply removed the old versions of their page from the live web, for example the Department of History, whose URL was: `http://www.dds.unibo.it/`.

Even if the University of Bologna homepage and all its subsections were not accessible through the Wayback Machine, during our research we discovered that its sub-domains (e.g. `http://www.dds.unibo.it/`) were available on the Internet Archive and have been constantly preserved in the last twenty years. However, discover their URLs is not a trivial issue. For example, the Department of Philosophy and Communication Studies were two different departments until 2012 and the Department of Philosophy used `http://www.filosofia.unibo.it` as a URL even before the transition to the "Sistema Portale d'Ateneo". However, in the 90s, this department used another URL for a couple of years, `http://www.sofia.philo.unibo.it` which, without the memories of the people who managed the sub-section at that time, would have been very difficult to discover.

In summary, we can identify a specific turning point in the history of the University of Bologna website. With the "Sistema Portale d'Ateneo" project and in particular with the standardization of departments pages which started in 2004[8], the website has been completely re-organized and the majority of the previous content on these pages has been deleted from the live web. However, if they have maintained the same URL or if the previous URL is known[9], the subdomain materials and their structure can be retrieved from the Internet Archive.

7.2.3 Exploring the "Sistema Portale"

Mauro Amico, head of the web-technologies department at CeSIA, offered a collection of seven .png images (see Appendix) that capture the most impor-

[8] `http://www.magazine.unibo.it/archivio/2004/dipartimenti`

[9] A web page archived in 2002 could help us identify the URL of each department in that year: `https://web.archive.org/web/20020224030346/http://alma2000.unibo.it/facolta/dipE.asp`

tant instances in the evolution of the organization of the homepage before the current layout (1996 – 2013)[10].

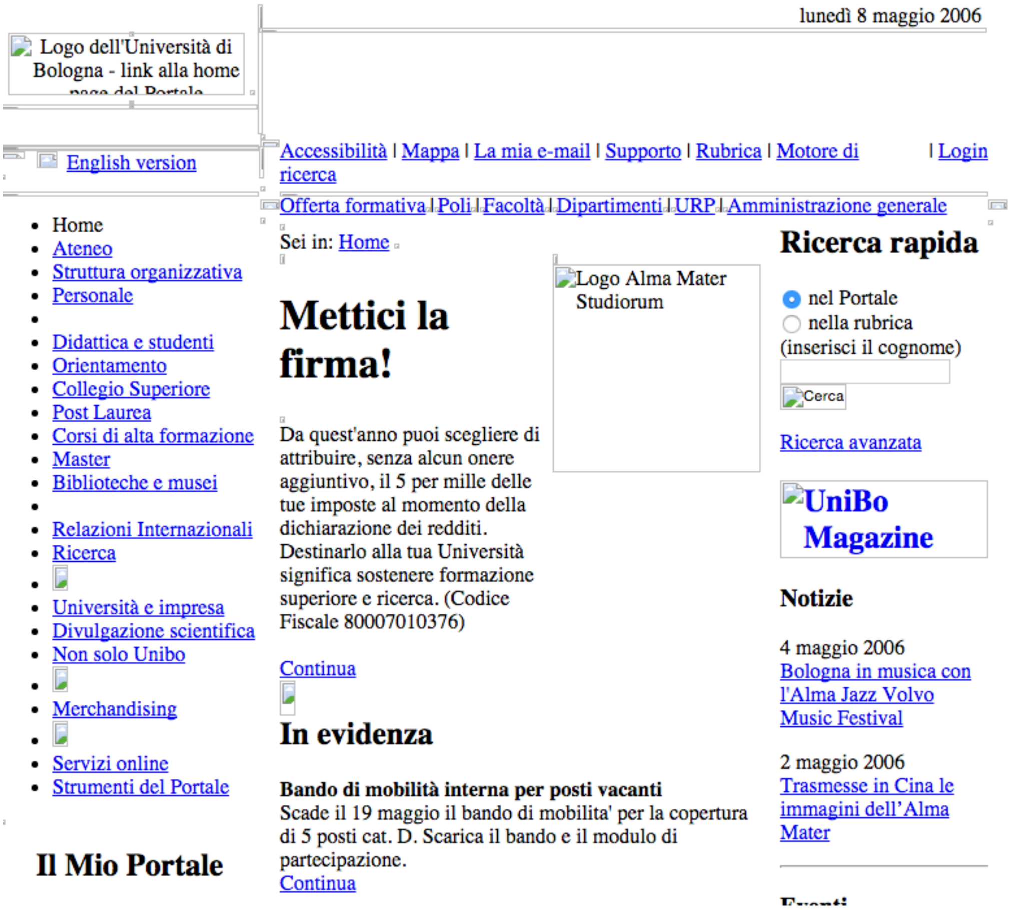
Looking at the snapshots after 2002 we can observe that, even if a few graphical adjustments were made (the Unibo-Magazine was introduced on the left in 2004; the search tool was repositioned in the center in 2006, etc.), the structure remained more or less the same until the July 2013, when the current interface was presented.

The present organization of the "Sistema Portale d'Ateneo" is the first one to be completely created by CeSIA without the supervision of Luca Garlaschelli and, along with a new graphic interface, its main characteristic is the fact that it offers for the first time the possibility of surfing the website as a specific user (a prospective student, a student, a private company, etc.) and it proposes different contents accordingly, thereby allowing better optimization and personalization of the website.

Even if these .png images give us a first idea of the different interfaces, to be able to explore again the old versions of the website, other services have to be employed. As a start, the Internet Memory Foundation offers the results of the 2006 national ".it" crawl online, but only a single snapshot of Unibo.it homepage is available (archived on the 8th of May 2006[11]) and that is not even completely preserved (see Figure 7.3). As previously noted, other national web archives could have captured, from time to time, parts of the Italian web sphere. Among them, we discovered that both the Portuguese (Arquivo) and Danish (Netarkivet) web archives have preserved parts of Unibo.it several times from 2006. These snapshots (example in Figure 7.4), collected during a visiting period at the Internet Centre of Aarhus University, allow us, for the first time, to explore and examine the differences in structure and content of the previous versions of the "Sistema Portale".

[10] They cover the periods: 01/1996-01/1998; 01/1998 – 09/1998; 09/1998-07/1999; 2002-2003; 2004-2206; 2006-2009; 2009-2013.

[11] `http://collection.europarchive.org/bncf/20060508021404/http://www.unibo.it/Portale/default.htm`



lunedì 8 maggio 2006

Logo dell'Università di Bologna - link alla home page del Portale

English version

Accessibilità | Mappa | La mia e-mail | Supporto | Rubrica | Motore di ricerca | Login

Offerta formativa | Poli | Facoltà | Dipartimenti | URP | Amministrazione generale

- Home
- Ateneo
- Struttura organizzativa
- Personale
-
- Didattica e studenti
- Orientamento
- Collegio Superiore
- Post Laurea
- Corsi di alta formazione
- Master
- Biblioteche e musei
-
- Relazioni Internazionali
- Ricerca
-
- Università e impresa
- Divulgazione scientifica
- Non solo Unibo
-
- Merchandising
-
- Servizi online
- Strumenti del Portale

Il Mio Portale

Sei in: Home

Metticí la firma!

Da quest'anno puoi scegliere di attribuire, senza alcun onere aggiuntivo, il 5 per mille delle tue imposte al momento della dichiarazione dei redditi. Destinarlo alla tua Università significa sostenere formazione superiore e ricerca. (Codice Fiscale 80007010376)

Continua

In evidenza

Bando di mobilità interna per posti vacanti
Scade il 19 maggio il bando di mobilita' per la copertura di 5 posti cat. D. Scarica il bando e il modulo di partecipazione.
Continua

Logo Alma Mater Studiorum

Ricerca rapida

● nel Portale
○ nella rubrica
(inserisci il cognome)

Cerca

Ricerca avanzata

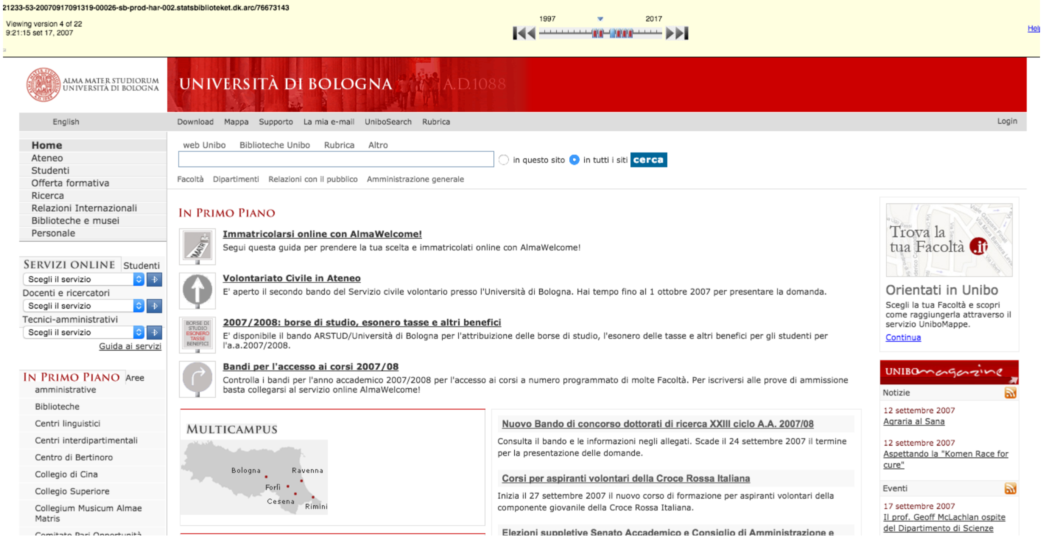
UniBo Magazine

Notizie

4 maggio 2006
Bologna in musica con l'Alma Jazz Volvo Music Festival

2 maggio 2006
Trasmesse in Cina le immagini dell'Alma Mater

Figure 7.3: The only, badly collected, snapshot of Unibo.it from the Internet Memory Foundation.



21233-53-20070917091319-00026-sb-prod-har-002.statsbiblioteket.dk.arc/76673143

Viewing version 4 of 22
9:21:15 set 17, 2007

IN PRIMO PIANO

Immatricolarsi online con AlmaWelcome!
Segui questa guida per prendere la tua scelta e immatricolati online con AlmaWelcome!

Volontariato Civile in Ateneo
E' aperto il secondo bando del Servizio civile volontario presso l'Università di Bologna. Hai tempo fino al 1 ottobre 2007 per presentare la domanda.

2007/2008: borse di studio, esonero tasse e altri benefici
E' disponibile il bando ARSTUD/Università di Bologna per l'attribuzione delle borse di studio, l'esonero delle tasse e altri benefici per gli studenti per l'a.a.2007/2008.

Bandi per l'accesso ai corsi 2007/08
Controlla i bandi per l'anno accademico 2007/2008 per l'accesso ai corsi a numero programmato di molte Facoltà. Per iscriversi alle prove di ammissione basta collegarsi al servizio online AlmaWelcome!

Nuovo Bando di concorso dottorati di ricerca XXIII ciclo A.A. 2007/08
Consulta il bando e le informazioni negli allegati. Scade il 24 settembre 2007 il termine per la presentazione delle domande.

Corsi per aspiranti volontari della Croce Rossa Italiana
Inizia il 27 settembre 2007 il nuovo corso di formazione per aspiranti volontari della componente giovanile della Croce Rossa Italiana.

Elezioni suppletive Senato Accademico e Consiglio di Amministrazione e

Orientati in Unibo
Scegli la tua Facoltà e scopri come raggiungerla attraverso il servizio UniboMappe.
Continua

12 settembre 2007
Agraria al Sana

12 settembre 2007
Aspettando la "Komen Race for cure"

17 settembre 2007
Il prof. Geoff McLachlan ospite del Dipartimento di Scienze

Figure 7.4: A snapshot of Unibo.it in 2006, from Netarkivet.

7.2.4 News From Its Recent Past

The use of primary sources from the digital archive of the newspaper *La Repubblica* has been of significant help in this study. As a matter of fact, these articles present an overview of the interactions between the university and its community, as well as insights on the key role of the website as an intermediary[12]. It has been found, for example, that in 2003 the university introduced, on its website, the digital edition of the student-guide of the city of Bologna, as also described in a news in the Unibo Magazine[13]. The guide was written by Umberto Eco, Carlo Lucarelli and other renowned professors and writers. This document provides a list of useful digital resources for new students, i.e. the platform "Flash Giovani", created with the support of the municipality of Bologna and focused on the cultural activities in the city and the website "Studenti.it" which has, in the last fifteen years, become one the most important Italian online communities for high school and university

[12] `http://ricerca.repubblica.it/repubblica/archivio/repubblica/2003/09/18/universita-insegna-la-dolce-vita-da.html?ref=search`

[13] `http://www.magazine.unibo.it/archivio/2003/09/22/guida-di-bologna`

students[14].

As described in the Unibo Magazine[15], since 2004 each professor has had a personal page, in which they publish course programs (and all additional materials, such as slides), their research interests and publications list[16].

Digital sources related to the recent years of the university also allow us to discover how in 2005 the future "Prorettore per la ricerca" Dario Braga underlined the importance of starting to teach courses in English (and also Chinese and Arabic) among his "proposals for the future"[17] or how five years later, during his administration, he actively discussed[18] in a Google Group newsletter the impact of the "Gelmini" school reform with a group of professors who were collectively termed as "Docenti preoccupati" ("worried professors")[19].

Moreover, these materials gave us insight into the activities of the "Centro Studi La Permanenza del Classico" whose director was the former Rector, Ivano Dionigi[20] and showed how the Unibo Magazine presented itself online in 2003 (with an interview[21] of the then Rector, Pier Ugo Calzolari, who spoke about the scarcity of funding for higher education and research in Italy).

[14] Currently it is one of the 200 most visited websites in Italy: http://www.alexa.com/siteinfo/studenti.it\#trafficstats

[15] http://www.magazine.unibo.it/archivio/2004/pagina-personale-docente/

[16] The main pages of professors have been also excluded from the Wayback Machine; other national web archives have preserved just a few of them.

[17] http://ricerca.repubblica.it/repubblica/archivio/repubblica/2005/04/13/proposte-per-ateneo-del-futuro.html?ref=search

[18] https://groups.google.com/forum/\#!topic/docentipreoccupati/WqOJqQzkPmU

[19] Another interesting source in order to study the experience of Dario Braga as Prorettore and its run for the future Rettore of the university, will be his personal blog: http://www.dariobraga.com/blog

[20] http://ricerca.repubblica.it/repubblica/archivio/repubblica/2006/10/25/umanisti-scienziati-insieme-alla-stessa-lezione.html?ref=search

[21] http://www.magazine.unibo.it/archivio/2003/11/11/intervista-al-rettore

Among all these different resources, one source deserves special mention. In May 2007, a group of activists decided to create a copy of the Unibo.it interface. They were demonstrating against the European Credit Transfer and Accumulation System (ECTS) for the evaluation of the number of hours of study. They believed that the university website could be the perfect target for their protest, in order to attract the attention of the institution. At the URL `http://www.unibologna.eu/` an identical version of the homepage was available, with the description of the reasons of the protest. In a couple of weeks, the website attracted a high number of visitors and most of all the attention of the university[22], which blocked the access to it from all its computers[23].

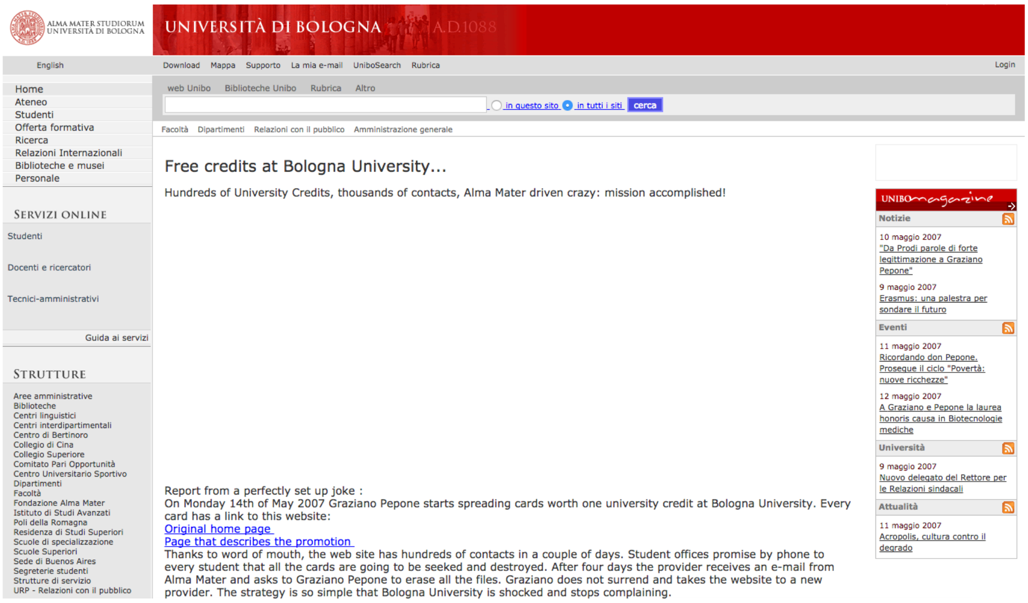
This source is not only important in our study as it documents a different and innovative way of conducting a protest against an academic institution, but as the fake-website has been preserved by the Internet Archive it also paradoxically offers a preserved version of Unibo.it, so that we can browse and study (see Fig. 7.5).

7.3 The history of www.unibo.it: 2002 - The Early 90s

Neither material on the live web nor documents in other national web archives are available for the first ten years of history of this website. For this reason, the second part of this study will mainly employ information from local and national newspapers, which have often described new services offered by the university to its community and will combine it with archive resources (in particular from the university yearbooks). As before, a pivotal role in this study has been played by the collection and critical selection of oral memories.

[22] `http://www.magazine.unibo.it/archivio/2007/attacco\_al\_portale`

[23] `http://ricerca.repubblica.it/repubblica/archivio/repubblica/2007/06/13/clonato-il-sito-dell-ateneo-per-protesta.html?ref=search`



Free credits at Bologna University...

Hundreds of University Credits, thousands of contacts, Alma Mater driven crazy: mission accomplished!

Report from a perfectly set up joke :
On Monday 14th of May 2007 Graziano Pepone starts spreading cards worth one university credit at Bologna University. Every card has a link to this website:
Original home page
Page that describes the promotion
Thanks to word of mouth, the web site has hundreds of contacts in a couple of days. Student offices promise by phone to every student that all the cards are going to be seeked and destroyed. After four days the provider receives an e-mail from Alma Mater and asks to Graziano Pepone to erase all the files. Graziano does not surrend and takes the website to a new provider. The strategy is so simple that Bologna University is shocked and stops complaining.

Notizie

10 maggio 2007
"Da Prodi parole di forte legittimazione a Graziano Pepone"

9 maggio 2007
Erasmus: una palestra per sondare il futuro

Eventi

11 maggio 2007
Ricordando don Pepone. Proseque il ciclo "Povertà: nuove ricchezze"

12 maggio 2007
A Graziano e Pepone la laurea honoris causa in Biotecnologie mediche

Università

9 maggio 2007
Nuovo delegato del Rettore per le Relazioni sindacali

Attualità

11 maggio 2007
Acropolis, cultura contro il degrado

Figure 7.5: The cloned version of Unibo.it, 2007.

7.3.1 Different Ways of Going Back in Time

In order to study the structure of the website before the "Sistema Portale d'Ateneo" several different sources have been employed, which will in turn help us in understanding what the website looked like, how it was used and how relevant it was in the academic digital "ecosystem".

In particular, as described earlier, the archive of the newspaper *La Repubblica* offers important information on how the website changed during the 90s. For example, it was discovered that the institution offered a free email account to all students from 2002[24] and it was the first Italian university which gave the possibility of paying fees online (2000);[25] moreover since 1999 some departments also guaranteed the possibility of enrolling for courses and exams online[26].

[24] http://ricerca.repubblica.it/repubblica/archivio/repubblica/2002/10/09/mail-gratuita-per-gli-studenti.html?ref=search

[25] http://ricerca.repubblica.it/repubblica/archivio/repubblica/2000/07/17/tasse-online-per-universita.html?ref=search

[26] http://ricerca.repubblica.it/repubblica/archivio/repubblica/1999/09/

Another interesting piece of news retrieved from the digital archive of *La Repubblica* is from October 2001, a few months before the project "Portale d'Ateneo" started. In those days, the University of Bologna website won the "WWW" prize from the Italian economic newspaper *Il Sole 24 Ore* for the best website in the category "School, university and research". At the ceremony Salvatore Mirabella, a technician who managed the website during the 90s, was also present[27].

However, as we can notice by looking at the images offered by CeSIA or by analyzing a few examples that are still available on the live web (i.e. `http://www2.unibo.it/annuari`)[28], before the "Sistema Portale d'Ateneo" project the homepage of Unibo.it was mainly an information page, presenting only a few links (See Fig. 7.6).

At the same time, consulting "The list and map of the Italian WWW servers"[29] created by Cilea and available from 1997 onwards on the Internet Archive[30], we can observe that several departments, faculties and research groups were already online and, as opposed to the relatively passive homepage, very active in the 90s. For example, we can retrieve all the information on courses in history since 1998[31], the organisation of the university astronomical observatory[32] and of the faculty of Engineering[33] since 1997, description on the

`13/finalmente-offerta-portata-di-mouse.html?ref=search`

[27] He was the head of "Urp – Servizio Web" , as described here: `http://www2.unibo.it/Annuari/Annu9901/Indice/parte2/parte2sez1/parte2sez1.html`

[28] It is important to notice that the page `http://www2.unibo.it` is not available on the live web anymore and it was excluded from the Wayback Machine.

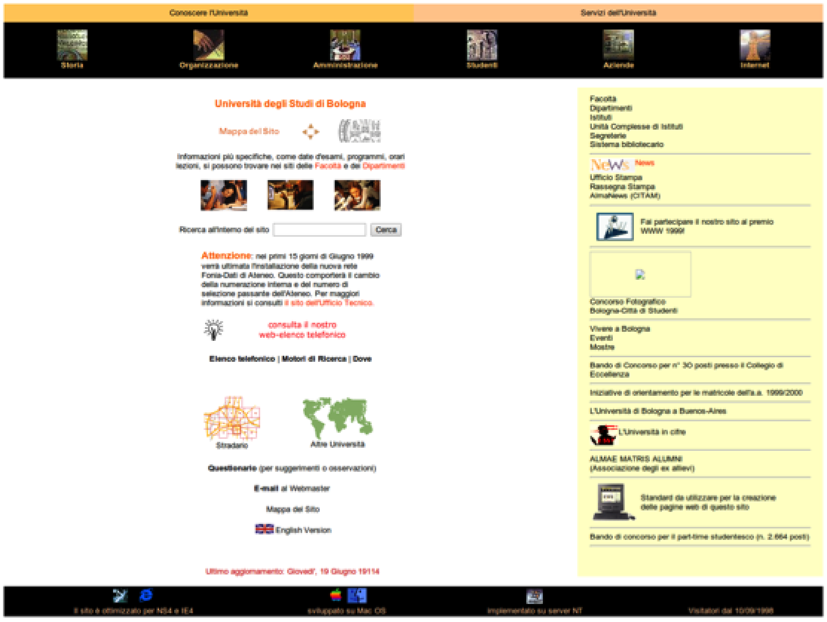
[29] As already remarked in the previous chapter, this is a useful starting point for every researcher who is interested in the past of the Italian web sphere.

[30] `http://web.archive.org/web/19971025045601/http://www.cilea.it/WWW-map/`

[31] `https://web.archive.org/web/19981206110539/http://www.dds.unibo.it/`

[32] `http://web.archive.org/web/19970114105744/http://www.bo.astro.it/`

[33] `https://web.archive.org/web/19970422153341/http://www.ing.unibo.it/`



Conoscere l'Università

Servizi dell'Università

Storia

Organizzazione

Amministrazione

Studenti

Aziende

Internet

Università degli Studi di Bologna

Mappa del Sito

Informazioni più specifiche, come date d'esami, programmi, orari lezioni, si possono trovare nei siti delle Facoltà e dei Dipartimenti

Ricerca all'interno del sito

Cerca

Attenzione: nei primi 15 giorni di Giugno 1999 verrà ultimata l'installazione della nuova rete Fonia-Dati di Ateneo. Questo comporterà il cambio della numerazione interna e del numero di selezione passante dell'Ateneo. Per maggiori informazioni si consulti il sito dell'Ufficio Tecnico.

consulta il nostro
web-elenco telefonico

Elenco telefonico | Motori di Ricerca | Dove

Stradario

Altre Università

Questionario (per suggerimenti o osservazioni)

E-mail al Webmaster

Mappa del Sito

English Version

Ultimo aggiornamento: Giovedi', 19 Giugno 19114

Facoltà
Dipartimenti
Istituti
Unità Complesse di Istituti
Segreterie
Sistema bibliotecario

NeWs News
Ufficio Stampa
Rassegna Stampa
AlmaNews (CITAM)

Fai partecipare il nostro sito al premio WWW 1999!

Concorso Fotografico
Bologna-Città di Studenti

Vivere a Bologna
Eventi
Mostre

Bando di Concorso per n° 30 posti presso il Collegio di Eccellenza

Iniziative di orientamento per le matricole dell'a.a. 1999/2000

L'Università di Bologna a Buenos-Aires

L'Università in cifre

ALMAE MATRIS ALUMNI
(Associazione degli ex allievi)

Standard da utilizzare per la creazione delle pagine web di questo sito

Bando di concorso per il part-time studentesco (n. 2.664 posti)

Il sito è ottimizzato per NS4 e IE4

sviluppato su Mac OS

implementato su server NT

Visitatori dal 10/09/1998

Figure 7.6: Unibo.it in 1999.

inter-faculty library since October 1996[34] (the entire system was created in 1993[35]), the digitisation of the students guide books carried out by the faculty of economics in 1994.[36]

These different pages were extremely useful for prospective and enrolled students[37]. For these reasons, they all evolved differently during the 90s and they are now interesting instances on how the departments of this university approached the World Wide Web.

In addition to examining departmental pages, there are many other ways of looking at the second half of the 90s' history of Unibo.it. We could follow the information related to AlmaNET, the university internal Internet connection,

[34] https://web.archive.org/web/19961031164155/http://liber.cib.unibo.it/

[35] https://web.archive.org/web/20010424195903/http://www.cilea.it/collabora/GARR-NIR/nir-it-2/atti/cib.html

[36] https://web.archive.org/web/20010424195734/http://www.cilea.it/collabora/GARR-NIR/nir-it-2/atti/didbo.html

[37] https://web.archive.org/web/19980117121355/http://caristudenti.cs.unibo.it/index.shtml

which in 1988 was established to connect three departments and was highly improved in 1996[38], thanks to the collaboration of Telecom Italia and under the supervision of CeSIA (which was created in 1994)[39]. Another perspective on the recent past of this institution could be developed examining the impact of the online service AlmaLaurea, presented in May 1998[40], which aimed at improving the relationship between the institution, its student community and the job-market.

