# Rediscovering the University of Bologna's digital past

Federico Nanni

PhD Student in Science, Technology and Society

University of Bologna

# My background

I studied **contemporary history** and **digital humanities**.

My main research interests are in the areas of **web historiography** and **computational history**.

# Born digital sources

As historians we have to:

- understand **how to deal** with born digital sources;

- **employ** them in order to:

  - answer relevant research questions;

  - imagine new ones.

# The web as a primary source: a necessity

"**Historians** could be **completely left out of post-1996 research**, letting everybody else do the work".

(Ian Milligan, at IIPC 2015)

## The web as a primary source: a duty

"The variety of historical evidence is nearly infinite. **Everything** that man says or writes, everything that he makes, everything he touches **can and ought to teach us about him**."

(Marc Bloch, 1949).

# Historical information retrieval in the digital age

Born digital documents bring with them two big issues
(Rosenzwieg, 2003):

- They are really **difficult to preserve** in their integrity (how do I find a website from '95?);

- They are **too many** (how can I find and study all the tweets related to a specific event?).

## How to deal with the scarcity of sources

In other fields researchers have been dealing with this issue for several years, for example:

- **Digital archivists** (Kahle, 1997; Schneider and Foot, 2004; Brügger, 2005);

- **Media studies researchers** (Brügger, 2008; Dougherty et al., 2010; Ankerson, 2012).

## How to deal with the scarcity of sources

Since November 1996 the **Internet Archive**, a San Francisco based digital library, has been preserving our digital past.



In 2001 the **Wayback Machine** was presented: a online platform for visualising snapshots of preserved websites and retrieving them with a **URL search tool**.

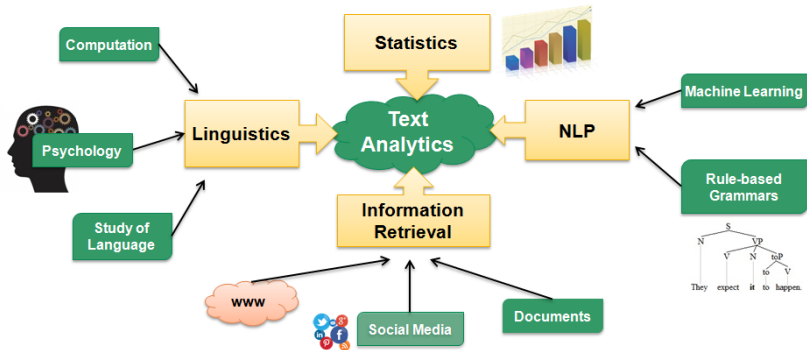## How to deal with the scarcity of sources

- **National libraries** have been preserving their national web sphere (UK, 2004; Denmak, 2005; France, 2006; Portugal, 2007);

- Since 2011 the **IIPC** (International Internet Preservation Consortium) has brought together archivists and researchers.

# How to deal with the abundance of sources

In other fields researchers have been dealing with this issue for several years, for example:

- Computational linguists (Manning and Schütze, 1999);

- Computer scientists (Blei et. al, 2003).

# How to deal with the abundance of sources



Image from: http://www.meaningcloud.com/blog/voice-of-the-customer-banking-industry/

## How to deal with the abundance of sources

- Be involved in the **NLP** and **Web Science** communities;

- Participate at their **conferences** (ACL, WWW, Web Science, ISWC);

- Receive **feedbacks** and establish **collaborations**.

# Doing Digital Humanities

This **new historical method** is the perfect example of a digital humanities practice:

- we use **computational technologies** (web archives, text mining methods...);

- **to deal with specific methodological challenges** (related to the intrinsic nature of born digital sources);

- in order to study **different humanities topics**.

# My research

Since October 2013 I've been conducting my research at CIS, the
**Centre for the History of Universities and Science**.

# Alma mater studiorum

The University of Bologna is considered to be the **world's oldest university** (established around 1088).

**Several different sources** have been used to **study its past** and the relationship between this institution, its large students community and the city of Bologna itself.

# My research goal

Born digital documents such as materials offered on **Unibo.it** or shared on **AlmaDL** could give historians new insight.

The aim of my PhD work is to offer a **different perspective** on the recent history of universities.

# Scarcity of sources

Reconstructing the digital past of the University of Bologna is particularly complicated, especially because:

- Unibo.it has been **completely redesigned** between 2002 and 2006 and the majority of the materials previously offered online are not available anymore;

- Italy <u>doesn't have</u> a national web archive;

- The University of Bologna **has been excluded** from the Internet Archive's **Wayback Machine**.