A third point of view might be focused on the relation and mutual influence between the university and the municipality, by considering the role played by the Internet. In fact, the city of Bologna and its citizens have a strong bond with innovation in computing technologies, with the municipality for instance having created one of the first civic-networks in the world in 1995, giving to all citizens free access to the Internet the very next year (Chiara, 1998). The early importance of the web for Bologna citizens appeared also in a 1996 article retrieved from *La Repubblica* digital archive. As mentioned in a piece of news[41], in November of that year, a digital discussion was censored for the first time in Italy and an entire mailing list named "Lisa" was completely closed. This happened on the Unibo server: CeSIA informed the professor of computer science Dario Maio of the presence of violent debates on the platform and the Department of Computer Science decided to drastically intervene.

As the article reported, these digital conflicts were probably related to the internal discussion of an Italian association named "La città invisibile"[42]. This association, comprising early Internet activists, was interested in shar-

[38] `https://web.archive.org/web/19990503060118/http://naomi.bo.astro.it/\~federico/almanet/smds/smdspage.html`

[39] Additional materials on this topic can be found in the bibliography dedicated to "Internet in Italy" edited by Riccardo Ridi on his website : `http://www.riccardoridi.it/esb/biblint/04.htm`

[40] `http://web.archive.org/web/20020225134330/http://www.almanews.unibo.it/Alma98.htm`

[41] `http://ricerca.repubblica.it/repubblica/archivio/repubblica/1996/11/24/troppe-parolacce-censurata-lisa.html`

[42] `http://www.citinv.it/intro.html`

ing the importance of digital cultures and rights. Among them there were also academics, for example Lucio Picci[43], at that time a young researcher at the University of Bologna[44].

It is evident that diachronically examining the digital alter ego of the institution and using these resources to extend our knowledge on its recent past is a complex challenge, which relies on both an interdisciplinary set of methodological approaches and specific research questions.

The examples presented above only highlight some specific perspectives on the topic, which we have encountered while examining the collected sources. While each one of the aspects previously presented would offer a different insight on the early use of the web by the institution, the last part of this section will focus on a different task: tracing the origin of the university website. This will help us in understanding the process of creation of Unibo.it: who was responsible for it, for what reasons was it created and in which context. Old websites hold interesting stories on their origins, which are often placed at the intersection of academic research, curiosity in advanced digital technologies (together with aspiration of contributing to them) and mutual human desire of communicating with others. Unibo.it is one of them.

7.3.2 At the Beginning of the Digital Era

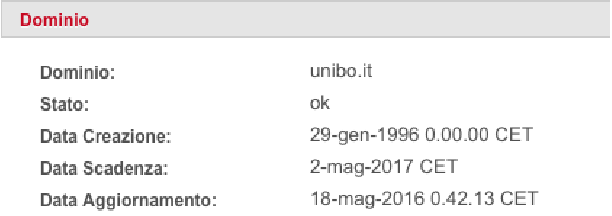
Tracing the first online presence of an entity such as the University of Bologna has not been an easy task. In Italy a list of .it server was initially maintained by the research center CNUCE (Centro Nazionale Universitario di Calcolo Elettronico) and is currently available on the website Registro.it[45]. However all early Italian websites (created before 1996) have a common creation date: 29-01-1996 (see Fig. 7.7).

The research group "GARR-Network Information Retrieval" organised a se-

[43] Currently professor of Political Economy: https://www.unibo.it/sitoweb/lucio.picci

[44] https://sites.google.com/site/lucioxpicci/storia

[45] http://www.nic.it/



Dominio

Dominio:	unibo.it
Stato:	ok
Data Creazione:	29-gen-1996 0.00.00 CET
Data Scadenza:	2-mag-2017 CET
Data Aggiornamento:	18-mag-2016 0.42.13 CET

Figure 7.7: Unibo.it creation date on Registro.it

ries of annual meetings in the early 90s[46], dedicated to the World Wide Web in Italy at the university level.
Consulting the proceedings of 1994, we learned[47] that Unibo.it was already active at least in the August of the year before. The fact that the university was fostering the use of the Internet and of web materials could be also deduced by consulting the 1993/94 yearbooks, where the importance of AlmaNet is mentioned, as well as the need of adopting emails as a form of communication and online databases as new resources.

During the first years of the World Wide Web Tim Berners-Lee curated a list of web-servers on the CERN website; the last update available is from late 1992[48]. Unibo.it is not mentioned in this list, but there is a link to another Italian research institution, the Physics Institute in Trieste.
Later, on the NCSA website, a specific section called "What's New!" published a list of the new servers on the web each month (from June 1993 to January 1996)[49]. By consulting it, some interesting information about specific sub-sections of Unibo.it was found: for example the "Bologna Astrophysics Preprints" has offered online, since November 1994, all the scientific publications of the Bologna Astronomical Observatory (OAB), the Astronomy

[46] `http://web.archive.org/web/19971025045540/http://www.cilea.it/GARR-NIR/`

[47] `http://web.archive.org/web/20010424195903/http://www.cilea.it/collabora/GARR-NIR/nir-it-2/atti/cib.html`

[48] `http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html`

[49] `http://web.archive.org/web/19961220071416/http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/archive-whats-new.html`

Department of Bologna University (DDA), the Radioastronomy Institute of CNR (IRA) and the TESRE Institute of CNR (ITE). However, for what concerns specifically the creation date of the website, in December 1993 a link to a map of all Italian web-server was published, but this link is not available anymore (it redirects to the 1997 version of the Cilea Map).
Summarising, by consulting born-digital materials as well as traditional archival sources, we know that Unibo.it was already available in the second half of 1993, and that the website was created after the end of 1992, according to the CERN web-server list.

University websites have been created most of the times by researchers working in fields where Internet was already actively adopted. For this reason, departments and research centers in computer science[50] and physics[51] are generally good starting points for discovering who created the website of a specific institution. However, in Bologna, the university website was created in a different place, namely at the Department of Mathematics, thanks to the collaboration between a Turkish professor who had at that time arrived from the United States, and a young Italian researcher.
The story of the origins of Unibo.it emerged in an interview conducted with Renzo Davoli. Davoli is currently a professor of Computer Science at the University of Bologna; in the early 90s he was working under the supervision of Ozalp Babaoglu, who arrived in Bologna in 1988 from Cornell University and wanted to use the Internet – among other things – to stay in touch with his colleagues and friends from abroad. Given the fact that Bologna did not have a Department of Computer Science at that time, Babaoglu and Davoli were working at the Department of Mathematics.
In 1988 the two of them established the second Italian node to the Internet[52],

[50] https://web.archive.org/web/19970518021303/http://www.utexas.edu/teamweb/history/

[51] https://swap.stanford.edu/19911206000000/http://slacvm.slac.stanford.edu/FIND/default.html

[52] The first node is from the research institute CNUCE, in Pisa, as described here: http://www.30annidirete.it/.

from the Department of Mathematics to CNUCE, in Pisa[53] and Davoli became the person in charge of the University TCP/IP network. They then became part of AlmaNet, the internal network initially established between the Departments of Mathematics, Engineering and Physics. In the following years, AlmaNet played an essential role in connecting university departments, especially the ones located outside the city.

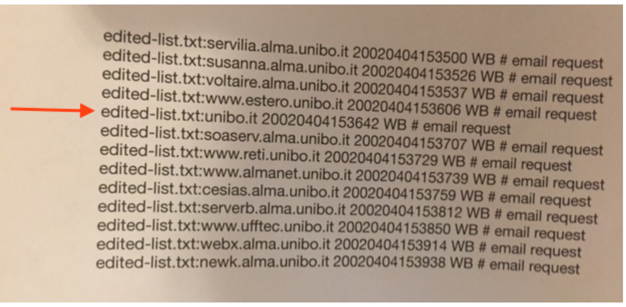
Interactions between the departments were again improved thanks to the advent of the World Wide Web. Departments were the first to be online and, once again, this was accomplished thanks to Babaouglu and Davoli. As a matter of fact, in July 1993, the two researchers registered and created the webpages of the domains `cs.unibo.it` (Computer Science) and `dm.unibo.it` (Department of Mathematics). This helped colleagues in other departments understand the huge potential of the web. Initially, Davoli and Babaoglu managed the main website as well, which then passed under the supervision of the Public Relation Office, and in particular of Salvatore Mirabella[54].

7.4 Working with the Internet Archive

In the previous sections, it has been highlighted how the digital past of an institution could be re-discovered without the prompt availability of Internet Archive snapshots. The following paragraphs will describe the work we conducted to understand why Unibo.it was excluded from the Wayback Machine. Before moving on, it is essential to mention that, when a website is excluded due to robots.txt (as the University of Trento website, between 2006 and 2010), its pages are not preserved by the Internet Archive. In the case of the exclusion issue concerning Unibo.it, as it will be presented in the next paragraphs, it was instead discovered that the Internet Archive has continued to preserve the website (despite what described in the FAQ section), which was simply not available for any kind of consultation through the Wayback Machine.

[53] This information is also offered by the The Internet Engineering Task Force (IETF): `https://tools.ietf.org/html/rfc1117#ref-AA62`

[54] `http://www.unibo.it/annuari/Annu9597/final/c2/p1/sp1/index.html`



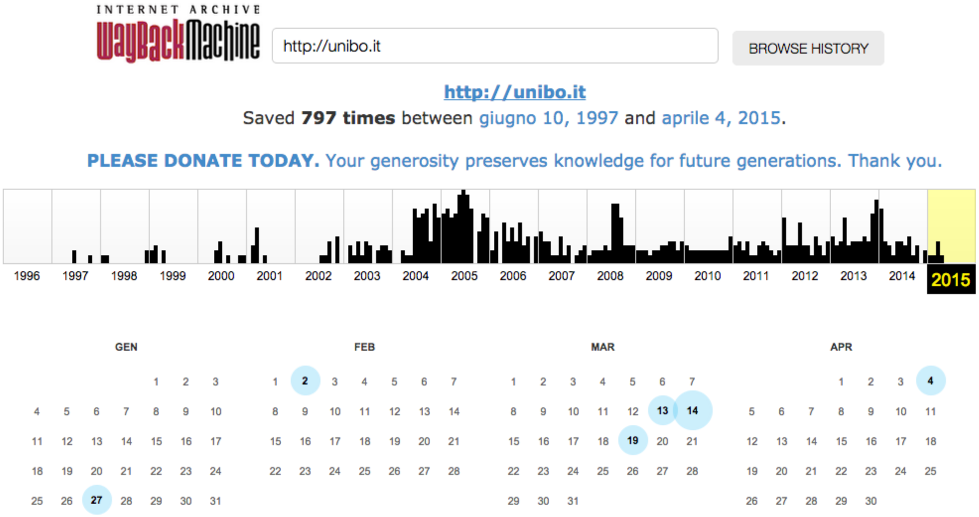
```
edited-list.txt:servilia.alma.unibo.it 20020404153500 WB # email request
edited-list.txt:susanna.alma.unibo.it 20020404153526 WB # email request
edited-list.txt:voltaire.alma.unibo.it 20020404153537 WB # email request
edited-list.txt:www.estero.unibo.it 20020404153606 WB # email request
edited-list.txt:unibo.it 20020404153642 WB # email request
edited-list.txt:soaserv.alma.unibo.it 20020404153707 WB # email request
edited-list.txt:www.reti.unibo.it 20020404153729 WB # email request
edited-list.txt:www.almanet.unibo.it 20020404153739 WB # email request
edited-list.txt:cesias.alma.unibo.it 20020404153759 WB # email request
edited-list.txt:serverb.alma.unibo.it 20020404153812 WB # email request
edited-list.txt:www.ufftec.unibo.it 20020404153850 WB # email request
edited-list.txt:webx.alma.unibo.it 20020404153914 WB # email request
edited-list.txt:newk.alma.unibo.it 20020404153938 WB # email request
```

Figure 7.8: The removal request received by the Internet Archive.

Given the specificity of the exclusion message, we decided to consult CeSIA, the team that has supervised Unibo.it during the last decades, regarding this issue. However, they did not submit any removal-request to the Internet Archive and they were not aware nor had any trace in digital and archival documents of anyone submitting it.
To clarify this issue, the Internet Archive team was then contacted. Thanks to the efforts of Mauro Amico (CeSIA), Raffaele Messuti (AlmaDL), Christopher Butler (Internet Archive) and Giovanni Damiola (Internet Archive), a collaboration with the Internet Archive started at the end of March 2015. As we contacted Butler, he told us that the Unibo.it case was similar to another recent case that involved the New York government websites[55].

With their help, we discovered that a removal request regarding the main website and a list of specific subdomains had been submitted to the Internet Archive in April 2002 (see Fig. 7.8). Thanks to this collaboration, the university website became available again on the Wayback Machine on the 13th of April 2015 (see Fig. 7.9). This also gives us the opportunity of attesting that the Internet Archive has kept preserving Unibo.it in the last fifteen years; the website was simply not available for any consultation. Additionally, having the website at our disposal, once again, gave me the opportunity of re-evaluating the findings of this study.

[55] http://www.villagevoice.com/news/why-were-new-york-government-websites-hidden-from-an-internet-archive-for-13-years-6721316



http://unibo.it

Saved **797 times** between giugno 10, 1997 and aprile 4, 2015.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.

Figure 7.9: Unibo.it is once again available on the Wayback Machine.

While the exclusion issue was solved, it was necessary to investigate further its causes. As it has already been described, in 2002 the administration of Unibo.it completely changed, during a general reorganisation of the digital presence of the university (the Portale d'Ateneo project). Therefore, while it is evident that this request was made by someone who was in the position to ask for the removal of the website[56] and who knew how the Internet Archive exclusion policy works[57], it still remains entirely unclear to us who, in that very same month, could have been in the position to submit this specific request, and the reasons behind it.

Even though several years have passed by, it was assumed that someone involved in the administration of the website would have remembered (or had traces in a backup) this email exchange with a team of digital archivists in San Francisco. Between April and June 2015 a last series of interviews was

[56] The Internet Archive says "the website owner" and, even if they happened to be not absolutely rigid on this point, it has to be someone at least involved in the management of the website.

[57] As he/she explicitly declared a specific list of subdomains to remove (as described above, the Internet Archive excludes urls and their subsections – not subdomains).

conducted with several people involved in the Unibo.it website, pre- and post the 2002 reorganisation. However, it was impossible to retrieve any information on this issue.

As the specificity of the request is the only hint that could help in identifying its author, we decided to analyse the different urls in more detail. The majority of them are server addresses (identified by "alma.unibo"), while the other pages are subdomains of the main website, for example estero.unibo.it (probably dedicated to international collaborations).
A few questions therefore remain unsolved: why would someone want to exclude exactly these pages and not all the department pages, which were active online, at that time? Why exactly these four subdomains were selected and not the digital magazine Alma2000 (alma2000.unibo.it) or the e-learning platform (www.elearning.unibo.it)?

7.5 Conclusion

The aim of this chapter has been to highlight both the issues and the potentialities of using born-digital documents to study the recent past of the University of Bologna. The main focus was to describe the methodological approach employed in order to reconstruct its website (which has been excluded for the last thirteen years from the Wayback Machine).
In doing so, the chapter underlined how its history is divided into two parts (before and after the setting up of the "Sistema Portale d'Ateneo") and how different sources (the yearbooks, materials from foreign web archives, document preserved by CeSIA, articles on local, national and digital newspaper) have been useful to improve our knowledge on the metamorphosis of this website (specifically, on the role of department pages). This work also examined how born-digital sources can offer new insights on common research topics related to the history of this university and its relation with the students' community and the city itself.

The different issues presented in this chapter highlight the need of an even more interdisciplinary approach for future historians. In the field of Internet studies and digital archiving, researchers are already discussing the importance of new ways of conceiving the retrieval, analysis, criticism and employment of born-digital primary sources. As historians, we should openly join this discussion with both theoretical contributions as well as concrete examples. As a matter of fact, these materials will sustain traditional historical research questions and will lead to an infinite number of new ones.

Chapter 8

Bridge: The Collected "New" Primary Sources

Let's make the best
of the situation.

This short chapter connects Part II & III of this thesis by offering an overview of the large collections of academic sources I retrieved from the reconstructed University of Bologna's website.

8.1 Introduction

As remarked before, historians of universities have focused so far on topics such as the role of academic institutions in the process of Nation building as well as on the bi-directional influence between governments and academia. Among these studies, another topic that has attracted vast attention is studying the way education is provided and how teaching and research are influenced by several political, economic and social factors. For example, in Rüegg (2004, 2011) a complete overview of the most recurrent teaching topics in European academic institutions and how and why they have changed in the last centuries is offered. Examining what has been taught and studied during the years at a specific academic institution and understanding the global and local reasons that triggered particular changes has attracted not only the attention of many historians of universities, but also of STS researchers as well as the scientometrics community.

8.2 An Overview of Academic Input and Output

The study of university websites and of academic born-digital sources, bring new materials to the surface that researchers can employ for studying in depth these topics. In the following pages, we give a general overview of the type of sources that we have collected when reconstructing Unibo.it and that could sustain future research on the recent past of this academic institution.

Syllabi In order to study what has been taught to a specific institution, traditionally historians obtain this information by examining different sources, such as student transcript of lessons and professor's notes[1] as well as university yearbooks and the widespread adoption of specific textbooks.
Dan Cohen remarked on the large still unexplored potential that online syllabi could offer to the study of the recent history of academic teaching (Cohen, 2006a, 2011). Its first digital-historical work on the topic has been published in 2005 in the Journal of American History (Cohen, 2005). The focus of the paper is to show how the teaching of history in U.S. universities is still strongly based on textbooks. These findings, which go in the opposite direction of the opinion of a round table published in the same journal four years before (Kornblith and Lasser, 2001), were made possible thanks to a large study of around 800 syllabi available on the web.
While Cohen's work is a first step in this direction, the potential of online syllabi goes beyond his study and could rapidly affect the practices, topics and findings of historians of universities as well as the scientometrics com-

[1] For example, a great resource to know what Professor Pasquini taught about Dante at the University of Bologna are the notes from its course from the academic year 1992-93 (Pasquini, 1993).

munity. During the reconstruction of the past of Unibo.it, we were in fact able to collect a vast collection of syllabi published online by the institution, which are in the number of thousands for each academic year[2].

Dissertation and Academic Articles Studying the output of academic institution has been the main focus of scientometrics studies. The topic attracted also the attention of historians and science and policy researchers, who aimed at understanding the forces that foster doing research on specific topics or the impact of the private sector on the research conducted in public universities. While in the last decades there has been a large body of work in scientometrics for establishing new quantitative measures for evaluating the scientific output of institutions, most of the methods applied exploit bibliometrics measures such as co-author networks and citation graphs. The downside of such techniques is that they obviously depend on the availability of the bibliometrics network data, which can often be difficult to obtain. However, if we consider the born-digital resources we collected in our work, such as Ph.D. dissertations and scientific articles, it is generally much easier to obtain the full-text of these publications[3] than the bibliometrics network data. For what concerns in particular academic dissertations, their prompt availability on digital databases such as university digital libraries permit to obtain a diachronic overview of the academic output of an institution in its entirety (Ramage, 2011).

Research Grants Much of the attention in the scientometric community has focused on quantifying the quality of research by studying scientific publications (Wagner et al., 2011). Database of grant awards, available on the live web or reconstructed using web archive snapshots, provide a unique opportunity for examining research inputs (grant proposals and awards) rather

[2] From the live version of the website is possible to promptly collect all syllabi from the academic year 2004-2005. Using Internet Archive snapshots we were able to identify all courses programs for several departments, for example we collected all syllabi of courses in History since 1998-99.

[3] As on DART-Europe: `http://www.dart-europe.eu`

than outputs (publications). As Nichols (2014) remarked in her study on the National Science Foundation awards:

> Research proposals contain a broader scope of data on the people, inputs and processes of science than is typically contained in publication data. Research publications report the narrow outcomes that emerge from ongoing research programs and limit the scope of the reporting to specific findings or results. Proposals, on the other hand, describe overarching research programs, which typically generate multiple publications.

By studying for example the content of the Horizon2020 calls we can examine how political and economic factors can influence the way a grant is shaped and presented. Additionally, project websites offer large amounts of academic information[4] regarding how the work is conducted in a department, who are the people involved, which new collaborations are established[5].

8.3 A Common Issue: Unstructured Texts

The different born-digital materials presented above are resources we collected during our research and that can sustain studies on the recent past of academic institutions from different points of view and with diverse goals. While collecting them could be done following procedures such as the ones described in previous chapters, being able of studying them is currently extremely challenging. As a matter of fact, these sources are, first of all, huge in numbers and second of all they often exist only in digital format. For example, when we consider the syllabi collected on Unibo.it we talk about thousands and thousands of documents. An even larger number emerges when we consider academic dissertations promptly offered on platform such as the portal DART (Digital Access to Research Theses - Europe), which

[4] To know more see the recent work by Bicho and Gomes (2016) titled “Preserving Websites Of Research and Development Projects”.

[5] For example, see the information offered here: `http://www.unibo.it/en/research/projects-and-initiatives/research-projects-horizon-2020-1`

aggregates university repositories of Ph.D. theses from all around Europe. The dissertations currently offered on DART are over 700.000.

It is important to bear in mind that abundance per se is not a problem when dealing with the recent past of academic institution. Scientometrics provides us of a large variety of methods for analysing large collections of publications, which can be adopted to get a first overview even by traditional qualitative historians. However, the issue that emerges when considering the sources previously introduced is the fact that they are all available in unstructured text[6], in opposition to the bibliographic network traditionally adopted in scientometrics.

This issue forces historians that intend to use these materials in considering a more interdisciplinary approach, that combines the traditional historical method with solutions from the field of natural language processing, where researchers have been dealing with the abundance of digital materials for decades. As we remarked in the first part of this thesis, while in traditional digital humanities project the use of text mining solutions has always been an option, in a constant comparison between close and distant reading, when researchers are dealing with collections of born-digital sources these techniques appear to be the only solution available. As a matter of fact, if we consider the previously described collections of Ph.D. theses, the only solution available for retrieving specific content is adopting a search tool. For this reason, in the next part and through a series of case study, we will remark on the importance of approaching, adopting, developing and testing text mining methods in a strongly critical way, always highlighting their potential and limits and constantly describing how/whether these approaches will impact on the way historians collect, analyse and select sources.

[6] Unstructured in the sense that it is information that either does not have a pre-defined data model or is not organised in a pre-defined manner. This presents irregularities and ambiguities that make it difficult to process using traditional programs as compared to data stored in fielded form in databases or annotated in documents.

Part III

How to Deal with Abundance

Chapter 9

Retrieval and Exploration of Academic Content

I've seen everything imaginable
pass before these eyes.

This chapter is based on my experience as a co-founder of the startup "FiND", which won in 2014 the Working Capital grant for innovative startups. The project was coordinated by Professors Maurizio Matteuzzi (supervisor of my thesis) and Giovanna Cosenza and was driven by the goal of improving the retrieval process of academic content on university websites, which I initially presented in Nanni (2014). This chapter also presents the initial findings of my research on using text mining methods for exploring born-digital collections of doctoral dissertations. I conducted these works during my internship at the Human Language Technology group of the Foundation Bruno Kessler (Trento).

9.1 Introduction

In the previous chapters we introduced the datasets we collected while reconstructing the University of Bologna website. As we already remarked, methods from the field of natural language processing (NLP), such as information retrieval approaches and text mining methods, are the only solutions at our disposal for dealing with the abundance of collected unstructured texts. For this reason, in this chapter we start by presenting two different ways of exploring and retrieving information from this collection.

In the first part, we remark on the fact that advanced NLP and information retrieval (IR) solutions could enhance the retrieval of academic content in the large number of syllabi we collected. We also present how we extracted information from multimedia materials, such as academic video lessons.

In the second part of the chapter, we offer an overview on the potential of LDA topic modeling for dealing with the vastness of materials offered by the university digital library. In particular, we conduct a series of text explorations over its collection of Ph.D. dissertations.

9.2 FiND: Semantic Retrieval of Academic Content

Retrieving academic content from university websites currently remains a difficult task, due to different reasons. If we consider for example the case of Unibo.it, we can notice that these materials are presented on the website in different format (text content, video lecture, etc.); additionally, the university currently provides only a basic string-matching search tool on the website, without any kind of advanced-search that employs metadata information, such as the year when the document was created or the type of document (syllabus, news, grant proposal, etc.).

The impact of this issue is multilayered, as it affects both normal users of the website (such as students) as well as companies and external researchers that intend to establish collaborations. Additionally, in the specific case of our research it also affects us, as it impedes us from conducting an in-depth analysis of the corpus.

In order to tackle this issue and enhance information retrieval on academic websites, in 2014 we started a research project at the department of Philosophy and Communication Studies. In the same year, we participated and won the Working Capital grant for innovative startups, financed by the Italian telecommunication company Telecom Italia. Thanks to this grant, we had the opportunity of implementing a prototype of a semantic search tool, that we called FiND, which goes beyond the simple string matching of standard search engines.

9.2.1 Approaches Adopted

In the research project that led to the development of FiND, we started by focusing on the retrieval of academic materials. On the University of Bologna website these contents are generally presented in two different formats: syllabi and video lectures.
In the following paragraphs we offer an overview of the computational solutions we adopted in order to *a)* associate each content with a descriptive text and *b)* use this text to improve the retrieval of the content, given a user query. While the implementation and case-study is focused on content in Italian, the same approach could be adopted to process content in a different language.

Speech to Text. The University of Bologna offers a series of video-lecture on its website (and Youtube account[1]), where researchers and professors give an overview of their work at the university, spanning from teaching to research. In order to process these materials, we adopted the Google speech-to-text API available through the tool Dictation[2]; additionally we worked on improving the transcription speed of this solution, becoming able to finally process a video in 1/5 of its time. During the research project a series of qualitative evaluations have been conducted on the obtained output; these studies have shown solid performances for the recognition of concepts and common nouns (e.g. history, science), while it encounters problems when dealing with proper or domain-specific nouns (while Darwin was correctly

[1] `https://www.youtube.com/user/UniBologna/`

[2] For more information see: `https://dictation.io/`

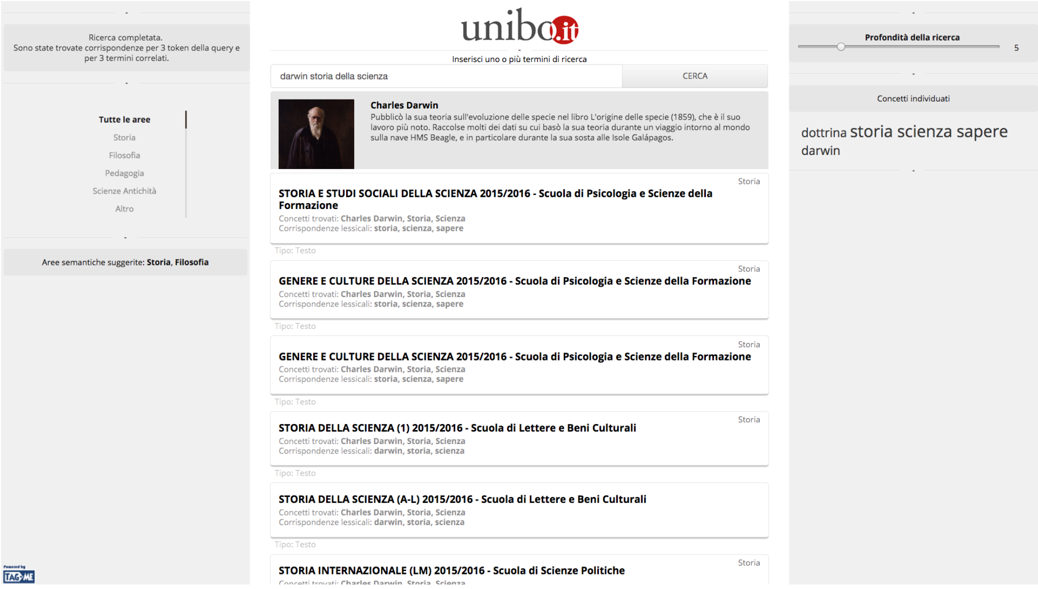
transcribed, terms such as “hepatitis B virus” were misspelled).

Text Pre-Processing. At this stage the different types of content are associated with a descriptive text: each syllabus has a textual description and videos have their transcription. In order to be retrieved, given a user query, documents have been tokenised, lemmatised and POS tagged. We did so in order to identify central concepts and verbs. The pipeline has been initially developed using TextPro (Pianta et al., 2008); in the second part of the project, we re-implemented both the part of speech tagger and lemmatiser.

Entity Linking. Next, we adopted an entity linking system (TagMe (Ferragina and Scaiella, 2010)) in order to identify relevant entities and concepts in the textual content. For example, we used it to link the expression “the great war” to the DBpedia entry “World_War_I”. This solution, which supported us in identifying and disambiguating relevant pieces of information, helped us in moving from string-match to semantic search.

Weighting We identified the most relevant terms, entities and concepts in each document by computing their occurrence in the document, weighted over all documents in the collection mentioning them, namely the term frequency–inverse document frequency (TF-IDF) weighting (Sparck Jones, 1972). Based on TF-IDF, the weight of a term (or of the linked concept/entity) will be the highest if it has a high frequency within a few documents. By converse, weights will be lower for those entities which occur fewer times in a document, or occur in many documents.

Query Preparation and Expansion. User queries were processed following the same steps, in order to identify relevant terms, concepts and entities. Next, queries were expanded through a series of resources, such as a manually compiled list of synonyms for nouns and verbs, and concepts related to the ones identified by using a simple relatedness measure (Witten and Milne, 2008).



Ricerca completata.
Sono state trovate corrispondenze per 3 token della query e per 3 termini correlati.

Aree semantiche suggerite: **Storia, Filosofia**

Inserisci uno o più termini di ricerca

darwin storia della scienza

Charles Darwin
Pubblicò la sua teoria sull'evoluzione delle specie nel libro L'origine delle specie (1859), che è il suo lavoro più noto. Raccolse molti dei dati su cui basò la sua teoria durante un viaggio intorno al mondo sulla nave HMS Beagle, e in particolare durante la sua sosta alle Isole Galápagos.

Storia
STORIA E STUDI SOCIALI DELLA SCIENZA 2015/2016 - Scuola di Psicologia e Scienze della Formazione
Concetti trovati: **Charles Darwin, Storia, Scienza**
Corrispondenze lessicali: **storia, scienza, sapere**
Tipo: Testo

Storia
GENERE E CULTURE DELLA SCIENZA 2015/2016 - Scuola di Psicologia e Scienze della Formazione
Concetti trovati: **Charles Darwin, Storia, Scienza**
Corrispondenze lessicali: **storia, scienza, sapere**
Tipo: Testo

Storia
GENERE E CULTURE DELLA SCIENZA 2015/2016 - Scuola di Psicologia e Scienze della Formazione
Concetti trovati: **Charles Darwin, Storia, Scienza**
Corrispondenze lessicali: **storia, scienza, sapere**
Tipo: Testo

Storia
STORIA DELLA SCIENZA (1) 2015/2016 - Scuola di Lettere e Beni Culturali
Concetti trovati: **Charles Darwin, Storia, Scienza**
Corrispondenze lessicali: **darwin, storia, scienza**
Tipo: Testo

STORIA DELLA SCIENZA (A-L) 2015/2016 - Scuola di Lettere e Beni Culturali
Concetti trovati: **Charles Darwin, Storia, Scienza**
Corrispondenze lessicali: **darwin, storia, scienza**
Tipo: Testo

Storia
STORIA INTERNAZIONALE (LM) 2015/2016 - Scuola di Scienze Politiche
Concetti trovati: **Charles Darwin, Storia, Scienza**

dottrina storia scienza sapere darwin

Figure 9.1: Results presented by FiND.

Cosine Similarity. Given a query, we ranked results measuring the cosine similarity between the TF-IDF vector representations of the query and the document.

9.2.2 Prototype

The developed platform[3] allows the user to retrieve syllabi and video lectures. Given a query, such as “darwin storia della scienza” (”darwin history of science”), the tool processes the query and offers a series of information, as shown in Figure 9.1. First of all, the platform shows a brief description from Wikipedia of the most relevant concept identified in the query (in this case “Charles Darwin”). Secondly, on the left it gives the user the possibility of filtering results based on specific disciplines (such as “History” or “Philosophy”). On the right it permits to tune the query expansion, adding or removing related concepts; in our case two concepts are added to the query: “knowledge” and “doctrine”, which have been found to be related to the

[3] A prototype is available at: `http://unibo.vfind.it/` or, in its initial version, at: `http://www.queryandfind.it/unibofind.html`

concept “history of science”.

9.3 Corpus Exploration with LDA

During the last decade, humanities scholars have experimented with the potential of different text mining techniques for exploring large corpora, from co-occurrence-based methods (Buzydlowski et al., 2002) to automatic keyphrase extraction (Hasan and Ng, 2014; Moretti et al., 2015) and sequence-labeling algorithms, such as named entity recognition (Nadeau and Sekine, 2007). Among all the different approaches, the Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003) has become in recent years one of the most employed unsupervised techniques (Meeks and Weingart, 2012). Humanities scholars appreciate its capacity of detecting the presence of a set of meaningful categories called “topics” in a collection of texts (Underwood, 2012; Bogdanov and Mohr, 2013; Jockers, 2014). Additionally, the DH community has often remarked LDA’s potential for serendipity (Alexander et al., 2014) and for distant reading analyses (Leonard, 2014; Graham et al., 2016), i.e. studies that move beyond text exploration.

Latent Dirichlet Allocation is a generative probabilistic model of a corpus, where each document is represented as a random mixture over latent topics and each topic is identified as a distribution over words. LDA can be considered as an improvement of the probabilistic latent semantic analysis (Hofmann, 1999), by adding the assumption that the topic distribution has a Dirichlet prior. Given the fact that LDA is an unsupervised approach it does not need training data; the user should simply specify in advance the number of topics (k) that he/she intends to identify and will receive as an output the generated topics (effectively sets of words, as can be seen for example in Figure 9.2).

Studying the distribution of the obtained topics in different corpora has been adopted for several purposes in digital humanities scholarships (Meeks and Weingart, 2013), from exploring large collections of documents (Yang et al., 2011) to highlighting differences in scientific publications (Hall et al., 2008).

Among the DH community, LDA topic modeling has attracted in particular the interest of the digital history community (Brauer and Fridlund, 2013). This fascination of digital historians for a natural language processing method is a very interesting fact, as traditionally this community has focused on digital preservation, public history and geographical information systems, rather than on text analysis (Robertson, 2016) (as we already remarked in the first chapters of this thesis). We argue that this change has happened because topic modeling proposes a solution to a precise need that brings together historians as well as political scientists (Grimmer and Stewart, 2013; Slapin and Proksch, 2014) and other researchers whose established methodologies rely on digging in large analogue archives. Topic modeling, in its simplicity of use[4] and well hidden complexity (Underwood, 2012; Weingart, 2012), represents that "compass" that a historian has always needed when examining a large collection of sources. As a matter a fact, it promises to promptly offer to the researcher: *a)* a general overview of a specific collection by capturing the different meanings of words, *b)* a clear division of the collection in subparts (i.e. topics) and *c)* a quantification of the relevance of each topic for each document.

9.3.1 Using LDA for Identifying Interdisciplinary Research

During the last decades, policy makers, government agencies, and private companies are trying to encourage more interdisciplinary research than ever before, with the goal of addressing complex challenges and accelerating innovations across industries and branches of knowledge (Holm et al., 2013). Interdisciplinary practices are sustained by dispensing grants, scholarships, and establishing direct collaborations with the academia.

As recently remarked in a special issue of Nature dedicated to the topic,[5] researchers involved in these projects are attempting greater and greater

[4] See for example `http://programminghistorian.org/lessons/topic-modeling-and-mallet`

[5] `http://www.nature.com/news/interdisciplinarity-1.18295`

challenges combining techniques from multiple disciplines, blurring the traditional borders between academic areas. At the same time, the creation of large interdisciplinary teams in order to attempt challenging projects has often attracted the attention of the media and consequently of society[6].

Historians of higher education, who aim to investigate for example whether rhetoric and funding have been playing a significant role on orienting the research focus of universities towards being more interdisciplinary, now have at their disposal digital databases of dissertations (such as the one collected from Unibo.it), which diachronically reflect academic outputs in their entirety. However, given the enormous amount of primary sources readily available, the identification of interdisciplinary theses in these corpora needs to be conducted adopting computational methods.
We decided to use LDA, given the fact that previous authors (Ramage, 2011; Nichols, 2014) have remarked on the potential of topic models for tracking interdisciplinary collaborations and extra-departmental interactions around specific topics through collections of academic materials[7].

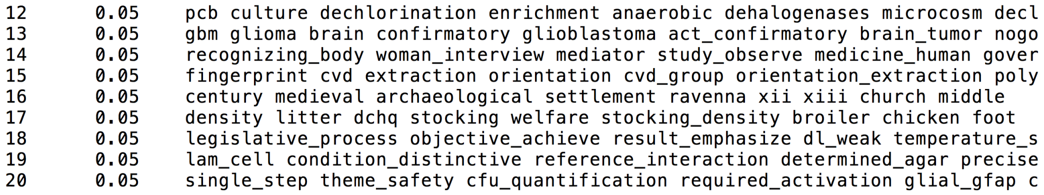
Dataset

In this research, LDA topic modeling has been adopted in a series of exploratory studies over the corpus of Ph.D. dissertations offered on the University of Bologna digital library platform[8] (Vignocchi et al., 2010).
This corpus provides access to a total of 2954 dissertations with an English abstract (published between 2007-2015). The mean length of these

[6] See for example the article "Why it's time to get real about interdisciplinary research" on the Guardian: `https://www.theguardian.com/science/political-science/2015/oct/14/why-its-time-to-get-real-about-interdisciplinary-research`

[7] Ramage (2011) in particular adopted an extended version of Labeled LDA for analysing interdisciplinary collaborations in the Stanford Corpus of Ph.D. Dissertations; in his case each thesis was associated with a set of related disciplines and sub-disciplines. Given the fact that the thesis offered on Unibo.it did not have any information regarding secondary disciplines, we could not use Labeled LDA; for this reason we opted for conducting a general study with standard LDA, and we examined the results filtering by the metadata at our disposal (namely: supervisors, department, main discipline, etc).

[8] `http://amsdottorato.unibo.it/`



```
12      0.05    pcb culture dechlorination enrichment anaerobic dehalogenases microcosm decl
13      0.05    gbm glioma brain confirmatory glioblastoma act_confirmatory brain_tumor nogo
14      0.05    recognizing_body woman_interview mediator study_observe medicine_human gover
15      0.05    fingerprint cvd extraction orientation cvd_group orientation_extraction poly
16      0.05    century medieval archaeological settlement ravenna xii xiii church middle
17      0.05    density litter dchq stocking welfare stocking_density broiler chicken foot
18      0.05    legislative_process objective_achieve result_emphasize dl_weak temperature_s
19      0.05    lam_cell condition_distinctive reference_interaction determined_agar precise
20      0.05    single_step theme_safety cfu_quantification required_activation glial_gfap c
```

Figure 9.2: Examples of topics generated using Mallet.

abstracts is around 300 tokens. Each dissertation has been assigned to one main discipline from the list of 28 disciplines defined by the Italian Ministry of Education, University, and Research[9]. As expected, the distribution of dissertations over disciplines is heavily skewed (ranging anywhere from 322 for medicine to 7 for oriental studies). The following example snippets of dissertation abstracts from three disciplines demonstrate the highly technical content of these abstracts.

> **Medicine.** *IL-33/ST2 axis is known to promote Th2 immune responses and has been linked to several autoimmune and inflammatory disorders, including inflammatory bowel disease (IBD), and evidence show that it can regulate eosinophils (EOS) infiltration and function.*
>
> **Computer Science.** *We have tried for the first time to explore the relation among mutations at the protein level and their relevance to diseases with a large-scale computational study of the data from different databases.*
>
> **History.** *The present study aims at assessing the innovation strategies adopted within a regional economic system, the Italian region Emilia-Romagna, as it faced the challenges of a changing international scenario.*

9.3.2 Topic Modeling in Action

In the next sub-section we report some of our findings on mining interdisciplinary research using topic models, which are obtained using the toolkit

[9] We will expand the description of the dataset in the next chapters.

Mallet[10], with 250 topics and 10.000 iterations.

Other authors have already remarked on the fact that LDA results are often difficult to interpret (Chang et al., 2009): for example, as you can see in Figure 9.2, while it is easy to detect that topic 16 is related to archeological studies related to the Middle Age in Ravenna, other topics (e.g., topic 14 and 17) are cryptic. To face this issue a common strategy is to employ visualisation techniques (Chuang et al., 2012a; Chaney and Blei, 2012). In our research, in order to understand *a)* how these topics are distributed across our dataset and *b)* whether some of these topics could guide us in retrieving interdisciplinary dissertations, we examine the output of Mallet using visualisations created adopting RapidMiner[11] and Gephi[12].

Visualisation of Results Using RapidMiner

Firstly we use RapidMiner, as it can be seen in Figure 9.3. On the X axis each column is associated with a different discipline (for example "agr" is Agriculture, "bio" is Biology, "m-sto" is History), on the Y axis the value indicates how relevant a specific topic (in this case topic 168: "fruit plant quality stress phenolic milk farm fatty_acid") is. Theses are represented by dots; they are associated with *a)* their discipline, for example a thesis in "Law" will be in the column "ius", and *b)* by a value that represents how relevant that topic is for that thesis.

As it can be seen, the topic 168 is very relevant for dissertations in Agriculture, but we can notice a series of outliers in other disciplines. In particular we marked one dissertation from the field of history; this is an interdisciplinary study titled "Agricultural Genetics and Plant Breeding in Early Twentieth-Century Italy"[13], which therefore is focused both on history and

[10] `http://mallet.cs.umass.edu/topics.php`; Mallet is one of the most adopted implementations of LDA topic models. It takes a collection of raw texts (i.e. .txt) as inputs and provides as output a list of topics (each of them described with a series of relevant words, as you can see in Figure 9.2) and their "relevance" for each of the documents..

[11] `https://rapidminer.com/`

[12] `https://gephi.org/`

[13] `http://amsdottorato.unibo.it/5680/`

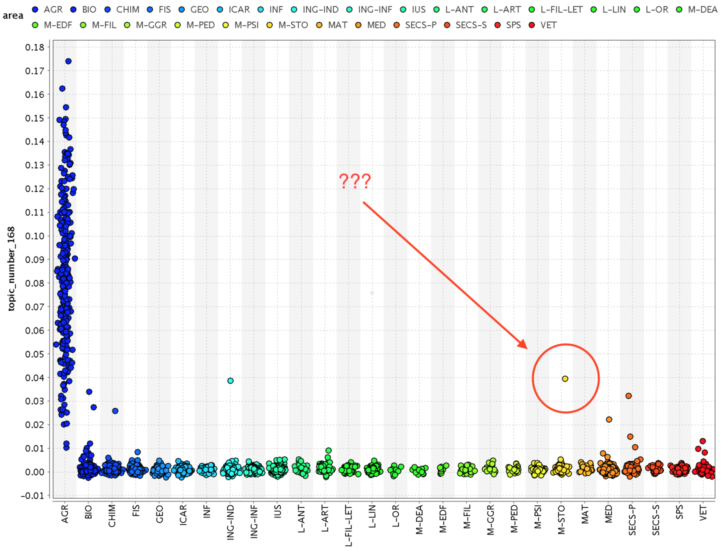


Figure 9.3: Identifying outliers in topic model results using Rapid Miner, topic words: *fruit plant quality stress phenolic milk farm.*

on agriculture.

A second example using Rapid Miner is presented in Figure 9.4. In this case, the topic 83 (topic words: "cell gene expression protein treatment role effect increase human") is highly relevant for many areas, such as Agriculture, Medicine, Veterinary and, especially, Biology. Also in this case, we highlighted an outlier, this time from the field of Computer Science ("inf"). This dissertation is again an interdisciplinary work (this time between computer science and biology), titled "Investigating the role of single point mutations in the human proteome: a computational study"[14].

Visualisation of Results Using Gephi

Motivated by these initial findings regarding the detection and retrieval of interdisciplinary works in the corpus, we decided to employ another tool to explore in a more fine grained way the distribution of topics in the collection:

[14] `http://amsdottorato.unibo.it/3363/`

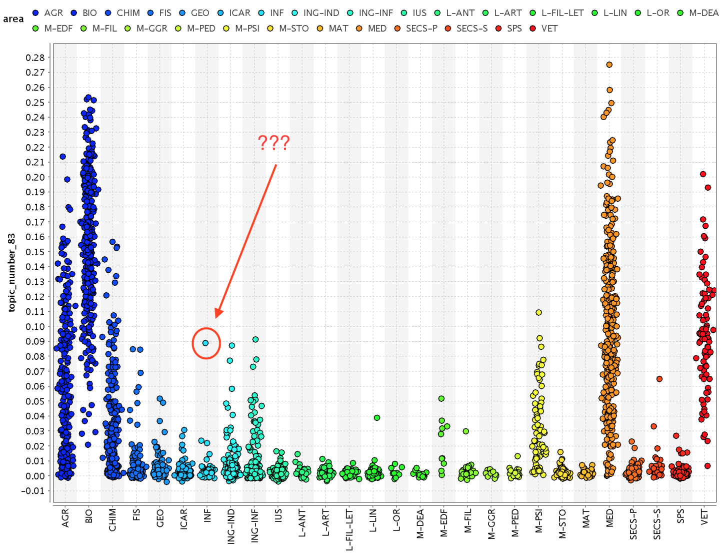


Figure 9.4: Identifying outliers in topic model results using Rapid Miner, topic words: *cell gene expression protein treatment role effect increase human.*

Gephi[15], a network analysis and visualisation tool.

In order to conduct this analysis we examined the relation between topics and theses in a different way. For each LDA topic, each thesis is associated with a value that represents how relevant that topic is for that thesis. Therefore each thesis can be represented as a vector of LDA topic values, for example:

$$thesis_A = [ValueTopic1, ValueTopic2, ValueTopic3, etc.]$$

Next, we computed a so-called "centroid" for each discipline, following what Chuang et al. (2012b) did in order to detect interdisciplinary theses in the Stanford Dissertation Corpus. Each centroid can be described as the "center of mass" of all the members belonging to that discipline, as it is represented by the mean of all vectors of theses in that discipline.

Finally, we computed the cosine similarity, which is a measure of similarity between two vectors, between the vector of each thesis and the 28 centroid vectors. This gave us the possibility of representing the similarity between

[15] https://gephi.org/

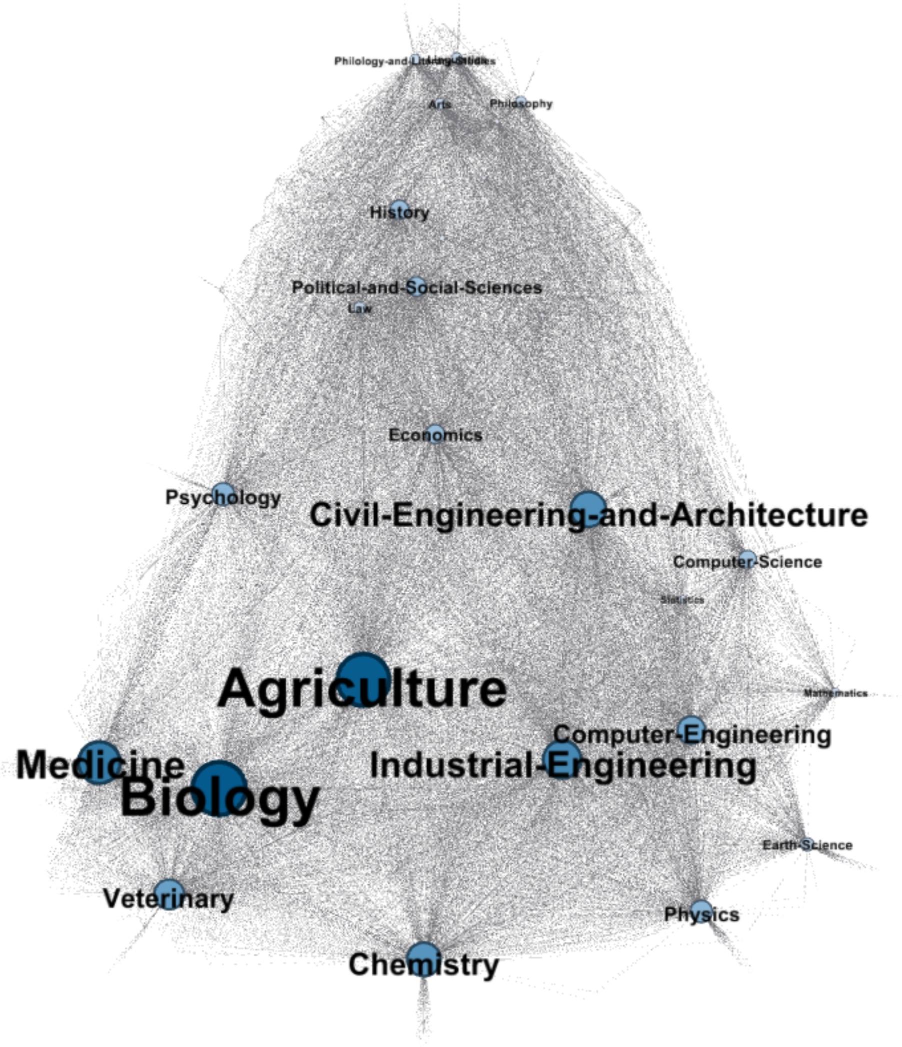


Figure 9.5: A visualisation of the entire collection using Gephi.

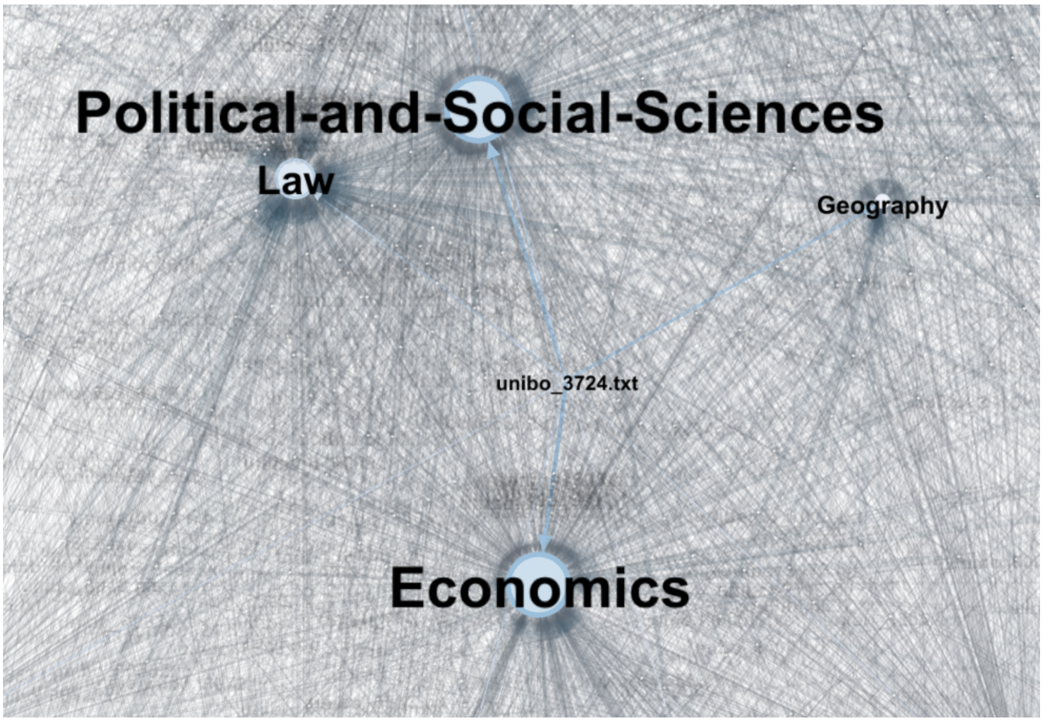


Figure 9.6: Zoom-in for exploring the connections of a specific thesis.

theses and disciplines in a graph through weighted nodes (using the cosine similarity value). A thesis is therefore connected with all centroids[16], and the weight of the connecting node represents how relevant is that discipline for that thesis.

In Figure 9.5 we present the obtained visualisation. The size of the centroids is given by the number of theses belonging to that discipline, the intensity of colour by the number of theses linked to it. The layout has been obtained by running the algorithm ForceAtlas.
Focusing on specific theses, we can see how their position often reveals interdisciplinary practices. For example, if we consider thesis_3724 (see Figure 9.6), which title is "Technological innovation in Emilia-Romagna: knowledge, practice, strategies"[17], it is indeed positioned closed to Law, Geography, Po-

[16] We adopted a threshold (0.3) in order to create the graph avoiding memory issues.

[17] `http://amsdottorato.unibo.it/3724/`

litical Science and Economics.