# No kidding

INTERNET ARCHIVE

**WayBackMachine**

| http://www.unibo.it/ |

Latest

Show All

## Sorry.

This URL has been excluded from the Wayback Machine.

# Specific research questions

- How can we retrieve and analyse these sources?

- Do they give us new insight?

- How could we change the way the historical method is taught?

# Different approaches

In order to face this issue and reconstruct the digital past of the University of Bologna I followed two paths:

- I collected information **from other sources** (Brügger, 2008);

- I **studied** and (almost) **solved** this exclusion-issue.

# Oral histories

In the last 20 years **four different teams** supervised the website:

- 2015 - 2011: CeSIA + AAGG.

- 2011 - 2002: Luca Garlaschelli - Portale d'Ateneo Project.

- 2002 - 1994: CeSIA - Salvatore Mirabella (URP).

- 1994 - 1993: Ozalp Babaoglu - Renzo Davoli (Department of Mathematics).

I worked in close contact with Mauro Amico, who has been at CeSIA since the early 2000.

# Finding documents in archives

No backups of the old versions of Unibo.it have been preserved by the teams who managed the platform.

- Mauro Amico gave me **seven .png images** which covered the most important layout changes between 1998 and 2009;

- One snapshot (2006) is available in the **Internet Memory Foundation** collections;

- **Netarkivet** archived a few time the Unibo website (2006 - 2012).

# Unibo in the Internet Memory Foundation

# Unibo in Netarkivet

## Finding documents in archives

- **Unibo sub domains** are available on the Wayback Machine;

- **The English version** of the website is available on the Wayback Machine (2004 - 2014);

- In **2007** during a student protest **the website was cloned** (unibologna.eu). This version is available on the Wayback Machine.

# Cloned version

## **Newspapers and other historical sources**

- Articles on Unibo.it in the digital archives of **Repubblica.it** (1996 - 2015) and of the local newspaper **Il Resto del Carlino** (1999 - 2015);

- News in **university digital magazines**: AlmaNews (1997 - 2013), Alma2000 (2000 - 2002) and AlmaMagazine (2002 - 2015);

- Pieces of information in **student forums** and in **Usenet discussion groups**;

- Cilea **list of web servers** (1997) and NCSA what's new! (1994).

# Summarizing
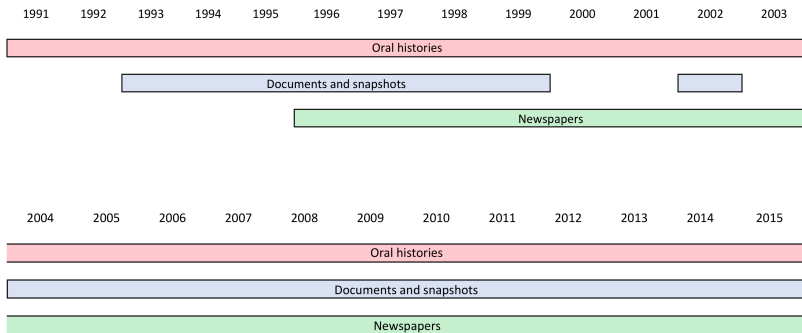
# The exclusion issue

In the FAQ section the Internet Archive clarified that a website is excluded if:

- It stops the IA crawlers by changing the **robot.txt file**;

- A website owner sends a **specific removal request**.

# Different messages

INTERNET ARCHIVE
WayBackMachine

http://portale.unitn.it/

Latest
Show All

Page cannot be crawled or displayed due to robots.txt.

See portale.unitn.it robots.txt page. Learn more about robots.txt.

INTERNET ARCHIVE
WayBackMachine

http://www.unibo.it

Latest
Show All

**Sorry.**

This URL has been excluded from the Wayback Machine.

# Getting in contact
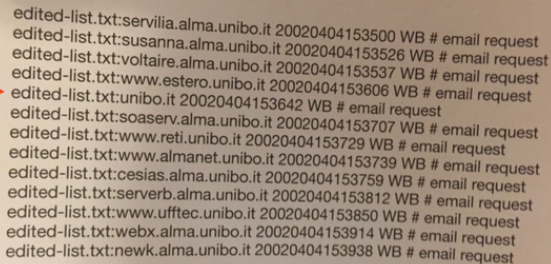
The Internet Archive has a specific email address for contacting them regarding this issues: **info@archive.org**.