9.3.3 It is Time for Tool Criticism

In the previous paragraphs we remarked on the fact that a series of visualisation techniques have the potential of supporting us in collecting, analysing and selecting interdisciplinary works beyond the barriers imposed by the unstructured nature of the examined documents (as Chuang et al. (2012b) already emphasised). Writing on the use of topic models for text exploration, Trevor Owens remarked:

> If you shove a bunch of text through MALLET and see some strange clumps clumping that make you think differently about the sources and go back to work with them, great. [...] If you aren't using the results of a digital tool as evidence then anything goes. (Owens, 2012)

However, when a researcher intends to move from exploratory studies over a collection of unstructured sources (such as the ones conducted in the previous pages) to use the obtained results as quantitative analyses (for example in tracking and comparing the growth in interdisciplinary collaborations between Italian universities over the last decades), he/she has to first define its assumptions and evaluate in a highly critical way the adopted method. This is fundamental, even if previous studies have already remarked on the potential of topic models for detecting interdisciplinary materials in corpora of scientific publications (Ramage, 2011; Chuang et al., 2012b; Nichols, 2014) and especially when the results of an exploratory study seem to be so promising.

For all these reasons, in the next chapters we will first discuss the importance of tool criticism in digital humanities research, in particular for what concerns the use of LDA. Next, we will conduct a quantitative evaluation on the task of interdisciplinary theses detection, comparing LDA to a series of other approaches. This will highlight the real quality of LDA topic model outputs for this task.

Chapter 10

Tool and Tool Criticism

Now I got a job
but it don't pay.

This chapter discusses how to critically assess the reliability of text mining methods. It has been initially conceived starting from the paper "Semi-supervised textual analysis and historical research helping each other: some thoughts and observations", which I wrote together with my co-supervisor Prof. Simone Paolo Ponzetto and Prof. Hiram Kümper. This article has been published in the 10th anniversary special issue of the International Journal of History and Arts Computing (previously History and Computing) dedicated to the topic "The Future of Digital Methods for Complex Datasets". I worked on this paper while being a visiting student at the Data and Web Science Group and at the Historical Institute of the University of Mannheim. This chapter also addresses the advantages and limits of Latent Dirichlet Allocation in humanities research. I already wrote about this topic in the paper "Entities as Topic Labels: Improving Topic Interpretability and Evaluability Combining Entity Linking and Labeled LDA", which appears in a special issue of the Italian Journal of Computational Linguistics, dedicated to the topic of "Digital Humanities". The paper is co-authored with Anne Lauscher, Pablo Ruiz Fabo and Prof. Simone Paolo Ponzetto.

10.1 Introduction

As presented in the previous chapter, text mining solutions offer fascinating ways of navigating large corpora, while information retrieval systems provide lists of documents that should satisfy our information need.

However, when dealing with these methods a few specific issues always emerge: first of all, these solutions are often *black-boxes* and it is not completely clear to the final user what happens in the background when he/she adopts a IR tool such as Apache Lucene[1] or a text mining solution like Mallet. Secondly, even though some of these techniques have been largely used for NLP tasks, it is not immediately obvious whether they will work with the same performances when adopted for digital humanities research tasks[2]. Finally, and most importantly, all computational tools are based on specific assumptions (for example: topics in text are composed by words that often co-occur together): before employing these methods it is important to understand whether the underlying assumption on which they have been developed is in line with the humanities assumption (is the definition of topic assumed in topic models in line with what we intend when we want to identify topics in a collection of doctoral dissertations?).

Given these issues, in this chapter we first introduce and discuss a series of established evaluation practices in the field of data science, which could be adopted in digital humanities research for assessing the performances and error-modes of specific tools. Next we present a precise overview of different solutions for evaluating topic modeling results. This will sustain us, in the next chapter, in establishing the reliability of topic modeling for identifying interdisciplinary dissertation in the Unibo.it corpus.

[1] `https://lucene.apache.org/`

[2] An interesting example is presented in (Ritze et al., 2009), where the Stanford NER is used to identify the names of ships in the MarineLives Corpus, a collection of manually-transcribed historical records of the English High Court of Admiralty between 1650 and 1669.

10.2 Critically Assessing Tools

In order to critically analyse the performance quality of tools, the two fundamental pre-steps are to *a)* precisely define our assumptions in advance and examine if they fit with the computer science assumptions behind a specific method and *b)* conduct a solid quantitative evaluation of the adopted method for a specific research task, in order to know its reliability and error-modes. As a matter of fact, without this step, we cannot be confident of the reliability of the obtained results and use them as quantitative evidence in support of a claim. In the next paragraphs, we offer an introductory example on the need of tool criticism and present a set of useful metrics, commonly adopted in the field of data science.

Imagine a setting where the purpose is to collect articles related to a specific subject (i.e. the political relations between the USA and Cuba) and in order to do that we adopt an off-the-shelf information retrieval tool, such as Lucene or Galago[3]. An IR system that adopts query likelihood for example assumes that our query is a sample drawn from a language model: given the query Q and a specific document D, the likelihood of "generating" Q the with a model is estimated, based on D. If we decide to adopt the specific IR system, it is important to bear in mind that the documents retrieved could be not the only articles about the subject, and that maybe they are not even about that specific subject at all – due to the errors in the NLP pre-processing step, or in the ranking process or even the automatic learning phase (if the system adopts machine learning). They could be for example articles related to a baseball match between the two countries. Therefore, if we want to use the collected data as evidence for supporting a specific argument or for confirming a hypothesis, we always have to know the quality of the adopted system first.

It is interesting to notice that this specific process would sound perfectly ordinary if we were not talking about machine learning methods, computers and algorithms. When a researcher wants to be sure on his/her viewpoint

[3] `https://sourceforge.net/p/lemur/wiki/Galago/`

('I believe this article is focused on the political relations between the USA and Cuba' [4]), she/he will ask other colleagues. The process described here is the same: we need evaluation metrics and human assessments (for example articles marked as 'being focused on the political relations between the USA and Cuba' and not) in order to confirm that our hypothesis (what the machine is showing us are articles related to the political relations between the USA and Cuba) is correct.

10.2.1 Manual Annotations and Inter-Annotator Agreement

Collecting human assessments for evaluating a system is a complicated and time-consuming process. During the last ten years, in computational linguistics, natural language processing and in industry the use of human (often non-expert) annotators for the construction of labeled datasets has become an established practice. As described in Snow et al. (2008), the Amazon Mechanical Turk 'is an online labor market where workers are paid small amounts of money to complete small tasks', for example annotating articles using a sets of prefixed topics[5]. However, it is also important to remark that the evaluation of a NLP tool for a specific task strongly depends on the annotation quality of the data. For this reason, especially when dealing with complex annotation tasks, it is crucial to rely on more than one annotator and assess the reliability of the annotations using a variety of measures aimed at quantifying the agreement rate among them (Artstein and Poesio, 2008). In our work, we adopt the Cohen's kappa coefficient (Cohen, 1968).

As humanists are working on extremely specific in-domain research tasks, they cannot rely on Amazon Mechanical Turk annotations as other researchers usually do. For solving this specific issue, they cannot even rely on computer scientists or data mining experts: they need the help of their peers for creat-

[4] The examples presented here describe an over simplified case study. However, the complexity of the evaluation process can easily be shown by turning to more complex, realistic tasks like, for example to identify how the different meanings of 'will' evolve within a reasonably sized historical corpus.

[5] To know more about it: `https://www.mturk.com/mturk/welcome`

ing gold standards of annotated data in order to test the employed methods.

10.2.2 Evaluation Metrics

After having manually annotated a sub-set of the collection with the correct answer given a question (e.g. this article is - or is not - focused on the political relations between USA and Cuba), different metrics could be adopted to assess the performance of a system. We report some of the most adopted metrics for classification and information retrieval in the next paragraphs, which have been adopted to evaluate the performance of LDA topic models in the next chapter.

Accuracy. The proportion of true results (both true positives and true negatives) among the total number of cases examined.

Precision. The fraction of retrieved documents that are relevant to the query (or to the class).

Recall. The fraction of documents that are relevant to the query (or to the class), which are successfully retrieved.

F1 Score. The harmonic mean of precision and recall.

The following measure can be adopted for studying the quality of an information retrieval system:

Mean Average Precision. MAP provides a single measure of quality across the retrieved results for different queries. For a single query, the Average Precision is the average of the precision value obtained for the set of top k documents after each of the relevant document is retrieved.

10.2.3 Re-Training

The measures described above, together with a gold standard of human annotations, could be directly adopted for evaluating the results produced by an off-the-shelf NLP tool. However, if the intention is to re-train a classifier with new data annotated for a specific task, other steps have to be considered.

Training, Testing, Validating. The new learning process is typically divided into two main phases, namely: i) a training phase, in which the new predictive model is learnt from the new labeled data; ii) a testing phase, in which the previously learnt model is applied to unseen data in order to quantity its predictive power, specifically its ability to generalise to data other than the labeled ones seen during training. Additionally, a validation phase can take place to fine-tune the model's parameters for the specific task or domain at hand. In order to perform these steps, the initial annotated data have to be divided in three parts: training, test and validation sets (following proportions such as 80%, 10%, 10%).

Cross Validation. In order to avoid overfitting, namely when a model begins to "memorise" training data rather than "learning" to generalise from trend, a common practice is to divide the annotated data in k folds and use them alternatively as training, test and validation sets.

In the next section, we offer an overview of established practices in natural language processing, developed precisely for improving and evaluating topic models results. Having these measures and practices in mind, in the next chapter we will measure the quality of LDA results for identifying interdisciplinary theses.

10.3 Extending and Evaluating LDA

In the last few years, LDA has been extensively applied in digital humanities (Meeks and Weingart, 2013), given its capacity of detecting the presence of meaningful topics in collections of texts (Underwood, 2012; Jockers, 2014).

However, the use of LDA in DH also highlights many of the flaws related to the uncritical use of computational methods in the discipline. Scholars, with little consideration on the statistical assumption on which LDA is based, seem to be attracted to it because it "yields intuitive results, generating what really feels like topics as we know them, with virtually no effort on the human side" (Weingart, 2011). Being an unsupervised machine learning technique, it requires little prior work from the humanist. Although LDA can help to categorise big amount of data, it can also generate ambiguous topics which makes it hard to use the produced results as evidence in humanities research (Rhody, 2012), often calling for a lot of additional work regarding their evaluation (Chang et al., 2009; Wallach et al., 2009) and interpretation (Newman et al., 2010; Lau et al., 2011; Hulpus et al., 2013; Nanni and Ruiz Fabo, 2016).

As a direct consequence of this fact, digital humanities scholars are often stuck in a situation where they adopt topic models because they have a strong need for the potential benefits offered by such a method, especially now that large collections of primary sources are available for the first time in digital format. However, at the same time, scholars cannot derive new humanities knowledge from adopting topic models, given the current limitations of the results obtained (Schmidt, 2012; Nanni et al., 2016b).

Aware of these limitations, the NLP community has intensively worked on extending and improving topic models, by making them able to integrate or predict additional pieces of information related to the document (such as metadata information). Additionally, several studies have focused on developing approaches for improving the interpretation of LDA outputs and systems for evaluating topic model results. In the following paragraphs we cover some of them most relevant solutions.

10.3.1 Extensions of LDA

One of the first extensions of LDA is the *author-topic model* (Rosen-Zvi et al., 2004). This approach includes a specific type of metadata, i.e. authorship information, by representing each author as a multinomial distribution over topics, while each topic is associated with a multinomial distribution over words. Given a collection of documents with multiple authors, each document is modelled as a distribution over topics that is a mixture of the distributions associated with the authors.

This approach was further extended to the *author-recipient-topic model*, for its application in social networks (McCallum et al., 2005). The model not only considers individual authors, but conditions jointly on the authors of messages and on their respective recipients. Consequently, a topic distribution is assigned to each author-recipient pair.

By considering as external information the citation graph of a collection of scientific publications, the *Citation Influence Model* (Dietz et al., 2007) is another extension of LDA that estimates the weight of edges, i.e. the strength of influence one publication has on another. The *topics over time* approach (Wang and McCallum, 2006) incorporates temporal information and aims to model topic structure by identifying how this structure changes over time. Newman et al. (2006) explored the relationship between *topics and entities* (persons, locations, organisations) and introduced methods for making predictions about entities and for modeling entity-entity relationships.

Another solution is called *Labeled LDA* (Ramage et al., 2009), a supervised extension of LDA, originally used for credit attribution (namely, connecting each word in a document with the most appropriate pre-defined meta-tags and viceversa). We saw above how different types of metadata have been used in various extensions of LDA. The type of metadata exploited by Labeled LDA is *keywords* associated to a document: by constraining Latent Dirichlet Allocation to define a one-to-one correspondence between LDA's latent topics and those keywords, the goal of Labeled LDA is to directly learn word-label correspondences (taking the keywords as labels). Labeled LDA has already shown its potential for fine grained topic modeling in computational social science (Zirn and Stuckenschmidt, 2014). Unfortunately, the method requires a corpus where documents are annotated with tags describing their content and this meta-information is not always readily available.

10.3.2 Improving the Interpretation of Topics

Even if in the last decade the research community has strongly focused on extending LDA, exploiting various kinds of external knowledge, it has also been remarked (Chang et al., 2009) that LDA results remain very difficult to interpret for humans. Chang et al. (2009) adopted the phrase "Reading tea leaves": no better image could be found to describe how complex it can be to interpret topic model results. Given these difficulties, several researchers in natural language processing have focused on facilitating the interpretability of LDA results using different means, such as topic labeling. One of the first papers that presented the task of topic labeling (Mei et al., 2007) addresses the issue as an optimisation problem involving a minimisation of the Kullback-Leibler divergence between word distributions, while maximising the mutual information between a label and a topic model. A later approach (Lau et al., 2011) proposed adopting external knowledge to label topics. More recently, (Hulpus et al., 2013) made use of structured data from DBpedia: the authors hypothesise that words co-occurring in text likely refer to concepts that belong closely together in the DBpedia graph. Using graph centrality measures, they show that they are able to identify the concepts that best represent the topics.

Inspired by Hulpus et al., we have recently presented a new method for improving the interpretability of topic model results, based on the combination of entity linking and Labeled LDA (Lauscher et al., 2016). Our approach identifies in an ontology a series of descriptive labels for each document in a corpus. Then it generates a specific topic for each label. Having a direct relation between topics and labels makes interpretation easier; using an ontology as background knowledge limits label ambiguity. As our topics are described with a limited number of clear-cut labels, they promote interpretability and support the quantitative evaluation of the obtained results. We have illustrated the potential of the approach by applying it to three datasets, namely the transcription of speeches from the European Parliament fifth mandate, the Enron Corpus and the Hillary Clinton Email Dataset.

10.3.3 Evaluating LDA Results

Topic models are not simply difficult to interpret, they are also extremely complex to evaluate. Wallach et al. (2009) pointed out clearly how, even if several papers have worked on improving topic models, no single research contribution before 2009 has explicitly addressed the task of establishing measures to evaluate LDA results. In order to fill this gap, they introduced two new ways of estimating the probability of held-out documents, while Mimno et al. presented a way of evaluating the coherence of the topics obtained (Mimno et al., 2011). In 2009, another highly relevant paper on the evaluation of topic models was published: this article, by Chang et al. (2009), presented the word-intrusion and the topic-intrusion tasks as evaluation practices. In these tasks, humans have to detect a word or a topic which does not fit with the rest.

10.4 When do LDA Results Become Evidence?

Topic models can be evaluated per-se, or their results can be measured against a gold standard. A way of doing it is to study the alignment between topic-results and classes in the dataset, or to use topic model outputs as features in a classification task and compare it with other solutions.

In the next chapter, we present an extended evaluation in order to assess the quality of topic modeling outputs for detecting the main and secondary disciplines of a thesis and for retrieving interdisciplinary dissertations. Only when the reliability of a method (such as LDA) is known, its error modes are reported and its performances are satisfactory, it will become possible to use this approach for generating quantitative evidence.

Chapter 11

Capturing Interdisciplinarity in Academic Abstracts

A gold sun in the sky
gleams.

In this chapter I compare the performance of LDA topic models with other text mining methods for the identification of interdisciplinary doctoral dissertations from text. This work is based on a paper I submitted to the 5th edition of the International Workshop on Mining Scientific Publication, hosted at the 2016 Joint Conference of Digital Libraries, and published in a special issue of the magazine D-Lib (Nanni et al., 2016a). I wrote this article under the supervision of Laura Dietz, Stefano Faralli, Goran Glavas and Prof. Simone Paolo Ponzetto during the second part of my visiting period as the Data and Web Science Group of the University of Mannheim.

11.1 Introduction

Even if interdisciplinarity is a recurrent topic in academia, defining it as a quantifiable property remains extremely challenging even today. In fact, as remarked by others (Wagner et al., 2011), this concept depends on the existence of a clear distinction between academic disciplines, which is still a disputed issue (Sugimoto and Weingart, 2015). In Repko (2008) a precise overview of the debate and a series of specific definitions are presented. In particular, Repko defines interdisciplinary research in the following way:

> A process of answering a question, solving a problem, or addressing a topic that is too broad or complex to be dealt with adequately by a single discipline, and draws on the disciplines with the goal of integrating their insights to construct a more comprehensive understanding. (Repko, 2008)

While, there is a large body of work in scientometrics focusing on new solutions for identifying, quantifying, and visualising the diffusion of interdisciplinary research in large collections of scientific works (Rafols and Meyer, 2010), most of these methods exploit bibliometric measures such as co-author networks and citation graphs. The downside of such techniques is that they obviously depend on the availability of the bibliometric network data, which can often be difficult to obtain. In fact, it is generally much easier to obtain the full-text of the publications (as from the digital library of the University of Bologna) than the bibliometrics network data, especially for comparisons across different research areas.

Previous attempts towards content-based interdisciplinarity detection mostly relied on unsupervised machine learning methods. As one prominent unsupervised technique, topic modeling has been often used to detect mixtures of discipline-specific terminologies (Ramage et al., 2011; Nichols, 2014), often in combination with partially supervised methods (Chuang et al., 2012b; Giannakopoulos et al., 2014). Noticing a general lack of gold standards and quantitative evaluations for the task of interdisciplinarity detection, the goal of this chapter is to fill this gap and provide answers to the following research questions:

- How well can the main discipline of scientific publications be predicted using supervised models with content-based features?

- In the case of interdisciplinary research, can similar techniques be used to robustly detect the secondary discipline(s)?
- Do prediction confidences of the main-discipline classifier contribute to the detection of interdisciplinary research (i.e., do they improve the performance of the second-step binary interdisciplinarity classifier)?

In this chapter, we provide an extensive comparison between the results produced by LDA topic models and by other standard approaches from the field of NLP for performing the task. In addition to this, we offer the examination of different text-based features, both lexical (TF-IDF-weighted term-vectors) and semantic (topic model distributions, word embeddings). Finally, we present a reliable method that, given the abstract of a scientific publication, is able to accurately predict its main discipline, a set of secondary disciplines, and to detect if it is interdisciplinary or not. Through this analysis, this chapter aims to be a contribution to the recent debate on the importance of tool criticism in digital humanities research and a step towards a *computational history* of interdisciplinarity in academic institutions.

11.2 Related Work

In the field of scientometrics (Van Raan, 1997), the use of bibliometrics approaches and network analysis techniques is the most common methodology for evaluating and quantitatively describing the scientific output (Lu and Wolfram, 2012). In particular, for measuring interdisciplinarity, researchers usually opt for numerous variants of citation and co-citation analyses (Wagner et al., 2011). As already remarked (Rafols and Meyer, 2010), the percentage of citations outside the main discipline of the citing paper is the most useful indicator of interdisciplinarity.

More recently, researchers started exploiting textual content of publications in scientometric studies and several methods using lexical and topic-based information were proposed. Dietz et al. (Dietz et al., 2007) extend Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to quantify the impact that research papers have on each others, whereas Gerrish et al. (Gerrish and Blei, 2010b) extended LDA to model authoritative publications. Furthermore, Lu et al. (Lu and Wolfram, 2012) compared LDA to co-citation methods for measuring the relatedness between authors and their research and remarked that the topic modelling approach produces a more complete map of these relations.

For what concerns in particular the automatic detection of interdisciplinarity, several studies observed the usefulness of topic modelling for the task (Ramage et al., 2011; Chuang et al., 2012b; Nichols, 2014). One of the few supervised approaches towards interdisciplinarity detection from text is proposed by Giannakopoulos et al. (Giannakopoulos et al., 2014). In this paper the researchers adopt a Naive Bayes classifier to establish the probability of each publication to belong to a specific discipline. The obtained results are used to visualise “distances” between academic fields, and are interpreted by the authors as signals of interdisciplinary practices.

In Chuang et al. (Chuang et al., 2012b), the authors employ the Rocchio Classifier (Joachims, 1996) and, as in Giannakopoulos et al. (Giannakopoulos et al., 2014), derive distances between dissertation abstracts and representations of disciplines with the goal of identifying and visualising interdisciplinary research in the Stanford corpus of doctoral dissertations.[1] These distances are obtained by two different vector-based representations of dissertations abstracts: (1) term-vectors, typically weighted according to the term frequency-inverse document frequency (TF-IDF) scheme (Salton and McGill, 1983) and (2) vectors capturing the distributions of latent topics in the abstract, obtained using the topic models. It has been observed that different representations of abstracts yield outputs that “seem plausible, but each makes different predictions” (Chuang et al., 2012b).

11.3 Methodology

Following Chuang et al. (2012b) and Nichols (2014), we studied a two-step pipeline where first the abstract is classified into one of many research disci-

[1] As presented here: `http://nlp.stanford.edu/projects/dissertations/`

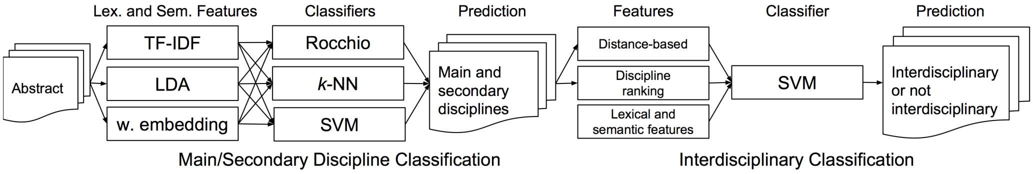

Lex. and Sem. Features
Classifiers
Prediction
Features
Classifier
Prediction
Abstract
TF-IDF
LDA
w. embedding
Rocchio
k-NN
SVM
Main and secondary disciplines
Distance-based
Discipline ranking
Lexical and semantic features
SVM
Interdisciplinary or not interdisciplinary
Main/Secondary Discipline Classification
Interdisciplinary Classification

Figure 11.1: Graphic representation of the adopted pipeline for interdisciplinarity detection.

plines, then the outputs are analysed to predict a binary interdisciplinarity attribute. The pipeline of the proposed approach is depicted in Figure 11.1.

11.3.1 Main/Secondary Discipline Classification

Our first task was to classify dissertation abstracts according to their main (primary) discipline. To this end, we experimented with several different classifiers for which we used the following set of lexical and semantic features:

TF-IDF. The TF-IDF-weighted term vector of the dissertation abstract, computed over the corpus of available dissertation abstracts.

LDA. The topical distribution (i.e., a topic-probability vector) of the abstract. The topic model was trained over the corpus of available dissertation abstracts. Chuang et al. (2012) already investigated these feature representations for detecting interdisciplinarity.

Embeddings. The semantic embedding of the dissertation abstract, which is computed as the element-wise average of the embeddings of the words the abstract contains. Let A be the set of unique words in the dissertation abstract. The embedding of the abstract (v_a) is then computed as follows:

$$v_a = \frac{1}{N} \sum_{w \in A} \text{freq}(w) \vec{v}_w$$

where freq(w) is the frequency with which word w occurs in the abstract, v_w is the embedding vector of the word w, and N is the total number of words in the dissertation abstract. We used the state-of-the-art GloVe word embeddings (Pennington et al., 2014) for the computation of the abstract embeddings. We experimented both with the pre-trained GloVe embeddings[2] and word embeddings trained by us on the CORE Corpus of scientific publications.[3]

Using a k-fold cross-validation setup, we experimented with three different supervised classification models using the above described set of features: (1) Rocchio classifier (Joachims, 1996), (2) k-Nearest Neighbors (k-NN) (Cover and Hart, 1967), and (3) Support Vector Machines (SVM) (Joachims, 1998).

Rocchio. In Rocchio classification, a centroid is computed from supervised discipline labels. This represents the center of mass of all the members. Given an unobserved abstract, this is classified by comparing it with each discipline centroid: the closest one is returned as the most suitable label. This method has been already adopted for the task in Chuang et al. (2012) and by us in the previous chapters.

k-NN. As the Rocchio classifier generalises each class to a single centroid, this approach could be sub-optimal when the classes are broad and general. For this reason, the second method we studied is the k-Nearest Neighbors classifier (k-NN). This is an alternative classification method that labels each observation with the majority class of the k most similar training documents.

SVM. The last method under study was a Support Vector Machine (SVM). SVMs are one the most adopted approaches for text classification tasks (Joachims, 1998). In this model, examples are represented as points in a multidimensional feature-space. Learning the classification consists in finding hyper-planes that separate the points while maximising the margin between the hyper-planes and the closest points. In this work, we trained a

[2] Trained on the merge of Wikipedia and Gigaword corpus: `http://nlp.stanford.edu/projects/glove`

[3] `https://core.ac.uk/intro/data_dumps`

series of binary classifiers which distinguish between one of the labels and the rest (one-versus-all) and finally assign each thesis to the classifier with the highest confidence.

In a multi-class classification setting, all three classifiers offered a mechanism for ranking the discipline classes for every given dissertation abstract instance: SVM and Rocchio directly provide confidence scores for each class, and for k-NN we rank the classes according to the number of neighbours (out of the total k) labeled with a particular class. We then used these class rankings, implicitly produced by the classifiers when predicting the main discipline of an abstract, as a method for capturing the secondary disciplines.

11.3.2 Interdisciplinarity Classification

We present here the features we adopted for the binary interdisciplinarity detection task.

Original. The feature vectors used for multi-class discipline classification can also be used in this secondary classifier (trained with different truth labels). These are TF-IDF term vectors, topic distributions and semantic abstract embeddings. These features capture general language patterns of abstracts, without any further discipline information.

Distance-based features. These features directly encode the confidences of the main-discipline classifier from the first step. Given a set of n disciplines, this will be a numerical vector of length n where an element at position m of the vector is the classifier's confidence that the m-th discipline is the main discipline of the abstract. The rationale behind this feature is that a more uniform confidence distribution across disciplines (when classifying the main discipline) is a stronger indicator of an interdisciplinary dissertation, whereas a very skewed confidence distributions should be more common for the monodisciplinary dissertations. This feature represents the assumption behind previous attempts to interdisciplinary detection described in literature (Chuang et al., 2012b; Nichols, 2014).