**Between October 2013 and February 2015** I tried several times to get in touch with them, both with the help of Cesia and Netarkivet.

However, until last March **we didn't receive any answer**.

# Working together

Thanks to the help of Raffaele Messuti (Unibo - AlmaDL) and Giovanni Damiola (Internet Archive) we discovered that a **specific request** was sent in **2002**.

# Working together

Thanks to the help of Mauro Amico (Unibo - CeSIA) and Chris Butler (Internet Archive) since the **13th of April** the website has been available again on the Wayback Machine.
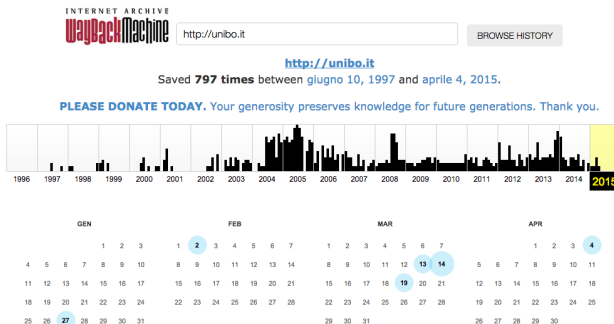
# From web history to the history of institutions

1) Discovering the **reasons for the removal request**.

2) Remarking how the **website** could become a **relevant source** for the history of academic institutions.

3) Collaborating with Renzo Davoli and Ozalp Babaoglu to know more about their role in the **arrival of the Internet and the WWW** at the University of Bologna.

4) Working on the **preservation of the Italian web sphere**.

# Abundance of sources

# Abundance of sources

The Alma Mater Digital Library of the University of Bologna offers all the **PhD Theses** published during the last 9 years (2007-2015).

This kind of corpus could be really useful for researchers interested in understanding **what has been studied** at a specific academic institution (Ramage et al., 2011).

# Big historical data

Over 3000 PhD dissertations.
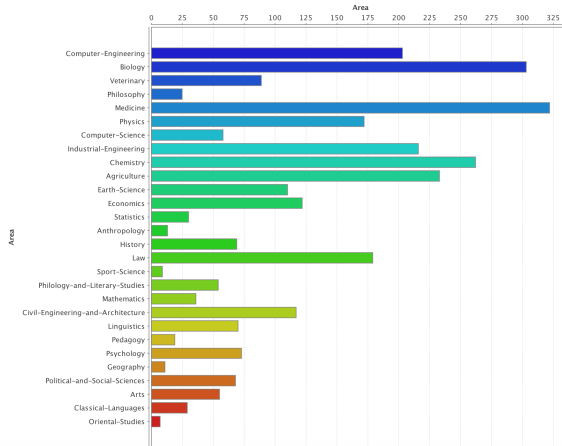


Labeled according to the Italian Scientific and Disciplinary Sectors.

# Mining PhD theses

- Identify the **most recurrent research topics**;

- **Detect interdisciplinary works**, which could be representative of collaborations between departments;

- **Diachronically analyse** these documents;

- Compare the **scientific production** of different **Italian universities**.

# Interdisciplinarity

"Interdisciplinarity is 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, 2012).

# Rocchio classification

In order to detect **interdisciplinary practices** we employed the Rocchio classification method.

For each discipline a **centroid**, which represents the center of mass of its members, is created.

We calculated the **similarity** between each thesis and all the centroids (cosine similarity, leave-one-out cross validation).

# How did it go?

1) Rocchio classification could be a useful way of conducting a first exploration of the corpus;

2) it detects **similarity between strings** and ignores relations between words (it's the classic example of a **bag-of-words** approach);

3) it is **not the most efficient way** of identifying **discipline-topics**.

# Brief introduction to topic models

Techniques for analyzing relationships between terms, documents and concepts became famous during the Nineties under the name of Latent Semantic Analysis.

Starting from them, Blei et al. in 2003 proposed a new method, called **Latent Dirichlet allocation** (LDA).

In LDA each document is seen as a mixture of topics and each topic is a mixture of words.