Discipline-ranking features. In some cases, the ranking of disciplines induced by confidence scores (i.e., primary discipline versus secondary discipline) might be more informative than the confidence values. Furthermore, some main disciplines might be more appropriate for interdisciplinary research (e.g., biology) than others (e.g., history). This is encoded in a numerical feature vector of length $n \times n$, encoding the matrix of 28 disciplines across 28 possible ranks, where within the row of rank r, we place a 1 in the column corresponding to discipline m on that rank. We motivate the usage of such a feature vector with the following observations:

- If the confidence scores are evenly distributed between distant disciplines (e.g., *computer science* and *biology*) this should be an indicator of interdisciplinarity of the dissertation;

- If the confidence scores are evenly distributed between closely-related disciplines (e.g., *history* and *political science*) this might indicate monodisciplinarity rather than interdisciplinarity. The balanced confidence distribution between the similar disciplines in that case is merely a result of main-discipline classifier's inability to make a confident decision because closely related disciplines share a lot of terminology (i.e., lexical features).

SVM. We trained a binary SVM classifier with the above-described features using ground truth annotations in a k-fold cross-validation setting.

11.4 Quantitative Evaluation

11.4.1 Dataset

Recently, Italian universities started to offer online institutional repositories of the doctoral theses defended at their institutions. Each publication is stored under legal deposit at the National Libraries of Florence and Rome

Table 11.1: The total number of abstracts for each discipline in our dataset (All) and the number of interdisciplinary (Int-Disc) and monodisciplinary (Mono-Disc) theses in our gold standard.

Discipline	All	Int-Disc	Mono-Disc
Agriculture	233	14	22
Anthropology	13	1	0
Arts	55	4	5
Biology	303	4	16
Chemistry	262	8	15
Civil Engineering and Architecture	117	7	11
Classical Languages	29	0	3
Computer Engineering	203	4	6
Computer Science	58	2	7
Earth Science	110	5	6
Economics	122	1	7
Geography	11	0	0
History	69	3	3
Industrial Engineering	216	8	12
Law	179	6	6
Linguistics	70	6	3
Mathematics	36	3	1
Medicine	322	3	17
Oriental Studies	7	0	0
Pedagogy	19	0	1
Philology and Literary Studies	54	1	0
Philosophy	25	2	6
Physics	172	4	18
Political and Social Sciences	68	0	3
Psychology	73	1	6
Sport Science	9	0	0
Statistics	30	2	0
Veterinary	89	4	5

and is uniquely identified by the National Bibliography Number (NBN) and the Digital Object Identifier (DOI).
One of the largest datasets available is offered by the Digital Library of the University of Bologna. When this research was conducted, it consisted of 4556 theses, defended between 2007 and 2015. Each of these theses is described with a series of metadata: title, author, short abstract (the majority of which are in English), names of the supervisors, etc. However, the bibliography is not available, therefore standard scientometric analyses for identifying interdisciplinarity cannot be conducted.
From this dataset, we selected all the theses with an abstract in English (2954) as our starting dataset, of which a subset yields our gold standard (see Section 3.2). The mean length of these abstracts is of 320 tokens.

Main Discipline Label. While it could be complex to select a primary discipline-identification (Wagner et al., 2011), our dataset contains an explicit mention of the main discipline of each dissertation in the field "Settore disciplinare" (Subject Area). The obtained labeled corpus contains 28 primary disciplines (statistics provided in Table 1).

Secondary Discipline Labels and Interdisciplinary Annotations. With the primary discipline label, we had a method for evaluating the performance regarding the first research question (predicting the main discipline). However, for the other research questions we needed to manually create a gold standard with annotations of other secondarily related fields, and a boolean flag of whether a thesis is interdisciplinary or not. The highly technical nature of the abstracts' content required expert annotators in different academic fields. In a survey, we asked supervisors of the dissertations to provide assessments for each of their supervised dissertations whether they consider it to be interdisciplinary and if so, to list the secondary disciplines of the dissertation. This resulted in a collection of 272 dissertations which are manually annotated for interdisciplinarity by experts (93 interdisciplinary and 179 monodisciplinary dissertations).

Table 11.2: Results on discipline classification (main: F1, secondary: MAP). Methods over which SVM TF-IDF achieves significant improvements (std error test) are marked with $^-$.

	MDC			SDR
	All	Mono-Disc	Int-Disc	Int-Disc
Random	0.04$^-$	0.04 $^-$	0.04 $^-$	0.10$^-$
Rocchio TF-IDF	0.71$^-$	0.84	0.62	**0.55**
Rocchio LDA	0.62$^-$	0.71$^-$	0.53$^-$	0.43$^-$
k-NN TF-IDF	0.67$^-$	0.69$^-$	0.57$^-$	0.27$^-$
SVM TF-IDF	**0.75**	**0.87**	**0.68**	**0.55**
SVM w. Emb.	0.68$^-$	0.8$^-$	0.62	0.52
SVM All feats.	**0.75**	**0.87**	**0.68**	**0.55**

In order to estimate the difficulty of detecting interdisciplinarity for humans, we asked two other expert annotators to re-annotate the 272 dissertations.[4] With respect to interdisciplinarity annotations, the annotators agreed with the dissertation supervisors in 82% of the cases, with the inter-annotator agreement in terms of Cohen's Kappa score being 0.63 (which is considered to be substantial agreement). In cases where an annotator disagreed with a supervisor, we kept the annotations assigned by the dissertation supervisor.

11.4.2 Experiments

We evaluated the three parts of our pipeline, main disciplinary classification, secondary discipline classification and interdisciplinarity detection, with model hyper-parameters optimised on a hold-out tuning set,[5] and studied the performance using 10-fold cross-validation.

Main discipline classification. In Table 11.2 we report (1) the classifi-

[4] According to their expertise, one for natural sciences, a second annotator for technical and social sciences.

[5] For SVM, best hyper-parameter C=1 in all experiments. For k-NN, optimal k varied in the range 30-250. When using LDA, we tested values for the parameter k (the number of topics) in range 50-1000. Best performance with 250 topics. When using the SVM, unless indicated otherwise, we use a linear SVM (i.e., no non-linear kernel).

cation results of the best performing methods for the main-discipline classification task (MDC), in terms of micro-averaged F_1 score and (2) the ranking performance for the assignment of secondary disciplines (SDR), measured in terms of mean-average precision (MAP) against the manually assigned secondary disciplines. In order to validate the underlying assumption that predicting the main discipline for interdisciplinary dissertations is more difficult than for monodisciplinary dissertations, we separately measured the performance of the models on the subset of monodisciplinary dissertations ("Mono-Disc") from the performance on the interdisciplinary dissertations ("Int-Disc"). As expected, the performance on the monodisciplinary subset is significantly higher than the performance on the interdisciplinary subset, which confirms our assumption that it is more difficult to detect the main discipline for interdisciplinary publications.

An important initial finding is that using lexical features (i.e., TF-IDF weighted term-vectors) within a SVM or a Rocchio classifier consistently outperformed other models which use semantic features, and in particular the results of LDA topic models.

Secondary discipline classification. The quality of the ranking of the secondary disciplines was evaluated only on the subset of the corpus consisting of 93 interdisciplinary dissertations ("Int-Disc"). The trends in the results are consistent to experiments on main-discipline classification: SVM and Rocchio classifier based only on lexical features outperform other combinations of classifiers and features (and in particular of LDA topic models).

Interdisciplinarity detection. The performance of different models on the interdisciplinarity detection task, in terms of micro-averaged F_1 score, is shown in Table 11.3. In all cases, we used a binary SVM for the classification, but varied different features sets in accordance with methods discussed for discipline classification (denoted in the row headers of Table 11.3) as follows. The first column (*orig*) presents results of the SVM model for interdisciplinarity classification that uses the same set of features as the main-

Table 11.3: Performance on the interdisciplinarity classification task (F1). Underlined method/features is significantly better than all other methods.

	Interdisciplinary detection				
	orig	dt	dr	dt+dr	all
Random	0.4	0.4	0.4	0.4	0.4
Rocchio TF-IDF	0.52	0.51	0.54	**0.56**	0.57
Rocchio LDA	0.52	**0.56**	0.45	0.53	0.59
k-NN TF-IDF	0.52	0.46	0.52	0.42	0.58
SVM TF-IDF	0.52	0.44	0.49	0.47	0.56
SVM w. Emb.	0.60	0.35	0.37	0.36	0.60
SVM All feats.	**0.74**	**0.56**	**0.55**	0.55	**0.70**

discipline classifier given by the corresponding table row (therefore row 2,4,5 are identical w.r.t. the orig column). The remaining columns present performance achieved when using the output of the discipline classifier as input features: (1) only distance-based features (*dt*), (2) only discipline-ranking features (*dr*), (3) distance-based and discipline-ranking features (*dt* + *dr*), and (4) distance-based and discipline-ranking features together with the original features used by the main-discipline classifier (*all* = *orig* + *dt* + *dr*). Our results clearly show that inferring interdisciplinarity from the results of the main-discipline classifier (i.e., using distance-based and discipline-ranking features employing LDA topic models outputs), as suggested in literature Chuang et al. (2012b), cannot be done robustly. Using the prediction confidences of the main-discipline classifier for interdisciplinarity classification (the scores in the *dt*, *dr*, and *dt*+ *dr* columns of Table 3) yields worse performance than using directly the same lexical and semantic features (*orig*) that were used for the main-discipline classifier. The best interdisciplinarity classification performance is achieved by the SVM model directly using lexical and semantic features (TF-IDF term vector, topic distribution, and abstract embedding), which is not exploiting in any way the confidence scores produced by the main-discipline classifier.

These results disprove the assumption from previous works (Ramage et al., 2011; Chuang et al., 2012b) that a notion of "distance" between publications and representation of disciplines are the best way for detecting interdisciplinarity. Our experiments reveal that the distinction between interdisciplinary and monodisciplinary practices can be better explained by the differences in the language used than by the distances to disciplines' classification centroids or hyperplanes. This might indicate that, instead of the ranking of involved disciplines, what really distinguishes interdisciplinary dissertations is their focus and the language in which they present and describe their research.

11.5 Conclusion

In this chapter we presented an extensive evaluation of text mining methods (in particular LDA topic models) for the task of interdisciplinary detection. Following the insights from related studies Chuang et al. (2012b), we frame the pipeline as a sequence of two supervised classification tasks: (1) classification of the main discipline of the publications, and (2) binary interdisciplinarity detection, with features based on results of the main-discipline classification as (additional) input. In both classification steps, the study included several sets of lexicalised and semantic features such as term-vectors, topic distributions, and semantic word embeddings.

The results of our experiments on the corpus of doctoral dissertation abstracts show that both robust identification of the main discipline and ranking of the secondary disciplines can be better achieved using a simple (and computationally inexpensive) classification model (e.g., Rocchio classifier with only TF-IDF weighted term-vectors) instead of employing complex solutions (e.g. LDA topic models outputs). Regarding the interdisciplinarity classification, the results of our experiments do not indicate that discipline classification is needed for interdisciplinarity detection, which is in contrast to findings of previous work. On the contrary, we show that the best performance on the interdisciplinarity detection task is achieved using the set of lexical and semantic features derived directly from the text. We speculate that the lexicalised classifier learns from cue words such as "interdisciplinary", "innovative" and "collaboration" as well as discipline-specific words together.

In addition to the more "technical" outcomes, a second important finding of this study is that, across all experiments, word-features are consistently strongly outperforming topic-features from LDA. This means that the results of topic modeling for this task, in strong contrast to findings of previous work, could not be considered as the most reliable evidence. The relevance of this finding goes beyond this specific work and could be extended to many kinds of humanities researches that employ text mining methods for generating quantitative evidence. The reliability of an approach can never be assumed in advance, on the contrary it has to be tested and proven, and its error modes have to be clearly understood before adopting it and interpreting its results for a humanities research task (as remarked in Traub and van Ossenbruggen, 2015; Nanni et al., 2016).

Chapter 12

Next Step: Enhancing the Use of Academic Collections

In the suburbs I,
I learned to drive.

This short chapter wraps-up Part II & III of this thesis and show the potential of combining born-digital sources and computational methods for the study of the recent past of European academic institutions.

12.1 Introduction

One of the central arguments of this dissertation is that historians who intend to use born-digital materials in their research should always take in great consideration the scarcity and abundance issues that emerge when dealing with them. As it has been described through a series of case studies, these intrinsic problems of born-digital sources are about to change completely the way we traditionally identify, analyse and select evidence in our work. The *leitmotiv* of this dissertation has therefore been that historians should consider and - in a highly critical way - combine the historical method with methodologies and solutions from other fields of study, in order to successfully deal with these issues and continue to pursue the study of the past.

While the focus of Part II of this thesis has been dedicated to the scarcity issue, in Part III we started by considering the large collections of born-digital academic materials now at disposal of historians of higher education. These collections, which comprise academic dissertations, scientific publications, syllabi, grant proposals, etc., are fully accessible only by using natural language processing approaches.
Given this fact, in the previous chapters we have strongly emphasised the importance of adopting computational methods in a highly critical way, to avoid misinterpreting the final results. In order to remark further on this important aspect, we presented a case-study based on the adoption of one of the most employed text mining techniques in digital humanities research, namely LDA topic modeling, for the identification and retrieval of interdisciplinary theses in a corpus of academic dissertations. Through this study, we showed how the results produced by LDA were less reliable than the one produced by much simpler, although less popular, methods (e.g., Rocchio TF-IDF).

The approach developed and described in the previous chapter can be now employed to support historians of higher education in their work. In the next pages, we will show how this solution has been adopted to enrich a large dataset of dissertation abstracts published by European institutions in the last 30 years and how the obtained information could be used in historical research.

12.2 Enriching the DART-Europe Corpus

DART-Europe (Digital Access to Research Theses - Europe) is a partnership of research libraries and library consortia who are working together in order to improve global access to European research theses. The online portal offers over 700.000 theses from 28 European countries and 596 universities. While this corpus presents an unprecedented amount of primary sources for future historians of higher education, the available collection has a specific limitation. Only a very small part of the dataset has metadata regarding the discipline of the thesis. This limits both the navigation of the corpus and impedes any type of diachronic and cross-country discipline-based comparative study (such as studying the changes in biological research across the last thirty years, as well as examining the difference in topics studied in political science, depending on the political situation of the country).

12.2.1 Method

In order to address this issue, we employed the method presented in the previous chapter for detecting the main and secondary disciplines of a dissertation (i.e., a SVM using TF-IDF vector representations of each abstracts). This supervised classifier has been initially trained on a subset of the DART-Europe dataset, which provides information regarding the main discipline (all theses from Italian universities with an abstract in English, which are 11726), and evaluated with 10-fold cross validation.
We obtained a micro F1-Score of 0.72, which is consistent with the results obtained in the previous chapter, where the classifier was trained and tested only on theses from the University of Bologna. In addition to this, as the presented task can also be considered as a ranking problem (i.e., at which position of the ranking produced by the SVM does the correct discipline appear?), we report the obtained Recall@1 (0.72) and Recall@2 (0.91). This means that the correct discipline appears to be in the top two results produced by our classifier in more than 90% of the cases.

Using this classifier, we identified the main and secondary disciplines in a subcorpus of 200.000 doctoral theses published between 1985 and 2016, which provide an abstract in English. The obtained resource[1] could support many studies of the recent past of academic institutions. In the next section, we present a small case-study, which shows the potential of these sources and

[1] `https://federiconanni.com/computational-turn/`

methods[2].

12.3 Exploring its Potential: Quantifying the Computational Turn

During the last decades, academia seems to have experienced an unstoppable growth in the adoption of digital methods. Many argue that the impact of the use of computational resources, infrastructures and approaches has challenged the traditional way we conduct research in sciences, social sciences and humanities (Berry, 2011; Kitchin, 2014).
The "computational turn" that we can notice by looking at the programs of traditional conferences or at the topics of research grants, is fostered by technological as well as political and cultural reasons. As a matter of fact, the continuous growth in availability of digital datasets (from public genome data to open data provided by public administrations to collections of textual data such as HathiTrust) together with the advancement in computational power and in machine learning solutions are playing a central role in fostering the adoption of digital technologies all around academia, from biology to sociology to history, and a re-thinking of our methodologies. Additionally, the rhetoric on the infinite potential of big data and artificial intelligence (fostered by computer science companies such as Google and Facebook and constantly emphasised by the media) has an impact on academia as well, by conditioning research focus, methods and collaborations.
However, while the rhetoric on the "computational turn" that academia is experiencing is easy to spot, what remains difficult to assess is whether it is true that academia as a whole has been actually adopting more digital resources and methods during the recent years. In order to address this

[2] This work (currently under peer-review) has been conducted in collaboration with Giacomo Nanni (Academy of Fine Arts - Bologna) and Giulia Paci (European Molecular Biology Laboratory - Heidelberg). In particular, I designed the study, collected the corpus, developed the method and conducted the experiments; Giacomo Nanni has worked on the visualisation of the results and Giulia Paci has overviewed the data-analysis step of the work.

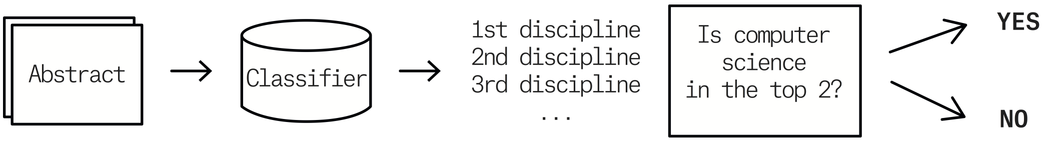


Figure 12.1: Schema of our pipeline.

question, we adopted the enriched version of the DART-Europe Corpus.

12.3.1 Approach

In order to determine whether a dissertation in our collection employs digital methodologies or not, we first identified its main disciplines. We did so by employing the prediction confidence of the first supervised classifier (main and secondary discipline classifier), which labels each thesis with a set of disciplines, based on the content of its abstract (in English).
If "Computer Science" (or "Computer Engineering") appeared to be one of the top two disciplines detected, we consider that thesis as having a "computational" aspect. From now on, we will call this thesis "computational".

12.3.2 Analyses

Before going through our results it is important to bear in mind that the dataset we employed offers only a partial view of the real output of European universities during the last thirty years. This is due to the fact that some universities do not appear in the DART dataset and for several (especially German) dissertations an abstract in English is not available. In addition to this, it is also important to consider that, even if the quality of our classifier is solid, it remains a machine learning approach with a margin of improvement. With this in mind, in the next pages we present our preliminary findings.

Is Academia experiencing a computational turn?

The first study we conducted aimed at assessing whether academia has been generally going through a computational turn. First of all, we detected that 9% of theses in our dataset have been labeled as "computational", meaning that they present a combination of Life Sciences, Social Sciences or Humanities and Computer Science topics. Moreover, as presented in Fig. 12.2, we noticed a constant growth of these theses over the decades, starting from the 2% to the 10% of the total.

However, if we look at the mean percentage of computational theses per area, we noticed that this differs greatly, with Life Sciences having a mean of 8.4% +/- 0.45% computational theses, Social Sciences 8.8%+/- 0.47% and Humanities 3.5% +/- 0.45%. More specifically, if we consider the single disciplines, we notice the prevalence of computational theses in Biology, Linguistics and Economics over History, Classical Languages and Anthropology.

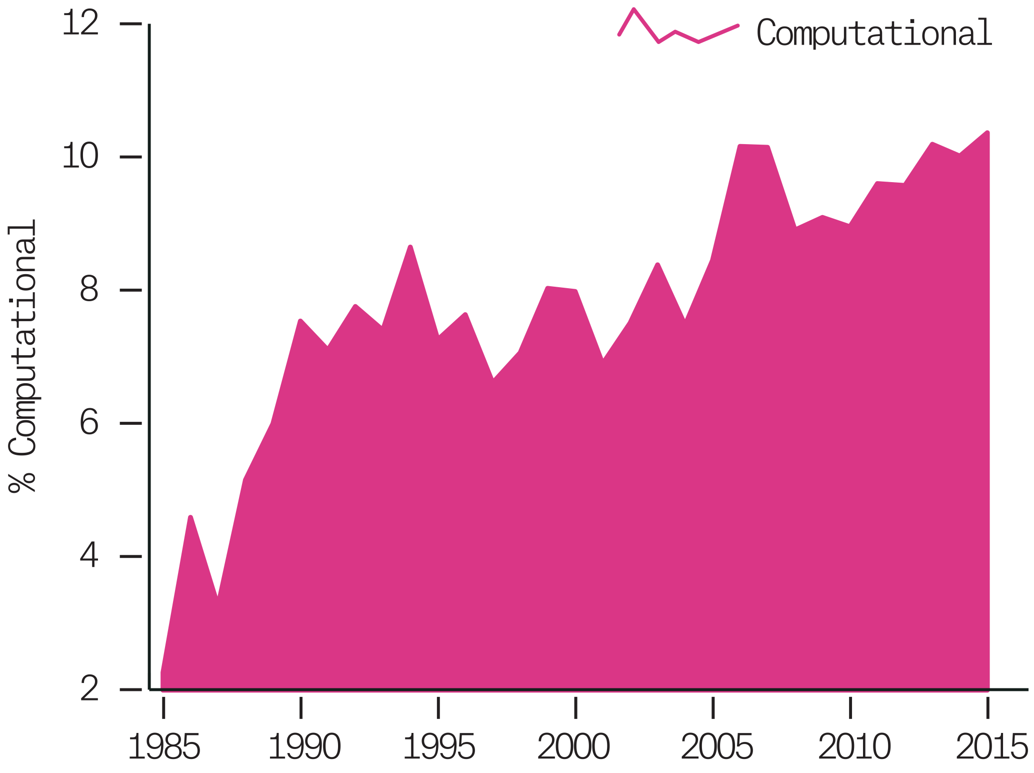


Figure 12.2: Growth of computational theses, between years 1985 and 2015.

What kind of universities are fostering it?

The dataset at our disposal offers different points of view for answering this question. As we wanted to recognise macro trends in the adoption of compu-

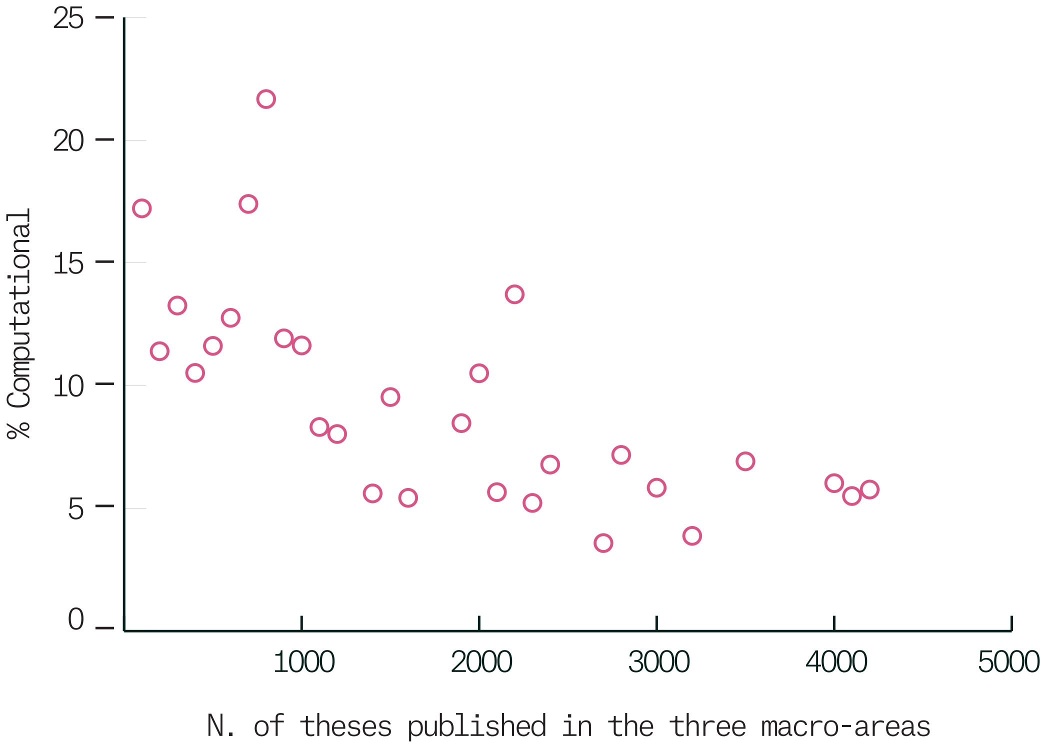


Figure 12.3: Average percentage of computational theses (y axis) vs number of theses published in the institution in the three considered macro-areas (x axis), binned in intervals of size=100.

tational methods in academia, we grouped disciplines considering the ERC domains. In particular, we focused here on disciplines belonging to Life Sciences, Social Sciences and Humanities. The results we report here are based on examining the relationship between the number of theses in Life Sciences, Social Sciences and Humanities published by a specific institution and the percentage of which are computational.

Two interesting aspects emerge from this preliminary study (see Fig. 12.3). First of all, we notice an inverse relationship between the number of theses published in Life Sciences, Social Sciences and Humanities by a specific institution and the percentage of which are computational. By examining the universities that publish more computational theses, we find that these are science and technology institutes such as the Telecom ParisTech, l'Ecole polytechnique fédérale de Lausanne (see also Fig. 12.4), TU Delft, l'Universitat Politècnica de Catalunya and the KTH Royal Institute of Technology in Stockholm. This seems to imply that, when research in Life Science, Humanities and Social Sciences is conducted at these institutions, it usually involves a computational aspect.

12.4 Towards a New Type of Comparative Study

In order to understand what kind of computational research is performed at these universities (is it really interdisciplinary research, or is it simply applying computational methods on a research task from a different field?) and which are the factors that foster these "interdisciplinary" projects, we envision further applications of this new type of comparative study between academic institutions (e.g., Fig. 12.4) using the extended DART-Europe dataset presented in this chapter.

This approach could support the community in understanding the role that private and public research funds have played in orienting academic research in this direction and how traditional institutions have been dealing with the advent of computational methods and the growth of their application in academia.

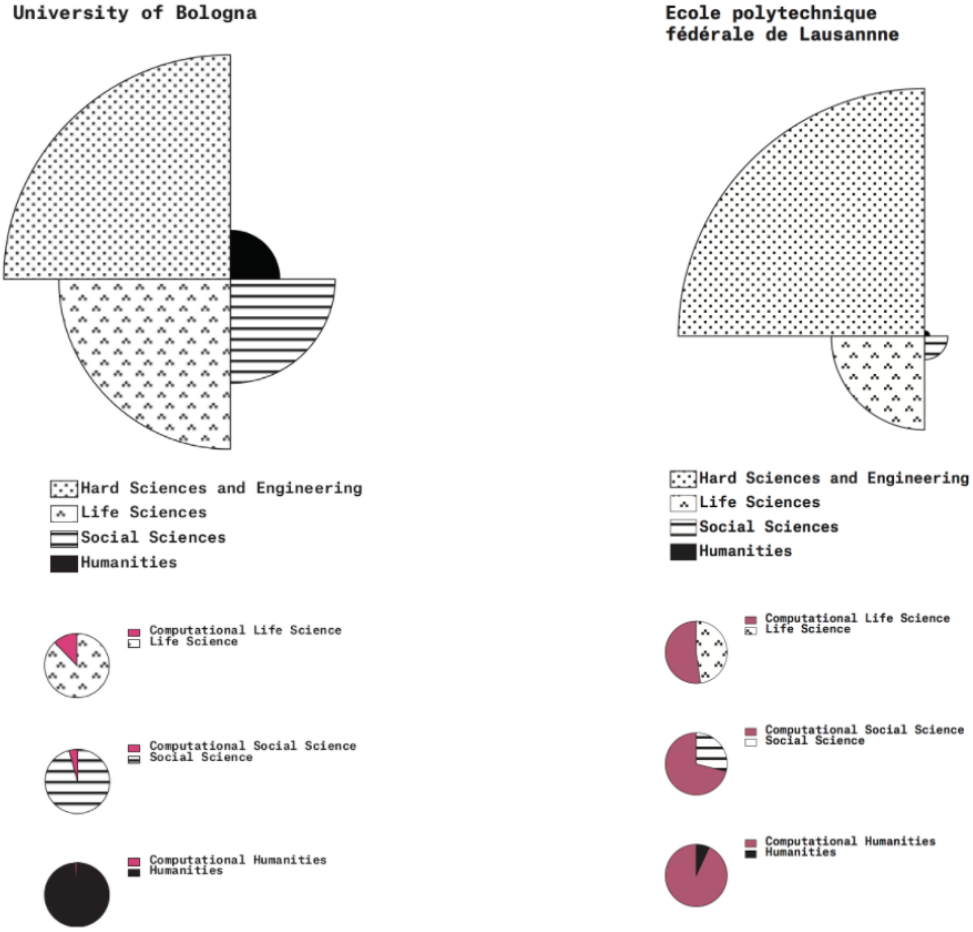


Figure 12.4: Distribution of theses per ERC domain (top graph) and corresponding distribution of computational theses for relevant domains. Comparison between University of Bologna and Polytechnique of Lausanne.

Conclusion: Envisioning the Future of the Craft

I guess you guys
aren't ready for that yet.
But your kids are gonna love it.

In recent years, researchers have argued that history, as other humanities disciplines, is reaching a turning point in its methodology (Nelson, 2016; Graham et al., 2016; Scheinfeldt, 2016). As a matter of fact, sustained by the efforts of many digitisation projects, the community has been employing more and more computational methods in order to examine these vast resources. This change in methodology has reopened a long-term debate on the ways textual evidence of the past can and shall be properly interpreted. This is noticeable when considering dedicated call for papers of history journals[3] as well as examining the activities of major associations (American-Historical-Association, 2015) and considering the large interest that concepts such as distant reading and text mining have attracted in the field[4].

While for the historical profession it is of course beneficial to re-discuss and criticise the validity of established practices of acquiring knowledge from sources, a central argument of this thesis is that the adoption of digitised

[3] See for example: http://www.officinadellastoria.info/magazine/images/stories/cfp_storiamedia/cfp_storiamedia.pdf; http://jah.oah.org/submit/digital-history-reviews/

[4] See for example: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/topic-modeling-the-past.1.html

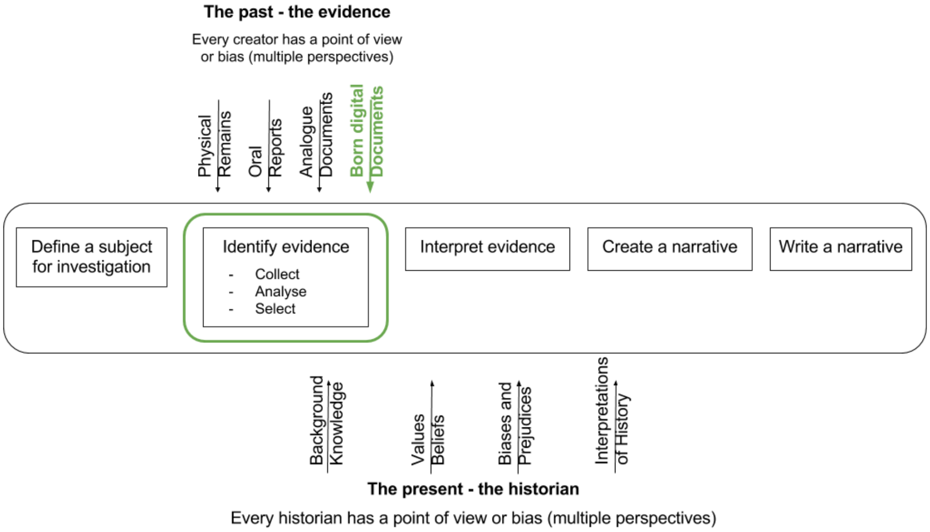


Figure 12.5: Graphical representation of the historical method.

datasets and computational methods cannot be considered by itself the triggering factor of a fundamental turning point in our profession. As a matter of fact, adopting or not adopting large-scale datasets of digitised sources, together with computational methods, will always remain a choice for the history scholar: Charles Darwin can still be studied without conducting text mining over the collections presented on Darwin Online[5], as well as the London of the XVIII Century can be examined without distant reading the Proceedings of the Old Bailey Online[6].

A Different Focus in Historiography

However, I do agree that history is about to face a drastic transition in its theories and methods, but I argued in Part I of this thesis that this will happen for a different reason. The triggering cause of this transition relies in the born-digital nature of the large majority of sources produced by contemporary societies. This change affects any type of documents we create and consume in our everyday life, from bureaucratic forms collected by the pub-

[5] http://darwin-online.org.uk/

[6] https://www.oldbaileyonline.org/

lic sector to newspapers articles, from mail correspondences to notes taken in school, and it is about to present all its consequences on historical research.

As I have discussed in the central parts of this dissertation, born-digital sources are way more complex to archive, collect, analyse and select compared to traditional materials. Websites (such as Unibo.it - see Part II), are large and variegated collections of documents, which are often not preserved in their entirety by web archive initiatives and that could be re-constructed only through the meticulous combination of various pieces of information from different sources. When a resource such as a website is finally re-created, this is usually so vast that computational technologies (i.e. natural language processing methods and information retrieval approaches) are necessary for identifying and retrieving specific documents (e.g. interdisciplinary dissertations - see Part III).

The methodological steps presented in this thesis for collecting, analysing and selecting born-digital documents require strong interdisciplinary competences and a highly critical attitude towards sources and methods. However, while the attention of historiography during the last decades has been almost solely dedicated to re-discussing the two central parts of the historical method, namely to what extent it is possible to discover the truth about the past through the interpretation of the sources and whether historians can convey this knowledge through an objective narration[7], the previous – essential – step of the craft has been kept in the shadow for too long.

For these reasons, I conclude this thesis first of all by remarking on the importance of restoring centrality in the historiographic debate to the ways sources are collected, analysed and selected, especially when these documents are born-digital materials.

[7] See for example Carr (1961); Evans (1997); Munslow (2006).

The Essential Lesson of the Cultural Turn

When I – as well as other researchers such as Graham et al. (2016), Nelson (2016) and Scheinfeldt (2016) – say that history is about to experience a deep change in its theories and methods, this implies that the long-term influence of postmodern ideas over the craft is reaching its end.

During this envisioned transition, it is essential that scholars will bear in mind the incommensurable importance that the cultural turn has had over historiography. The post-structuralist and deconstructionist theories proposed by Barthes (1967), Derrida (1967) and Lyotard (1984) and applied in works of scholars such as Foucault (1975) and Said (1979) have completely revolutionised the craft. Today, the fact that even the most scrupulous historical work is considered as only one of the different subjective perspectives on a topic (Munslow, 2006), is an essential component of the way scholars approach and debate the study of the past.

However, in the age of unlimited abundance of primary sources (most of them in the hands of private companies such as Google and Facebook (Mayer-Schönberger and Cukier, 2013) and government agencies (Greenwald, 2014; Berendt et al., 2015)), where all our online activities are monitored, all our digital conversations are preserved, all our physical activities are tracked, all our political, cultural, sexual preferences are constantly examined, the cliometric idea of discovering the profound laws that model human history could soon re-emerge in the historical community. This is noticeable already, if we consider the interest that private companies attract on themselves every time they present resources or tools for supporting the study of the past (Michel et al., 2011; Marshall and Shipman, 2014; Zimmer, 2015).

With this context emerging in front of us, I conclude by remarking that history scholars should continue pursuing the profound and radical post-modern critique against the existence of a unique "truth" about the past. This is in fact the essential starting point for critically approaching the vastness of born-digital data.

The Next Generation of Historians

In the next decades, studying the past will become more and more challenging. New sources are rapidly substituting traditional materials, however their materiality is extremely ephemeral and this makes them very difficult to preserve in their entirety. At the same time, when these resources are collected, they are in an incommensurable number, compared to traditional archive materials, and this forces researchers to employ computational methods to study them. In this complex scenario, I conclude this thesis by raising the more pressing question that resulted from my work: what is the best way of preparing the new generations of historians to the craft?

In recent years, the digital history community has already offered many educational activities on computational methods to history students.
From workshops to panels, from courses to summer schools, from tutorials to hackathons, these courses have almost always been focused on presenting the potential of new resources, tools and platforms to the history students, following an attitude which has been branded as “more hack, less yack” (Nowviskie, 2014).

While offering hands-on experiences with computational tools is important in order to introduce history students to the digital humanities, in this thesis I have often remarked on the fact that a critical approach is strongly needed in order to properly deal with born-digital sources and computational methods. For this reason, my initial answer to the question is that students should first of all be guided in shaping their research topics and receive the preparation necessary to support a critical analysis of the born-digital documents and computational methods at their disposal. This will be the imperative premise for a new generation of historians who will be able to go beyond the naïve adoption of sources and tools and will instead critically employ them to answer established research questions and to raise many new ones.

Acknowledgements

Clear eyes,
full hearts.

I started thinking about the topic of this thesis in 2012, during Prof. Francesca Tomasi's course on Computer Science for Archive Keeping, and I finished writing about it at Frankfurt airport, on March 4th 2017, while coming back from a visiting period at the University of New Hampshire focused on outlining the future steps of my research.
In the last five years, many researchers, colleagues and friends have helped me to shape and sharpen this study. To all of them, I am extremely grateful.

First of all, I am wholeheartedly thankful to my supervisors: Professor Maurizio Matteuzzi, who incessantly supported my desire of studying how to extend the historical method with a large set of interdisciplinary methodologies, and Professors Simone Paolo Ponzetto and Niels Brügger, who hosted me during my visiting periods at the Universities of Mannheim and Aarhus and opened the doors of their research groups to a traditionally trained historian eager to learn as much as he can.
In particular, I am deeply grateful to the enormous support and to the many academic opportunities Prof. Ponzetto has offered me while being a visiting scholar at the Data and Web Science (DWS) Group: I would not be so passionate about my research without these experiences.

I am - and will always be - absolutely grateful to the incredible mentoring and support that Professor Laura Dietz has offered me during the last two years. It is thanks to her if today I am not afraid to consider myself a data scientist. And thanks to Ben Gamari, who still believes that one day I will code in Haskell.

I am also very thankful to the many great colleagues I have worked with during the last three years, starting from the people at the International Centre for the History of Universities and Science. In particular, I am deeply grateful to the precise advice and strongly critical observations that Professors Giuliano Pancaldi and Anna Guagnini offered me and to the support my colleague and friend Anwesha Chakraborty has always given me.

I will always be indebted to Professor Bernardo Magnini, Dr. Anne-Lyse Minard, Dr. Manuela Speranza, Dr. Sara Tonelli and to the entire Human Language Technology and Digital Humanities Groups of the Foundation Bruno Kessler for constantly feeding up my curiosity on natural language processing while being a visiting researcher in their group.

Thanks to my colleagues of the startup FiND: it was a very formative experience and I am grateful to have shared it with you. Thanks to Nabil Arafin, Giulia Benotto, Vincenzo Caruso, Gianmarco Nicoletti, Jacopo Tagliabue and all the others at Awhy - I have learned a lot while working with you.

To all the professors and researchers at DWS: I know that it is not so common for a computer science group to have a historian rumbling around the corridors incessantly asking for advice, thank you for being so supportive. In particular thanks to Stefano Faralli, Andreas Fleig, Anna Lisa Gentile, Ioana Hulpus, Petar Ristoski, Michael Schuhmacher, Sanja Stajner, Lydia Weiland and Caecilia Zirn for your help and patience.
Thanks to Erman Acar, who knows why tea is better than coffee, and to Petar Petrovski, who is always there when I feel the need to discuss about something; you guys are the best office-mates I have ever had.
Thanks to Goran Glavas, for sharing with me his absolute passion for NLP

and for teaching me the only formula of linear algebra I currently know. And thanks to Anne Lauscher, the first computer science student I have ever supervised: you made my job way too easy.

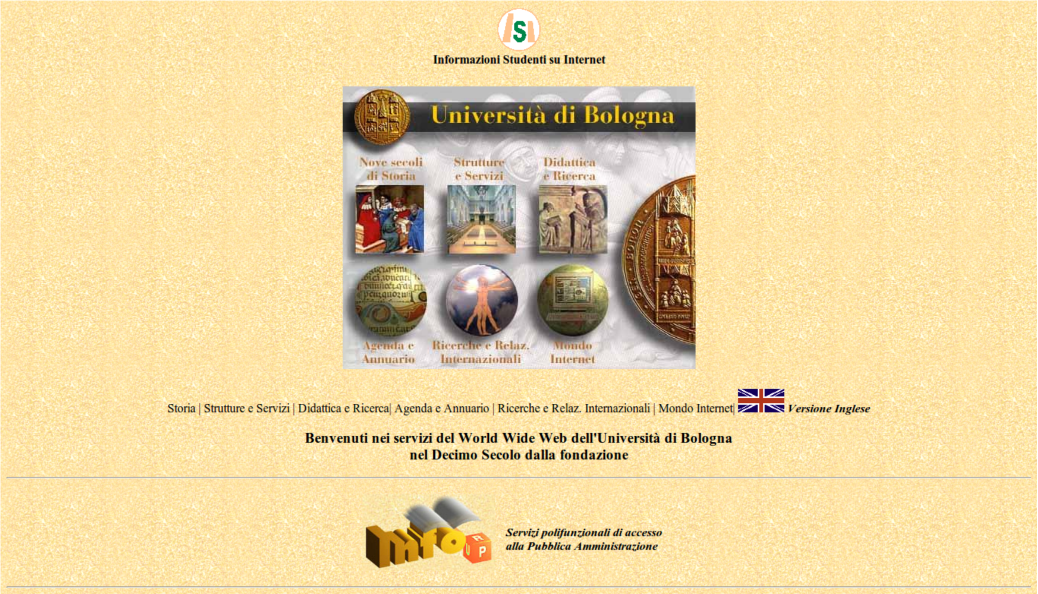
I am also thankful for the possibility to shape my research with the advice of great researchers, colleagues and friends such as Anat Ben David, Francesco Bianchini, Rudi Bonfiglioli, Marco Buechler, Dino Buzzetti, Charlotte Colding Smith, Josh Cowls, Meghan Dougherty, Greta and Emily Franzini, Hiram Kümper, Lorenzo Mantovani, Nikolay Marinov, Ian Milligan, Janne Nielsen, Pablo Ruiz Fabo, Alessandro Niccolò Tirapani, Francesca Tomasi, Jane Winters and many, many others.

Thanks to my family for welcoming me home even after having eaten Hawaii pizza (albeit only once) . Thanks to my brothers Niccolò and Giacomo - *mi mancate parecchio.* Thanks to my old friends from Bologna and to my new friends in Heidelberg and Mannheim, for having the patience of listening to me talking for half an hour (at the very least) on the future of historical research: I owe you a beer.

Finally, I am absolutely grateful to the infinite support, thoughtful advice and incredible trust in my capabilities that Giulia Paci has provided me since the day we first met. I have always believed that I am able to see two steps ahead, thank you for telling what awaits for me on the third.

Appendix

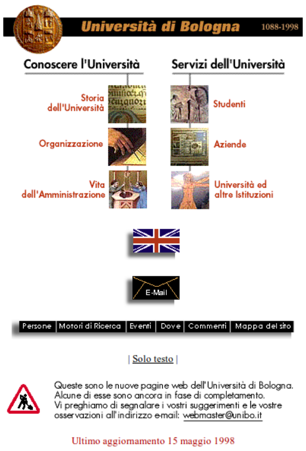
Snapshots of Unibo.it over time, provided by Mauro Amico (CeSIA).


Informazioni Studenti su Internet
Università di Bologna
Nove secoli di Storia
Strutture e Servizi
Didattica e Ricerca
Agenda e Annuario
Ricerche e Relaz. Internazionali
Mondo Internet
Storia | Strutture e Servizi | Didattica e Ricerca| Agenda e Annuario | Ricerche e Relaz. Internazionali | Mondo Internet|
Versione Inglese

Benvenuti nei servizi del World Wide Web dell'Università di Bologna nel Decimo Secolo dalla fondazione

Servizi polifunzionali di accesso alla Pubblica Amministrazione

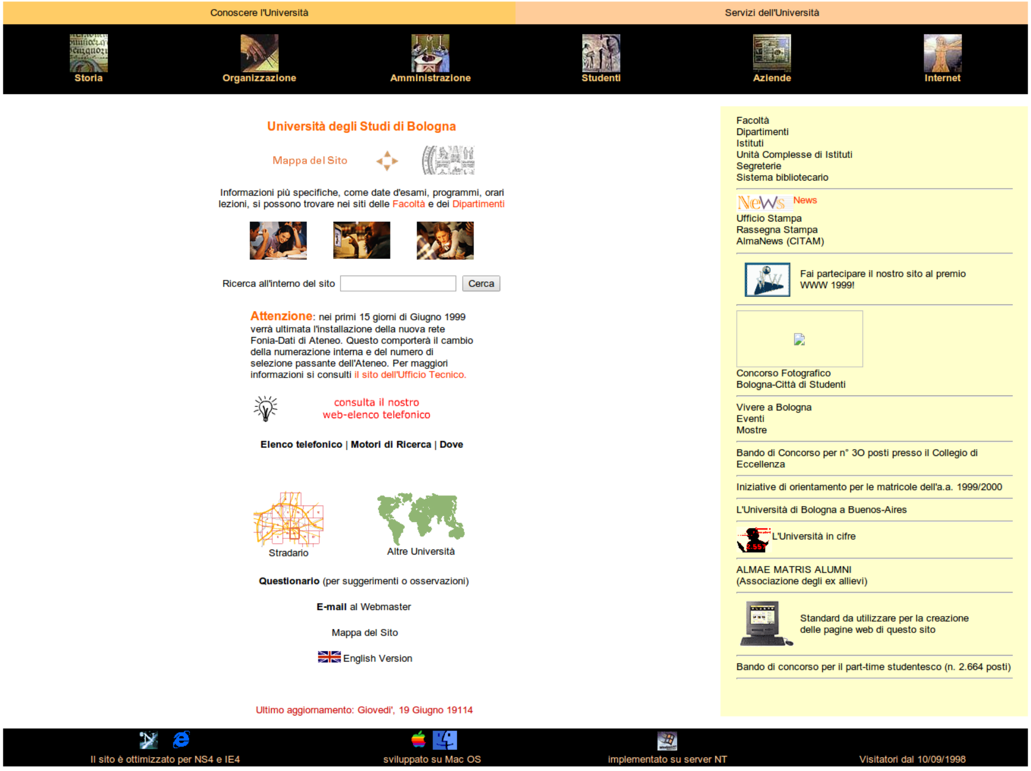
Figure A.1: Unibo.it homepage between 1996 and 1997.



Queste sono le nuove pagine web dell'Università di Bologna.
Alcune di esse sono ancora in fase di completamento.
Vi preghiamo di segnalare i vostri suggerimenti e le vostre osservazioni all'indirizzo e-mail: webmaster@unibo.it

Ultimo aggiornamento 15 maggio 1998

Figure A.2: Unibo.it homepage in 1998.



Informazioni più specifiche, come date d'esami, programmi, orari lezioni, si possono trovare nei siti delle Facoltà e dei Dipartimenti

Attenzione: nei primi 15 giorni di Giugno 1999 verrà ultimata l'installazione della nuova rete Fonia-Dati di Ateneo. Questo comporterà il cambio della numerazione interna e del numero di selezione passante dell'Ateneo. Per maggiori informazioni si consulti il sito dell'Ufficio Tecnico.

consulta il nostro
web-elenco telefonico

Questionario (per suggerimenti o osservazioni)

E-mail al Webmaster

Mappa del Sito

English Version

Ultimo aggiornamento: Giovedi', 19 Giugno 19114

Facoltà
Dipartimenti
Istituti
Unità Complesse di Istituti
Segreterie
Sistema bibliotecario

Ufficio Stampa
Rassegna Stampa
AlmaNews (CITAM)

Fai partecipare il nostro sito al premio WWW 1999!

Concorso Fotografico
Bologna-Città di Studenti

Vivere a Bologna
Eventi
Mostre

Bando di Concorso per n° 3O posti presso il Collegio di Eccellenza

Iniziative di orientamento per le matricole dell'a.a. 1999/2000

L'Università di Bologna a Buenos-Aires

L'Università in cifre

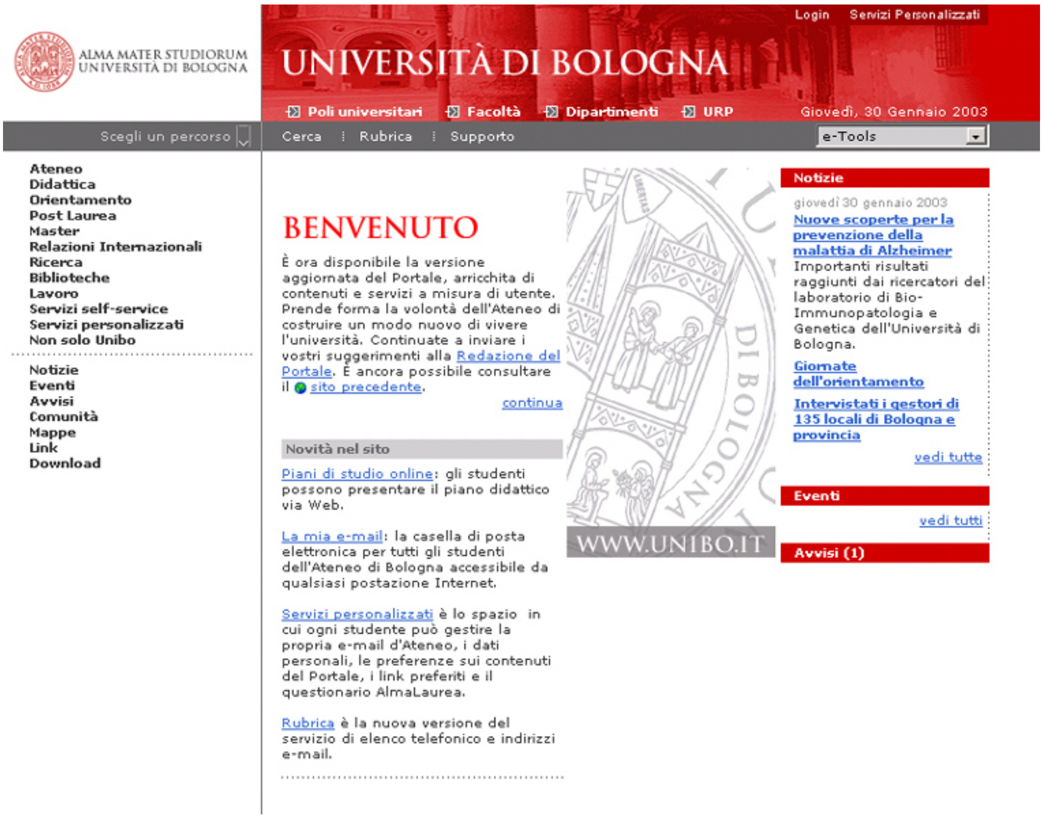
ALMAE MATRIS ALUMNI
(Associazione degli ex allievi)

Standard da utilizzare per la creazione delle pagine web di questo sito

Bando di concorso per il part-time studentesco (n. 2.664 posti)

Il sito è ottimizzato per NS4 e IE4
sviluppato su Mac OS
implementato su server NT
Visitatori dal 10/09/1998

Figure A.3: Unibo.it homepage in 1999.



BENVENUTO

È ora disponibile la versione aggiornata del Portale, arricchita di contenuti e servizi a misura di utente. Prende forma la volontà dell'Ateneo di costruire un modo nuovo di vivere l'università. Continuate a inviare i vostri suggerimenti alla Redazione del Portale. È ancora possibile consultare il sito precedente.

continua

Novità nel sito

Piani di studio online: gli studenti possono presentare il piano didattico via Web.

La mia e-mail: la casella di posta elettronica per tutti gli studenti dell'Ateneo di Bologna accessibile da qualsiasi postazione Internet.

Servizi personalizzati è lo spazio in cui ogni studente può gestire la propria e-mail d'Ateneo, i dati personali, le preferenze sui contenuti del Portale, i link preferiti e il questionario AlmaLaurea.

Rubrica è la nuova versione del servizio di elenco telefonico e indirizzi e-mail.

Notizie

giovedì 30 gennaio 2003

Nuove scoperte per la prevenzione della malattia di Alzheimer

Importanti risultati raggiunti dai ricercatori del laboratorio di Bio-Immunopatologia e Genetica dell'Università di Bologna.

Giornate dell'orientamento

Intervistati i gestori di 135 locali di Bologna e provincia

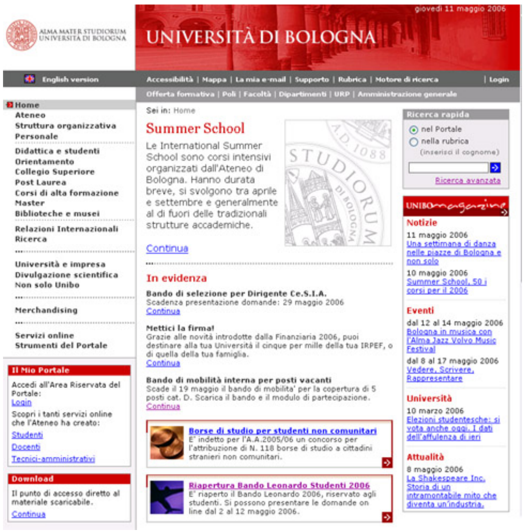
vedi tutte

Eventi

vedi tutti

Avvisi (1)

Figure A.4: Unibo.it homepage between 2002 and 2003.



Sei in: Home

Summer School

Le International Summer School sono corsi intensivi organizzati dall'Ateneo di Bologna. Hanno durata breve, si svolgono tra aprile e settembre e generalmente al di fuori delle tradizionali strutture accademiche.

Continua

In evidenza

Bando di selezione per Dirigente Ce.S.I.A.
Scadenza presentazione domande: 29 maggio 2006
Continua

Mettici la firma!
Grazie alle novità introdotte dalla Finanziaria 2006, puoi destinare alla tua Università il cinque per mille della tua IRPEF, o di quella della tua famiglia.
Continua

Bando di mobilità interna per posti vacanti
Scade il 19 maggio il bando di mobilita' per la copertura di 5 posti cat. D. Scarica il bando e il modulo di partecipazione.
Continua

Borse di studio per studenti non comunitari
E' indetto per l'A.A.2005/06 un concorso per l'attribuzione di N. 118 borse di studio a cittadini stranieri non comunitari.

Riapertura Bando Leonardo Studenti 2006
E' riaperto il Bando Leonardo 2006, riservato agli studenti. Si possono presentare le domande on line dal 2 al 12 maggio 2006.

Il Mio Portale

Accedi all'Area Riservata del Portale:
Login

Scopri i tanti servizi online che l'Ateneo ha creato:
Studenti
Docenti
Tecnici-amministrativi

Download

Il punto di accesso diretto al materiale scaricabile.
Continua

Notizie

11 maggio 2006
Una settimana di danza nelle piazze di Bologna e non solo

10 maggio 2006
Summer School, 50 i corsi per il 2006

Eventi

dal 12 al 14 maggio 2006
Bologna in musica con l'Alma Jazz Volvo Music Festival

dal 8 al 17 maggio 2006
Vedere, Scrivere, Rappresentare

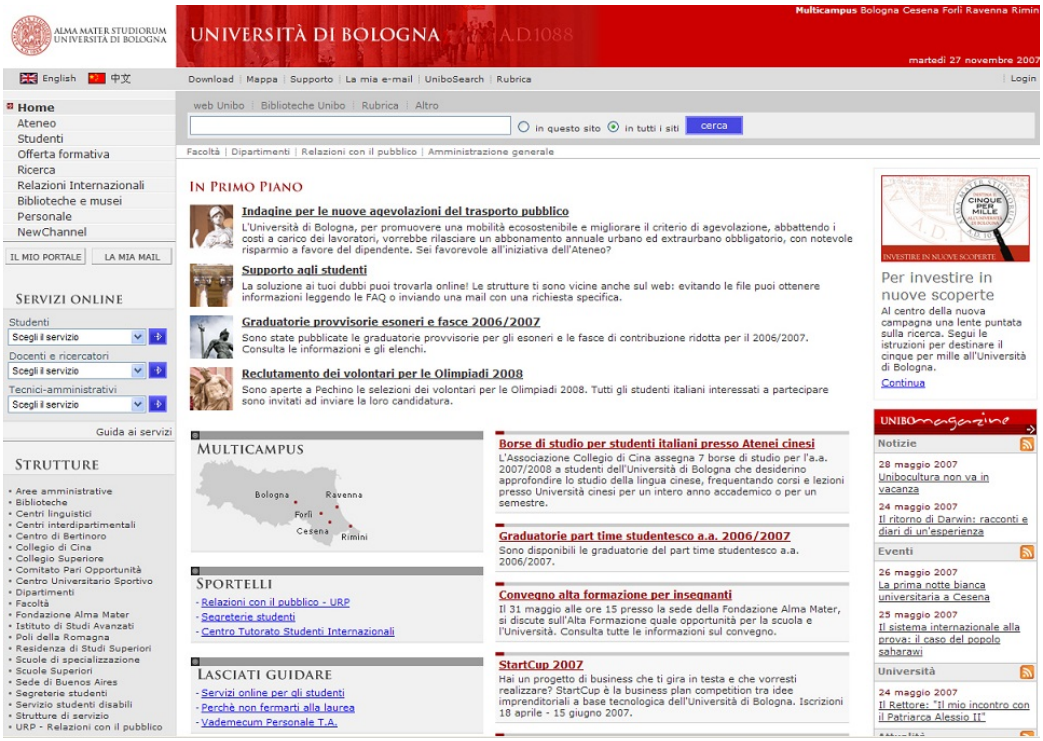
Università

10 marzo 2006
Elezioni studentesche: si vota anche oggi. I dati dell'affulenza di ieri

Attualità

8 maggio 2006
La Shakespeare Inc. Storia di un intramontabile mito che diventa un'industria.

Figure A.5: Unibo.it homepage between 2004 and 2006.


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
UNIVERSITÀ DI BOLOGNA A.D.1088
Multicampus Bologna Cesena Forlì Ravenna Rimini
martedì 27 novembre 2007
English
中文
Download | Mappa | Supporto | La mia e-mail | UniboSearch | Rubrica
Login
web Unibo | Biblioteche Unibo | Rubrica | Altro
○ in questo sito ● in tutti i siti
cerca
Facoltà | Dipartimenti | Relazioni con il pubblico | Amministrazione generale
Home
Ateneo
Studenti
Offerta formativa
Ricerca
Relazioni Internazionali
Biblioteche e musei
Personale
NewChannel
IL MIO PORTALE
LA MIA MAIL
Scegli il servizio
Guida ai servizi
Bologna
Ravenna
Forlì
Cesena
Rimini
CINQUE PER MILLE
INVESTIRE IN NUOVE SCOPERTE
UNIBOmagazine

SERVIZI ONLINE

Studenti

Docenti e ricercatori

Tecnici-amministrativi

STRUTTURE

- Aree amministrative
- Biblioteche
- Centri linguistici
- Centri interdipartimentali
- Centro di Bertinoro
- Collegio di Cina
- Collegio Superiore
- Comitato Pari Opportunità
- Centro Universitario Sportivo
- Dipartimenti
- Facoltà
- Fondazione Alma Mater
- Istituto di Studi Avanzati
- Poli della Romagna
- Residenza di Studi Superiori
- Scuole di specializzazione
- Scuole Superiori
- Sede di Buenos Aires
- Segreterie studenti
- Servizio studenti disabili
- Strutture di servizio
- URP - Relazioni con il pubblico

IN PRIMO PIANO

Indagine per le nuove agevolazioni del trasporto pubblico
L'Università di Bologna, per promuovere una mobilità ecosostenibile e migliorare il criterio di agevolazione, abbattendo i costi a carico dei lavoratori, vorrebbe rilasciare un abbonamento annuale urbano ed extraurbano obbligatorio, con notevole risparmio a favore del dipendente. Sei favorevole all'iniziativa dell'Ateneo?

Supporto agli studenti
La soluzione ai tuoi dubbi puoi trovarla online! Le strutture ti sono vicine anche sul web: evitando le file puoi ottenere informazioni leggendo le FAQ o inviando una mail con una richiesta specifica.

Graduatorie provvisorie esoneri e fasce 2006/2007
Sono state pubblicate le graduatorie provvisorie per gli esoneri e le fasce di contribuzione ridotta per il 2006/2007. Consulta le informazioni e gli elenchi.

Reclutamento dei volontari per le Olimpiadi 2008
Sono aperte a Pechino le selezioni dei volontari per le Olimpiadi 2008. Tutti gli studenti italiani interessati a partecipare sono invitati ad inviare la loro candidatura.

MULTICAMPUS

SPORTELLI

- Relazioni con il pubblico - URP
- Segreterie studenti
- Centro Tutorato Studenti Internazionali

LASCIATI GUIDARE

- Servizi online per gli studenti
- Perchè non fermarti alla laurea
- Vademecum Personale T.A.

Borse di studio per studenti italiani presso Atenei cinesi
L'Associazione Collegio di Cina assegna 7 borse di studio per l'a.a. 2007/2008 a studenti dell'Università di Bologna che desiderino approfondire lo studio della lingua cinese, frequentando corsi e lezioni presso Università cinesi per un intero anno accademico o per un semestre.

Graduatorie part time studentesco a.a. 2006/2007
Sono disponibili le graduatorie del part time studentesco a.a. 2006/2007.

Convegno alta formazione per insegnanti
Il 31 maggio alle ore 15 presso la sede della Fondazione Alma Mater, si discute sull'Alta Formazione quale opportunità per la scuola e l'Università. Consulta tutte le informazioni sul convegno.

StartCup 2007
Hai un progetto di business che ti gira in testa e che vorresti realizzare? StartCup è la business plan competition tra idee imprenditoriali a base tecnologica dell'Università di Bologna. Iscrizioni 18 aprile - 15 giugno 2007.

Per investire in nuove scoperte

Al centro della nuova campagna una lente puntata sulla ricerca. Segui le istruzioni per destinare il cinque per mille all'Università di Bologna.

Continua

Notizie

28 maggio 2007
Unibocultura non va in vacanza

24 maggio 2007
Il ritorno di Darwin: racconti e diari di un'esperienza

Eventi

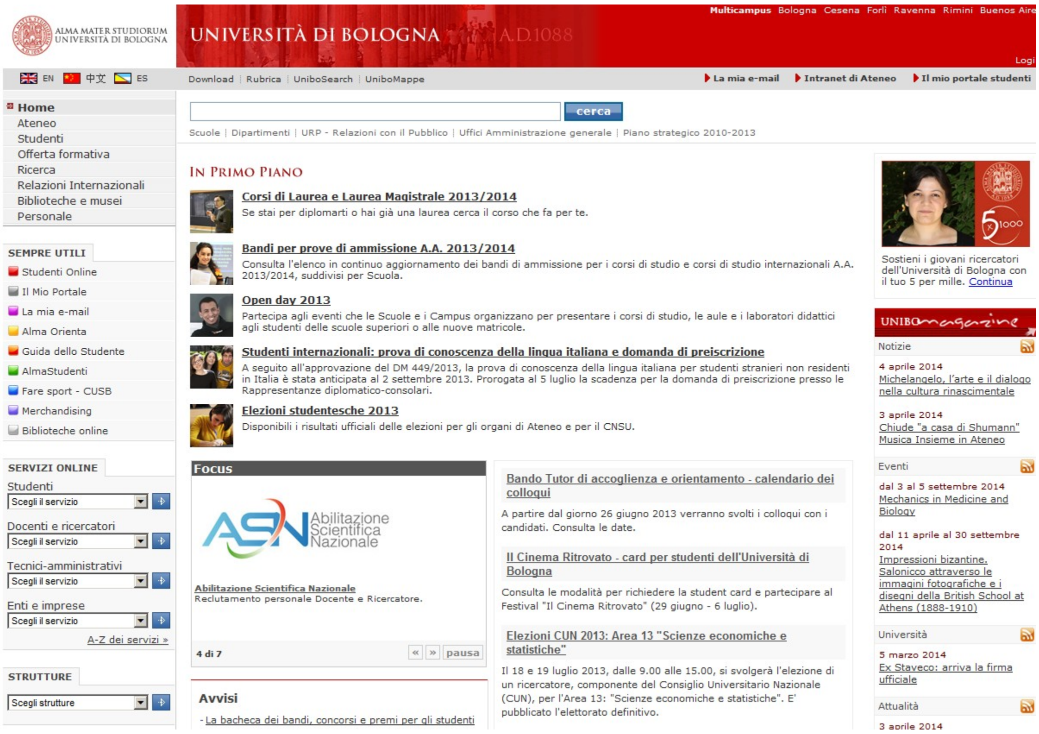
26 maggio 2007
La prima notte bianca universitaria a Cesena

25 maggio 2007
Il sistema internazionale alla prova: il caso del popolo saharawi

Università

24 maggio 2007
Il Rettore: "Il mio incontro con il Patriarca Alessio II"

Figure A.6: Unibo.it homepage between 2006 and 2009.



IN PRIMO PIANO

Corsi di Laurea e Laurea Magistrale 2013/2014
Se stai per diplomarti o hai già una laurea cerca il corso che fa per te.

Bandi per prove di ammissione A.A. 2013/2014
Consulta l'elenco in continuo aggiornamento dei bandi di ammissione per i corsi di studio e corsi di studio internazionali A.A. 2013/2014, suddivisi per Scuola.

Open day 2013
Partecipa agli eventi che le Scuole e i Campus organizzano per presentare i corsi di studio, le aule e i laboratori didattici agli studenti delle scuole superiori o alle nuove matricole.

Studenti internazionali: prova di conoscenza della lingua italiana e domanda di preiscrizione
A seguito all'approvazione del DM 449/2013, la prova di conoscenza della lingua italiana per studenti stranieri non residenti in Italia è stata anticipata al 2 settembre 2013. Prorogata al 5 luglio la scadenza per la domanda di preiscrizione presso le Rappresentanze diplomatico-consolari.

Elezioni studentesche 2013
Disponibili i risultati ufficiali delle elezioni per gli organi di Ateneo e per il CNSU.

Focus

Abilitazione Scientifica Nazionale

Abilitazione Scientifica Nazionale
Reclutamento personale Docente e Ricercatore.

Bando Tutor di accoglienza e orientamento - calendario dei colloqui
A partire dal giorno 26 giugno 2013 verranno svolti i colloqui con i candidati. Consulta le date.

Il Cinema Ritrovato - card per studenti dell'Università di Bologna
Consulta le modalità per richiedere la student card e partecipare al Festival "Il Cinema Ritrovato" (29 giugno - 6 luglio).

Elezioni CUN 2013: Area 13 "Scienze economiche e statistiche"
Il 18 e 19 luglio 2013, dalle 9.00 alle 15.00, si svolgerà l'elezione di un ricercatore, componente del Consiglio Universitario Nazionale (CUN), per l'Area 13: "Scienze economiche e statistiche". E' pubblicato l'elettorato definitivo.

Avvisi
- La bacheca dei bandi, concorsi e premi per gli studenti

Sostieni i giovani ricercatori dell'Università di Bologna con il tuo 5 per mille. Continua

Notizie

4 aprile 2014
Michelangelo, l'arte e il dialogo nella cultura rinascimentale

3 aprile 2014
Chiude "a casa di Shumann" Musica Insieme in Ateneo

Eventi

dal 3 al 5 settembre 2014
Mechanics in Medicine and Biology

dal 11 aprile al 30 settembre 2014
Impressioni bizantine. Salonicco attraverso le immagini fotografiche e i disegni della British School at Athens (1888-1910)

Università

5 marzo 2014
Ex Staveco: arriva la firma ufficiale

Attualità

3 aprile 2014

Figure A.7: Unibo.it homepage between 2009 and 2013.

Bibliography

Acerbi, A., Lampos, V., Garnett, P., and Bentley, R. A. (2013). The expression of emotions in 20th century books. *PloS one*, 8(3):e59030.

Adair-Toteff, C. (2014). Statistical origins of the "protestant ethic". *Journal of Classical Sociology.*

Ainsworth, S. G., Nelson, M. L., and Van de Sompel, H. (2015). Only one out of five archived web pages existed as presented. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 257–266. ACM.

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

Allington, D., Brouillette, S., and Golumbia, D. (2016). Neoliberal tools (and archives): A political history of digital humanities. *Los Angeles Review of Books.*

American-Historical-Association (2015). Guidelines for the professional evaluation of digital scholarship by historians.

Anderson, A., McFarland, D., and Jurafsky, D. (2012). Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21. Association for Computational Linguistics.

Ankerson, M. S. (2012). Writing web histories with an eye on the analog past. *New Media & Society*, 14(3):384–400.

Arnaudo, E. (2013). *Biomedicine and pain.* PhD thesis, University of Bologna.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Assmann, A. (2006). The printing press and the internet: From a culture of memory to a culture of attention. *Globalization, cultural identities, and media representations*, pages 11–25.

Au Yeung, C.-m. and Jatowt, A. (2011). Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1231–1240. ACM.

Ayers, E. L. and Rubin, A. S. (2000). *Valley of the Shadow: Two Communities in the American Civil War.* WW Norton & Company.

Baldissara, L. (1994). *Per una città più bella e più grande: il governo municipale di Bologna negli anni della ricostruzione (1945-1956).* il Mulino.

Banerjee, R. (2013). *An Innovation System Perspective on Adaptation Strategies to Climate Variability and Water Management in India.* PhD thesis, University of Bologna.

Barbagli, M., Colombo, A., and Orzi, R. (2009). *Gli studenti e la città: primo rapporto sugli studenti dell'Università di Bologna.* Bologna University Press.

Barthes, R. (1967). Discourse on history. *Social Science Information*, 6(4):65–75.

Bastian, M., Heymann, S., Jacomy, M., et al. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8:361–362.

Bellettini, A. and Tassinari, F. (1984). *La città ei gruppi sociali: Bologna fra gli anni cinquanta e settanta.* Clueb.

Ben-David, A. (2011). The emergence of the palestinian web-space: a digital history of a digital landscape. In *MiT7 Unstable Platforms: The Promise and Peril of Transition.*

Ben-David, A. (2016). What does the web remember of its deleted past? an archival reconstruction of the former yugoslav top-level domain. *New Media & Society.*

Ben-David, A. and Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. *Alexandria: The Journal of National and International Library and Information Issues*, 25(1-2):93–111.

Benvenuti, N. and Morriello, R. (2006). *Gestione delle raccolte e cooperazione nella biblioteca ibrida. Atti del convegno (Firenze, 13 ottobre 2005).* Firenze University Press.

Berendt, B., Büchler, M., and Rockwell, G. (2015). Is it research or is it spying? thinking-through ethics in big data ai and other knowledge sciences. *KI-Künstliche Intelligenz*, 29(2):223–232.

Bergamin, G. (2012). La raccolta dei siti web: un test per il dominio "punto it". *DigItalia*, 2:170–174.

Bergamin, G. and Messina, M. (2012). Magazzini digitali: dal prototipo al servizio. *DigItalia*, 1:115–122.

Berry, D. M. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, 12(0):2.

Bicho, D. and Gomes, D. (2016). Preserving websites of research and development projects. In *International Conference on Digital Preservation.*

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia – A crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Blevins, C. (2014). Space, nation, and the triumph of region: A view of the world from houston. *Journal of American History*, 101(1):122–147.

Blevins, C. (2015). The perpetual sunrise of methodology. Personal website: `http://www.cameronblevins.org/posts/perpetual-sunrise-methodology/`.

Bloch, M. (1949). *The historian's craft.* Manchester University Press.

Bogdanov, P. and Mohr, J. W. (2013). Topic models. what they are and why they matter. *Poetics*, 31:545–569.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Bornmann, L. and Leydesdorff, L. (2014). Scientometrics in a changing research landscape. *EMBO reports*, 15(12):1228–1232.

Boschetti, F., Romanello, M., Babeu, A., Bamman, D., and Crane, G. (2009). Improving ocr accuracy for classical critical editions. In *Research and Advanced Technology for Digital Libraries*, pages 156–167. Springer.

Braudel, F. (1972). *The Mediterranean and the Mediterranean world in the age of Philip II.* Univ of California Press.

Brauer, R. and Fridlund, M. (2013). Historicizing topic models, a distant reading of topic modeling texts within historical studies. In *International Conference on Cultural Research in the context of "Digital Humanities", St. Petersburg: Russian State Herzen University*.

Brizzi, G. P., Marini, L., and Pombeni, P. (1988). *L'Università a Bologna: maestri, studenti e luoghi dal XVI al XX secolo*. Cassa di Risparmio.

Brockliss, L. W. (1978). Patterns of attendance at the university of paris, 1400–1800. *The Historical Journal*, 21(03):503–544.

Bromham, L., Dinnage, R., and Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature*, 534(7609):684–687.

Brügger, N. (2005). *Archiving Websites. General Considerations and Strategies*. The Centre for Internet Research.

Brügger, N. (2008). The archived website and website philology. *Nordicom Review*, 29(2):155–175.

Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, 11(1-2):115–132.

Brügger, N. (2010). *Web history*. Peter Lang.

Brügger, N. (2011a). Digital history and a register of websites. In *Long History of New Media*. Peter Lang.

Brügger, N. (2011b). Web archiving—between past, present, and future. *The handbook of Internet studies*, 11:24.

Brügger, N. (2012a). Historical network analysis of the web. *Social Science Computer Review*.

Brügger, N. (2012b). Web historiography and internet studies: Challenges and perspectives. *New Media & Society*.

Brügger, N. (2012c). When the present web is later the past: Web historiography, digital history, and internet studies. *Historical Social Research/Historische Sozialforschung*, pages 102–117.

Brügger, N. (2016). Digital humanities in the 21st century: Digital material as a driving force. *Digital Humanities Quarterly*, 10(3).

Büchler, M., Franzini, G., Franzini, E., and Moritz, M. (2014). Scaling historical text re-use. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 23–31. IEEE.

Buechler, M., Heyer, G., and Gründer, S. (2008). eAQUA–bringing modern text mining approaches to two thousand years old ancient texts. In *Proceedings of e-Humanities–An Emerging Discipline, workshop at the 4th IEEE International Conference on e-Science*.

Burke, P. (2008). *What is cultural history*. Polity.

Busa, R. (1980). The annals of humanities computing: The index thomisticus. *Computers and the Humanities*, 14(2):83–90.

Bush, V. (1945). As we may think. *The atlantic monthly*, 176(1):101–108.

Buzydlowski, J. W., White, H. D., and Lin, X. (2002). Term co-occurrence analysis as an interface for digital libraries. In *Visual interfaces to digital libraries*, pages 133–144. Springer.

Carr, E. H. (1961). *What is history?* Cambridge University Press.

Chaitin, D. (2016). Obama scolds reporters for twitter obsession. *Washington Examiner*.

Chakraborty, A. and Nanni, F. (2017). The changing digital faces of science museums: A diachronic analysis of museum websites. In Brügger, N., editor, *Web 25—Histories from the first 25 years of the World Wide Web*. Peter Lang.

Chaney, A. J.-B. and Blei, D. M. (2012). Visualizing topic models. In *ICWSM*.

Chang, A. X., Spitkovsky, V. I., Manning, C. D., and Agirre, E. (2016). Evaluating the word-expert approach for named-entity disambiguation. *arXiv preprint arXiv:1603.04767*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296.

Chiapparini, M. (1995). Interattività e nuove tecnologie: il caso di internet. Master's thesis, University of Bologna.

Chiara, S. (1998). La telematica e la città. il progetto iperbole a bologna. Master's thesis, University of Bologna.

Chuang, J., Manning, C. D., and Heer, J. (2012a). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM.

Chuang, J., Ramage, D., Manning, C., and Heer, J. (2012b). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM.

Cohen, D. (2005). By the book: Assessing the place of textbooks in us survey courses. *The Journal of American History*, 91(4):1405–1415.

Cohen, D. (2006a). From babel to knowledge: Data mining large digital collections. *D-Lib Magazine*, 12(3):3.

Cohen, D. (2006b). When machines are the audience. Personal website: `http://www.dancohen.org/2006/03/02/when-machines-are-the-audience/`.

Cohen, D. (2010a). Initial thoughts on the google books ngram viewer and datasets. Personal website: `http://www.dancohen.org/2010/12/19/initial-thoughts-on-the-google-books-ngram-viewer-and-datasets/`.

Cohen, D. (2010b). Searching for the victorians. Personal website: `http://www.dancohen.org/2010/10/04/searching-for-the-victorians/`.

Cohen, D. (2011). A million syllabi. Personal website: `http://www.dancohen.org/2011/03/30/a-million-syllabi/`.

Cohen, D., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Taylor, A. M., Thomas, W. G., and Turkel, W. J. (2008). Interchange: The promise of digital history. *The Journal of American History*, 95(2):452–491.

Cohen, D., Gibbs, F., Hitchcock, T., Rockwell, G., Sander, J., Shoemaker, R., Sinclair, S., Takats, S., Turkel, W. J., Briquet, C., et al. (2011). Data mining with criminal intent. *Final white paper.*

Cohen, D. and Rosenzweig, R. (2006). *Digital history: a guide to gathering, preserving, and presenting the past on the web*, volume 28. University of Pennsylvania Press Philadelphia.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Committee, R. (1985). *Report of the Steering Committee for Efficiency Studies in Universities, Committee of Vice-Chancellors and Principals.*

Cornolti, M., Ferragina, P., and Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 249–260. ACM.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

Crane, G. (2006). What do you do with a million books? *D-Lib magazine*, 12(3):1.

Crocetti, D. (2011). *Medicalizing gender: from intersex to DSD, from the laboratory to patient groups*. PhD thesis, University of Bologna.

Crymble, A. (2015a). A comparative approach to identifying the irish in long eighteenth-century london. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 48(3):141–152.

Crymble, A. (2015b). Historians are becoming computer science customers – postscript. Personal website: `https://ihrdighist.blogs.sas.ac.uk/2015/06/24/historians-are-becoming-computer-science-customers-postscript/`.

Crymble, A. (2015c). *Surname Analysis, Distant Reading, and Migrant Experience: The Irish in London, 1801-1820*. PhD thesis, King's College London (University of London).

Crymble, A., Gibbs, F., Hegel, A., McDaniel, C., Milligan, I., Posner, M., and Turkel, W. J. (2012). The programming historian.

Dalbello, M. (2011). A genealogy of digital humanities. *Journal of Documentation*, 67(3):480–506.

Davis, C. (2014). Archiving the web: A case study from the university of victoria. *Code4Lib Journal*.

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Scarecrow Press.

de Wied, D. et al. (1991). Postgraduate research training today: emerging structures for a changing europe. *Netherlands Ministry of Education and Science, The Hague*.

Derrida, J. (1967). *Of grammatology*. JHU Press.

Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning*, pages 233–240.

Dill, D. D. and Soo, M. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher education*, 49(4):495–533.

Dougherty, J. and Nawrotzki, K. (2013). *Writing history in the digital age.* University of Michigan Press Ann Arbor.

Dougherty, M., Meyer, E. T., Madsen, C. M., Van den Heuvel, C., Thomas, A., and Wyatt, S. (2010). Researcher engagement with web archives: State of the art. *Joint Information Systems Committee Report.*

Dröscher, A. (2002). *Le facoltà medico-chirurgiche italiane: 1860-1915: repertorio delle cattedre e degli stabilimenti annessi, dei docenti, dei liberi docenti e del personale scientifico.* CLUEB.

Eisenstein, E. L. (1980). *The printing press as an agent of change*, volume 1. Cambridge University Press.

Elton, G. R. (1967). *The practice of history.* Wiley-Blackwell.

Engels, F. (1850). *The peasant war in Germany.* Foreign Languages Publishing House.

Evans, R. J. (1997). *In defence of history.* Granta Books.

Ewing, T., Gad, S., Housman, B., Kerr, K., Pencek, B., and Ramakrishnan, N. (2014). Mining coverage of the flu: Big data's insights into an epidemic. *Perspectives on History*, 52(2).

Ferragina, P. and Scaiella, U. (2010). TagMe: on-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

Fielden, J., Lockwood, G., and Nind, R. (1973). *Planning and management in universities: a study of British universities.* Chatto & Windus.

Findlen, P., Edelstein, D., and Coleman, N. (2011). Mapping the republic of letters. `https://republicofletters.stanford.edu`.

Fischer, C. S. (2013). Digital humanities, big data, and ngrams. *Boston Review.*

Fish, S. (2012a). The digital humanities and the transcending of mortality. *New York Times.*

Fish, S. (2012b). Mind your p's and b's: The digital humanities and interpretation. *New York Times.*

Foot, K. and Schneider, S. (2010). Object-oriented web historiography. In Brügger, N., editor, *Web History.* Peter Lang.

Foot, K., Schneider, S. M., Dougherty, M., Xenos, M., and Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 us electoral web sphere. *Journal of Computer-Mediated Communication*, 8(4):0–0.

Foucault, M. (1972). The archaeology of knowledge, trans. *AM Sheridan Smith (New York: Pantheon, 1972)*, 24.

Foucault, M. (1975). *Discipline & punish: The birth of the prison*. Vintage.

Fox, R. and Guagnini, A. (1993). *Education, technology and industrial performance in Europe, 1850-1939*. Cambridge University Press.

Frana, P. L. (2004). Before the web there was gopher. *IEEE Annals of the History of Computing*, pages 20–41.

Friedberg, E. and Musselin, C. (1987). The academic profession in france. *The Academic Profession: National, Disciplinary, and Institutional Settings*, pages 93–122.

Gaddis, J. L. (2002). *The landscape of history: How historians map the past*. Oxford University Press.

Gandolfi, G. (2011). *Imagines illustrium virorum: la collezione dei ritratti dell'Università e della Biblioteca Universitaria di Bologna*. Clueb.

Gebeil, S. (2014). Les mémoires de l'immigration maghrébine sur le web français (1996-2013). *Migrations societe*, pages 165–179.

Gerrish, S. and Blei, D. M. (2010a). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 375–382.

Gerrish, S. and Blei, D. M. (2010b). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning*, pages 375–382.

Giannakopoulos, T., Foufoulas, I., Stamatogiannakis, E., Dimitropoulos, H., Manola, N., and Ioannidis, Y. (2014). Discovering and visualizing interdisciplinary content classes in scientific publications. *D-Lib Magazine*, 20(11):4.

Gibbs, F. and Owens, T. (2012). The hermeneutics of data and historical writing. *Writing History in the Digital Age*, 159.

Gibbs, F. W. and Cohen, D. J. (2011). A conversation with data: Prospecting victorian words and ideas. *Victorian Studies*, 54(1):69–77.

Giles, J. (2012). The digital lab. *Nature*, 481(7382):430.

Glass, B. (1979). Milestones and rates of growth in the development of biology. *Quarterly Review of Biology*, pages 31–53.

Gomes, D., Miranda, J., and Costa, M. (2011). A survey on web archiving initiatives. In *Research and advanced technology for digital libraries*, pages 408–420. Springer.

Gorsky, M. (2015). Into the dark domain: The uk web archive as a source for the contemporary history of public health. *Social History of Medicine*, 28(3):596–616.

Grafton, A. (2011). Loneliness and freedom. *Perspectives on History.*

Graham, S., Milligan, I., and Weingart, S. B. W. (2016). *Exploring big historical data: The historian's macroscope.* Imperial College Press.

Greenwald, G. (2014). *No place to hide: Edward Snowden, the NSA, and the US surveillance state.* Macmillan.

Greif, A. (1997). Cliometrics after 40 years. *The American Economic Review*, 87(2):400–403.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297.

Gross, P. R. and Levitt, N. (1996). Higher superstition: The academic left and its quarrels with science. *ISIS-International Review Devoted to the History of Science and its Cultural Influence*, 87(2):323–327.

Guagnini, A. (1988). Higher education and the engineering profession in italy: The scuole of milan and turin, 1859–1914. *Minerva*, 26(4):512–548.

Guston, D. H. and Keniston, K. (1994). *The fragile contract: University science and the federal government.* MIT Press.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., and Margetts, H. (2014). Mapping the uk webspace: Fifteen years of british universities on the web. In *Proceedings of the 2014 ACM conference on Web science*, pages 62–70. ACM.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics.

Harvey, D. (1989). The condition of postmodernity: An enquiry into the origins of social change. *Malden, MA: Blackwell.*

Hasan, K. S. and Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1262–1273.

Hockey, S. (2004). The history of humanities computing. *A companion to digital humanities*, pages 3–19.

Hockx-Yu, H. (2014). Access and scholarly use of web archives. *Alexandria: The Journal of National and International Library and Information Issues*, 25(1-2):113–127.

Hof, R. D. (2013). Deep learning. *Technology Review*, 116(3):32–36.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57.

Holm, P., Goodsite, M. E., Cloetingh, S., Agnoletti, M., Moldan, B., Lang, D. J., Leemans, R., Moeller, J. O., Buendía, M. P., and Pohl, W. (2013). Collaboration between the natural, social and human sciences in global change research. *Environmental science & policy*, 28:25–35.

Holzmann, H., Goel, V., and Anand, A. (2016a). Archivespark: Efficient web archive access, extraction and derivation. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 83–92. ACM.

Holzmann, H., Nejdl, W., and Anand, A. (2016b). The dawn of today's popular domains: A study of the archived german web over 18 years. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 73–82. ACM.

Hoppe, R. (2005). Rethinking the science-policy nexus: from knowledge utilization and science technology studies to types of boundary arrangements. *Poiesis & Praxis*, 3(3):199–215.

Houston, N. M. (2014). Toward a computational analysis of victorian poetics. *victorian studies*, 56(3):498–510.

Howell, B. A. (2006). Proving web history: How to use the internet archive. *Journal of Internet Law*, 9(8):3–9.

Howell, M. C. and Prevenier, W. (2001). *From reliable sources: An introduction to historical methods*. Cornell University Press.

Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pages 465–474.

Huurdeman, H. C., Ben-David, A., Kamps, J., Samar, T., and de Vries, A. P. (2014). Finding pages on the unarchived web. In *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*, pages 331–340. IEEE.

Ide, N. and Véronis, J. (1995). *Text encoding initiative: Background and contexts*, volume 29. Springer Science & Business Media.

Iggers, G. G. (2005). *Historiography in the twentieth century: From scientific objectivity to the postmodern challenge.* Wesleyan University Press.

Iori, L. (2013). *Agricultural genetics and plant breeding in early Twentieth-Century Italy.* PhD thesis, University of Bologna.

Jackson, A. (2015). Ten years of the uk web archive: what have we saved? Personal website: `http://anjackson.net/2015/04/27/what-have-we-saved-iipc-ga-2015`.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. *Proc. of EuroVis—STARs*, pages 83–103.

Joachims, T. (1996). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, DTIC Document.

Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with many Relevant Features.* Springer.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history.* University of Illinois Press.

Jockers, M. L. (2014). Topic modeling. In *Text Analysis with R for Students of Literature*, pages 135–159. Springer.

Jockers, M. L. (2015). Revealing sentiment and plot arcs with the syuzhet package. Personal website: `http://www.matthewjockers.net/2015/02/02/syuzhet/`.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334.

Kaufman, M. (2015). Everything on paper will be used against me: Quanmatifying kissinger. Personal website: `http://blog.quantifyingkissinger.com/`.

Kirschenbaum, M. (2012). What is digital humanities and what's it doing in english departments? *Debates in the digital humanities*, 3.

Kirschenbaum, M. G. (2007). The remaking of reading: Data mining and the digital humanities. In *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Baltimore, MD.*

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):2053951714528481.

Kleinveldt, L. and Booysen, D. (2015). Collaboration and social networks between chemistry researchers: what does this mean for academic libraries? *International Association of University Libraries.*

Knowles, A. K. (2008). Gis and history. *Placing history: How maps, spatial data, and GIS are changing historical scholarship*, pages 1–13.

Knowles, A. K. and Hillier, A. (2008). *Placing history: how maps, spatial data, and GIS are changing historical scholarship.* ESRI, Inc.

Kornblith, G. and Lasser, C. (2001). Teaching the american history survey at the opening of the twenty-first century: a round table discussion. *The Journal of American History*, 87(4):1409–1441.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* University of Chicago press.

Kunny, T. (1997). A digital dark ages? challenges in the preservation of electronic information of electronic information. In *63rd IFLA Conuncil and General Conference.*

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242.

LaFrance, A. (2015). Raiders of the lost web. *The Atlantic.*

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Harvard university press.

Latour, B. (1991). *We have never been modern.* Harvard University Press.

Latour, B. and Woolgar, S. (1979). *Laboratory life: The construction of scientific facts.* Princeton University Press.

Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1536–1545.

Lauscher, A., Nanni, F., Fabo, P. R., and Ponzetto, S. P. (2016). Entities as topic labels: Combining entity linking and labeled lda to improve topic interpretability and evaluability. *Italian Journal of Computational Linguistics.*

Lebow, R. N. (1990). Domestic politics and the cuban missile crisis: The traditional and revisionist interpretations reevaluated. *Diplomatic History*, 14(4):471–492.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Leonard, P. (2014). Mining large datasets for the humanities. *IFLA WLIC*, pages 16–22.

Lepore, J. (2015). The cobweb: Can the internet be archived? *The New Yorker*.

Liu, A. (2012). The state of the digital humanities a report and a critique. *Arts and Humanities in Higher Education*, 11(1-2):8–41.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Lu, K. and Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, 63(10):1973–1986.

Lyman, P. and Kahle, B. (1998). Archiving digital cultural artifacts. *D-Lib Magazine*, 4(7).

Lyotard, J.-F. (1984). *The postmodern condition: A report on knowledge*, volume 10. U of Minnesota Press.

Macgregor, H. C. (2010). The biological sciences. In Rüegg, W., editor, *A history of the university in Europe*. Cambridge University Press.

Macleod, R. and Moseley, R. (1978). Breadth, depth and excellence: Sources and problems in the history of university science education in england, 1850-1914. *Studies in Science and Education*.

Maemura, E., Becker, C., and Milligan, I. (2016). Understanding computational web archives research methods using research objects. *Proceedings of the International Conference on Big Data*.

Mahoney, M. S. (1988). The history of computing in the history of technology. *Annals of the History of Computing*, 10(2):113–125.

Manning, C. D. (2016). Computational linguistics and deep learning. *Computational Linguistics*.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2:460–475.

Marche, S. (2012). Literature is not data: Against digital humanities. *Los Angeles Review of Books*, 28.

Marshall, C. C. and Shipman, F. M. (2014). An argument for archiving facebook as a heterogeneous personal store. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 11–20. IEEE Press.

Martini, F. (2012). *Tracciature Digitali: la conoscenza nell'era informazionale*. PhD thesis, University of Bologna.

Marx, K. (1867). *Capital*. Harmondsworth: Penguin/New Left Review.

Masanès, J. (2006). *Web archiving*. Springer.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mazzetti, S. (1848). *Repertorio di tutti i professori antichi, e moderni della famosa Università, e del celebre Istituto delle scienze di Bologna con in fine Alcune aggiunte e correzioni alle opere dell'Alidosi, del Cavazza, del Sarti, del Fantuzzi, e del Tiraboschi compilati da Serafino Mazzetti bolognese archivista arcivescovile*.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2005). The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. *Computer Science Department Faculty Publication Series*.

McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

McCloskey, D. N. (1978). The achievements of the cliometric school. *The Journal of Economic History*, 38(01):13–28.

Medland, W. J. (1990). The cuban missile crisis: Evolving historical perspectives. *The History Teacher*, 23(4):433–447.

Meeks, E. and Weingart, S. B. (2012). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1):1–6.

Meeks, E. and Weingart, S. B. (2013). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1):2–1.

Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 490–499.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8.

Merriman, B. (2015). A science of literature. *Boston Review*.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Milligan, I. (2012). Mining the 'internet graveyard': Rethinking the historians' toolkit. *Journal of the Canadian Historical Association/Revue de la Société historique du Canada*, 23(2):21–64.

Milligan, I. (2016a). Lost in the infinite archive: The promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing*, 10(1):78–94.

Milligan, I. (2016b). Web archives and born-digital sources workshop: Challenges, future steps, and the field. Personal website: `https://ianmilligan.ca/2016/06/10/web-archives-and-born-digital-sources-workshop-challenges-future-steps-and-the-field/`.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

Mitkov, R. (2005). *The Oxford handbook of computational linguistics*. Oxford University Press.

Momack, D. (2003). *Who Owns History?* Cabinet Magazine.

Moretti, F. (2000). Conjectures on world literature. *New left review*, 1:54.

Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history*. Verso.

Moretti, F. (2013). *Distant reading*. Verso Books.

Moretti, G., Sprugnoli, R., and Tonelli, S. (2015). Digging in the dirt: Extracting keyphrases from texts with kd. *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, page 198.

Moretti, G., Tonelli, S., Menini, S., and Sprugnoli, R. (2014). Alcide: an online platform for the analysis of language and content in a digital environment. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 270–274. Pisa University Press.

Moreux, J.-P. (2016). Data mining historical newspaper metadata. In *Proceedings of the IFLA International News Media Conference.*

Moritz, M., Wiederhold, A., Pavlek, B., Bizzoni, Y., and Büchler, M. (2016). Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1849–1859.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Munslow, A. (2006). *Deconstructing history.* Taylor & Francis.

Murphy, J., Hashim, N. H., and O'Connor, P. (2007). Take me back: validating the wayback machine. *Journal of Computer-Mediated Communication*, 13(1):60–75.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective.* MIT press.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nanni, F. (2013). L'archiviazione delle pagine dei quotidiani online. *Diacronie. Studi di Storia Contemporanea.*

Nanni, F. (2014). Managing educational information on university websites: a proposal for unibo.it. In *2nd Annual Conference of the Italian Association for Digital Humanities (Aiucd 2013): Collaborative Research Practices and Shared Infrastructures for Humanities Computing*, pages 279–286. CLEUP.

Nanni, F. (2015). Historical method and born-digital primary sources: A case study of italian university websites. *Officina della Storia.*

Nanni, F. (2017). Reconstructing a website's lost past - methodological issues concerning the history of www.unibo.it. *Under peer review.*

Nanni, F., Dietz, L., Faralli, S., Glavaš, G., and Ponzetto, S. P. (2016a). Capturing interdisciplinarity in academic abstracts. *D-Lib Magazine*, 22(9/10).

Nanni, F., Kümper, H., and Ponzetto, S. P. (2016b). Semi-supervised textual analysis and historical research helping each other: Some thoughts and observations. *International Journal of Humanities and Arts Computing*, 10(1):63–77.

Nanni, F. and Ruiz Fabo, P. (2016). Entities as topic labels: Improving topic interpretability and evaluability combining entity linking and labeled LDA. *Digital Humanities 2016*.

Negrini, D. (1998). *Repertorio nazionale degli storici dell'università: 1993-1997*. CLUEB.

Nelson, R. K. (2010). Mining the dispatch. *Mining the dispatch*.

Nelson, R. K. (2016). Digital humanities as appendix. *American Quarterly*, 68(1):131–136.

Nesbit, S. and Ayers, E. L. (2013). *Visualizing Emancipation*. University of Richmond, Digital Humanities Lab.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *HLT: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the national science foundation. *Scientometrics*, 100(3):741–754.

Nielsen, J. (2014). *DR's undervisning på tværs af medier: En historisk undersøgelse af mediesamspil*. PhD thesis, Department of Aesthetics and Communication, Faculty of Arts, Aarhus University.

Niu, J. (2012). An overview of web archiving. *D-Lib magazine*, 18(3):2.

Noiret, S. (2009). The digital historian's craft and the role of the european history primary sources (ehps) portal. *Archivi & Computer. Automazione e Beni Culturali*.

Noiret, S. (2015). Digital public history: bringing the public back in. *Public History Weekly*, 3.

North, D. C. and Thomas, R. P. (1973). *The rise of the western world: A new economic history*. Cambridge University Press.

Novick, P. (1988). *That noble dream: The'objectivity question'and the American historical profession*. Cambridge University Press.

Nowviskie, B. (2014). On the origin of "hack" and "yack". *Journal of Digital Humanities*, 3(2):3–2.

Owens, T. (2012). Discovery and justification are different: Notes on science-ing the humanities. Personal Website: `http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/`.

Owens, T. (2015). Digital sources & digital archives: The evidentiary basis of digital history (draft). Personal Website: `http://www.trevorowens.org/2015/12/digital-sources-digital-archives-the-evidentiary-basis-of-digital-history-draft/`.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: bringing order to the web. *Technical Report*.

Panajoli, T. (2012). *Una prospettiva storico-filosofica sull'evoluzione del curricolo di fisica nella scuola: problemi e innovazioni pedagogiche*. PhD thesis, University of Bologna.

Pancaldi, G. (1993a). *Le Università E Le Scienze: Prospettive Storiche E Attuali: Relazioni Presentate Al Convegno Internazionale, Bologna, 18 Settembre 1991*. Università di Bologna.

Pancaldi, G. (1993b). Vito volterra: Cosmopolitan ideals and nationality in the italian scientific community between the belle époque and the first world war. *Minerva*, 31(1):21–37.

Pancaldi, G. (2006). Wartime chemistry in italy: Industry, the military, and the professors. In *Frontline and Factory: Comparative Perspectives on the Chemical Industry at War, 1914–1924*, pages 61–74. Springer.

Parolini, G. (2013). *Making Sense of Figures: Statistics, Computing and Information Technologies in Agriculture and Biology in Britain, 1920s-1960s*. PhD thesis, University of Bologna.

Parsons, K. M. (2003). *The science wars: Debating scientific knowledge and technology*. Prometheus Books.

Pasquini, E. (1993). *Intertestualità e intratestualità nella "Commedia" dantesca - La tradizione del Novecento poetico. Appunti dalle lezioni del corso monografico 1992-93*. C.U.S.L.

Paterson, T. G. and Brophy, W. J. (1986). October missiles and november elections: The cuban missile crisis and american politics, 1962. *The Journal of American History*, 73(1):87–119.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*, volume 14, pages 1532–1543.

Perkel, J. M. (2015). The trouble with reference rot. *Nature*, 521(7550):111–112.

Perkin, H. (2007). History of universities. In *International handbook of higher education*, pages 159–205. Springer.

Pianta, E., Girardi, C., and Zanoli, R. (2008). The textpro tool suite. In *Proceedings of LREC*.

Piazza, S. (2013). *La valutazione della ricerca scientifica: Uno studio empirico nelle Scienze umane*. PhD thesis, University of Bologna.

Pickering, A. (1992). *Science as practice and culture*. University of Chicago Press.

Pitt, L., Berthon, P., and Berthon, J.-P. (1999). Changing channels: the impact of the internet on distribution strategy. *Business Horizons*, 42(2):19–28.

Pittinsky, M. S. (2003). *The wired tower: Perspectives on the impact of the internet on higher education*. FT Press.

Posner, M. (2015). Humanities data: A necessary contradiction. Personal website: `http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/`.

Rafols, I. and Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2):263–287.

Ramage, D. (2011). *Studying People, Organizations, and the Web with Statistical Text Models*. PhD thesis, Stanford University.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 248–256.

Ramage, D., Manning, C. D., and Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 457–465. ACM.

Rao, D., McNamee, P., and Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

Repko, A. F. (2008). Defining interdisciplinary studies. *Interdisciplinary Research: Process and Theory*.

Rhody, L. (2012). Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1):19–35.

Richardson, W. (1999). Historians and educationists: the history of education as a field of study in post-war england part ii: 1972–96. *History of education*, 28(2):109–141.

Rifkin, J. (2011). *The third industrial revolution.* Palgrave MacMillan.

Ritze, D., Zirn, C., Greenstreet, C., Eckert, K., and Ponzetto, S. P. (2009). Named entities in court: The marinelives corpus. In *Language Resources and Technologies for Processing and Linking Historical Documents and Archives-Deploying Linked Open Data in Cultural Heritage–LRT4HDA Workshop Programme*, page 26.

Robertson, S. (2014a). Chnm's histories: Digital history & teaching history. Personal website: `http://drstephenrobertson.com/blog-post/digital-history-teaching-history/`.

Robertson, S. (2014b). The differences between digital history and digital humanities. Personal website: `http://drstephenrobertson.com/blog-post/the-differences-between-digital-history-and-digital-humanities`.

Robertson, S. (2016). *The Differences between Digital Humanities and Digital History*, pages 289–307. University of Minnesota Press.

Rockwell, G. (2006). Tapor: Building a portal for text analysis. *Mind Technologies; Humanities Computing and the Canadian Academic Community*, pages 285–299.

Romagnoli, G. (18/08/1994). Noi, i sovversivi del computer. *La Stampa.*

Romano, A. (2007). *Gli statuti universitari: tradizione dei testi e valenze politiche.* CLUEB.

Rorty, R. (1992). *The linguistic turn: Essays in philosophical method.* University of Chicago Press.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494.

Rosenzweig, R. (2003). Scarcity or abundance? preserving the past in a digital era. *The American Historical Review*, 108(3):735–762.

Rothman, J. (2014). An attempt to discover the laws of literature. *The New Yorker.*

Rüegg, W. (2004). *A history of the university in Europe: Volume 3, universities in the nineteenth and early twentieth centuries (1800–1945)*. Cambridge University Press.

Rüegg, W. (2011). *A history of the university in Europe: Volume 4, Universities Since 1945*. Cambridge University Press.

Rydberg-Cox, J. (2005). *Digital libraries and the challenges of digital humanities*. Elsevier.

Said, E. (1979). Orientalism. 1978. *New York: Vintage*, 1994.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Salustri, S. (2010). *Un ateneo in camicia nera. L'Università di Bologna negli anni del fascismo*. Carocci.

Samar, T., Huurdeman, H. C., Ben-David, A., Kamps, J., and de Vries, A. (2014). Uncovering the unarchived web. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1199–1202. ACM.

Scheinfeldt, T. (2016). Sunset for ideology, sunrise for methodology? *Debates in Digital Humanities 2016*.

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1):49–65.

Schmidt, B. M. (2016). Do digital humanists need to understand algorithms? *Debates in Digital Humanities 2016*.

Schreibman, S., Siemens, R., and Unsworth, J. (2004). *A companion to digital humanities*. John Wiley & Sons.

Schulz, K. (2011). What is distant reading. *The New York Times*, 24.

Scott, P., Richards, E., and Martin, B. (1990). Captives of controversy: The myth of the neutral social researcher in contemporary scientific controversies. *Science, Technology & Human Values*, 15(4):474–494.

Seefeldt, D. and Thomas, W. G. (2009). What is digital history? *Perspectives on history*, 47(5).

Serafini, M. (2011). *Technological innovation in Emilia-Romagna: knowledge, practice, strategies*. PhD thesis, University of Bologna.

Shafer, R. J. (1974). *A Guide to Historical Method*. Dorsey Press.

Sharfman, E. C. D. (2015). The Development of the Printing Press and the Decline of the Chronicle as Historical Method. *Inquiries Journal/Student Pulse.*

Sinclair, S., Rockwell, G., et al. (2012). Voyant tools. `http://voyant-tools.org`.

Skouvig, L. (2016). Web-archives and big data: managing the messiness. *Big data - small meaning - global discourses.*

Slapin, J. B. and Proksch, S.-O. (2014). Words as data: Content analysis in legislative studies. *The Oxford Handbook of Legislative Studies*, page 126.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Sprugnoli, R., Tonelli, S., Marchetti, A., and Moretti, G. (2015). Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, page fqv027.

Srinivasan, V. (2016). The internet archive: Bricks and mortar version. *Scientific American.*

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706.

Sugimoto, C. R. and Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation*, 71(4):775–794.

Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities.*

Svensson, P. (2012). Beyond the big tent. *Debates in the Digital Humanities*, 36:49.

Swafford, A. (2015). Problems with the syuzhet package. Personal website: `https://annieswafford.wordpress.com/2015/03/02/syuzhet/`.

Tammaro, A. M. (2006). *Biblioteche digitali in Italia: scenari, utenti, staff e sistemi informativi*. Fondazione Rinascimento digitale.

Thaller, M. (1991). The historical workstation project. *Computers and the Humanities*, 25(2-3):149–162.

Thaller, M. (2012). Controversies around the digital humanities: An agenda. *Historical Social Research/Historische Sozialforschung*, pages 7–23.

Thelwall, M. and Vaughan, L. (2004). A fair history of the web? examining country balance in the internet archive. *Library & information science research*, 26(2):162–176.

Thomas, W. (2004). Computing and the historical imagination. *A companion to digital humanities*, pages 56–68.

Traub, M. C. and van Ossenbruggen, J., editors (2015). *Proceedings of the Workshop on Tool Criticism in the Digital Humanities*.

Trevelyan, G. M. (1927). *The Present Position of History: An Inaugural Lecture*. Longmans, Green.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Turkel, W. J. (2008). Towards a computational history. Personal website: `http://digitalhistoryhacks.blogspot.de/2008/07/towards-computational-history.html`.

Underwood, T. (2012). Topic modeling made just simple enough. Personal Website: `https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/`.

Valentini, A. (04/05/2011). Il pioniere del web che spalancò all'italia le vie del cyberspazio. *Il Tirreno*.

Van Raan, A. (1997). Scientometrics: State-of-the-art. *Scientometrics*, 38(1):205–218.

Vignocchi, M., Bergamin, G., and Messuti, R. (2010). Tesi di dottorato: stato dell'arte, iniziative in corso, scenari possibili. *Bibliotime*, 13(3).

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., and Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature. *Journal of Informetrics*, 5(1):14–26.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM.

Wang, X. and McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433.

Weber, M. (1905). *The Protestant Ethic and the Spirit of Capitalism: and other writings.* Penguin.

Weber, M. S. (2014). Observing the web by understanding the past: Archival internet research. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 1031–1036. International World Wide Web Conferences Steering Committee.

Webster, P. (2015). Will historians of the future be able to study twitter? Personal website: `https://peterwebster.me/2015/03/06/future-historians-and-twitter/`.

Weil, V. (2002). Making sense of scientists' responsibilities at the interface of science and society. *Science and Engineering Ethics*, 8(2):223–227.

Weingart, S. (2011). Topic modeling and network analysis. Personal website: `http://www.scottbot.net/HIAL`.

Weingart, S. (2012). Topic modeling for humanists: A guided tour. Personal Website: `http://www.scottbot.net/HIAL/index.html@p=19113.html`.

Wilkens, M. (2013). The geographic imagination of civil war-era american fiction. *American Literary History*, 25(4):803–840.

Witten, I. and Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30.

Worboys, M. (2011). Practice and the science of medicine in the nineteenth century. *Isis*, 102(1):109–115.

Wrigley, E. A. (1973). *Identifying people in the past.* Edward Arnold London.

Wyatt, S., Milojević, S., Park, H. W., and Leydesdorff, L. (2015). Quantitative and qualitative sts: The intellectual and practical contributions of scientometrics. *Available at SSRN 2588336.*

Yang, T.-I., Torget, A. J., and Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104. Association for Computational Linguistics.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23(3):327–343.

Zhai, C. and Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410. ACM.

Zimmer, M. (2015). The twitter archive at the library of congress: Challenges for information practice and information policy. *First Monday*, 20(7).

Zirn, C. and Stuckenschmidt, H. (2014). Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, 90:38–53.