# Topic models in digital humanities

This technique and other **distant reading methods** are more and more applied in Digital Humanities.
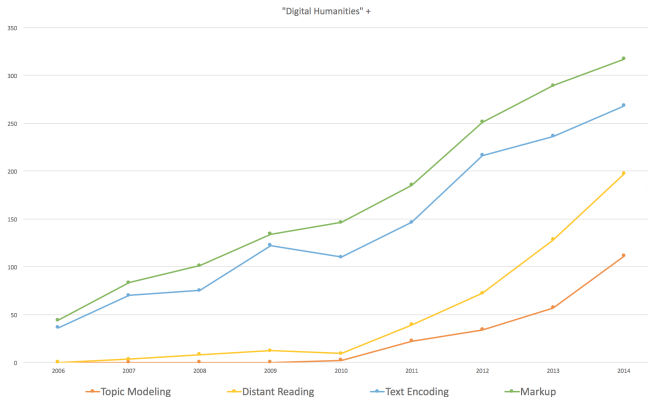


By the number of results per year on Google Scholar.

# Applying topic modeling

The process of doing topic modeling is divided in different tasks:

- determine the number of topics;

- train the model on a specific corpus corpus;

- apply the model (on this corpus or on a different one).

# Issues with topic modeling

The process of performing topic modeling presents different issues:

- we have to select a number of topics;

- we cannot define the topics in advance;

- results are not always understandable (Chang, 2009; Lau, 2014).

# Results

This is how the theses behave according to one of the most important topics in Agriculture:



Topic words: fruit plant quality stress phenolic milk farm fatty_acid breeding

## Results

This is how the theses behave according to one of the most
important topics in Agriculture:



Topic words: fruit plant quality stress phenolic milk farm fatty_acid
breeding

# Results

**Agricultural Genetics and Plant Breeding in Early Twentieth-Century Italy**

Documenti full–text disponibili:

**Abstract**

This thesis is about plant breeding in Early 20th–Century Italy. The stories of the two most prominent Italian plant–breeders of the time, Nazareno Strampelli and Francesco Todaro, are used to explore a fragment of the often–neglected history of Italian agricultural research. While Italy was not at the forefront of agricultural innovation, research programs aimed at varietal innovation did emerge in the country, along with an early diffusion of Mendelism. Using philosophical as well as historical analysis, plant breeding is analysed throughout this thesis as a process: a sequence of steps that lays on practical skills and theoretical assumptions, acting on various elements of production. Systematic plant–breeding programs in Italy started from small individual efforts, attracting more and more resources until they became a crucial part of the fascist regime's infamous agricultural policy. Hybrid varieties developed in the early 20th century survived World War II and are now ancestors of the varieties that are still cultivated today. Despite this relevance, the history of Italian wheat hybrids is today largely forgotten: this thesis is an effort to re–evaluate a part of it. The research did allow previously unknown or neglected facts to emerge, giving a new perspective on the infamous alliance between plant–breeding programs and the fascist regime. This thesis undertakes an analysis of I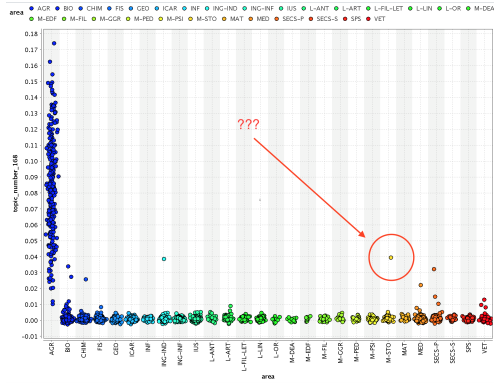talian plant–breeding programs as processes. Those processes had a practical as well as a theoretical side, and involved various elements of production. Although a complete history of Italian plant breeding still remains to be written, the Italian case can now be considered along with the other case–studies that other scholars have developed in the history of plant breeding. The hope is that this historical and philosophical analysis will contribute to the on–going effort to understand the history of plants.

| | |
|---|---|
| **Tipologia del documento:** | Tesi di dottorato |
| **Autore:** | Iori, Luca |
| **Relatore:** | Pancaldi, Giuliano |
| **Correlatore:** | Leonelli, Sabina |
| **Dottorato di ricerca:** | Science, technology, and humanities |
| **Scuola di dottorato:** | Scienze umanistiche |
| **Ciclo:** | 25 |
| **Coordinatore:** | Pancaldi, Giuliano |
| **Settore disciplinare:** | Area 11 – Scienze storiche, filosofiche, pedagogiche e psicologiche > M–STO/05 Storia della scienza e delle tecniche |
| **Settore concorsuale:** | Area 11 – Scienze storiche, filosofiche, pedagogiche, psicologiche > 11/C – Filosofia > 11/C2 Logica, storia e filosofia della scienza |

# Results

This is how the theses behave according to one of the most important topics in Biology:
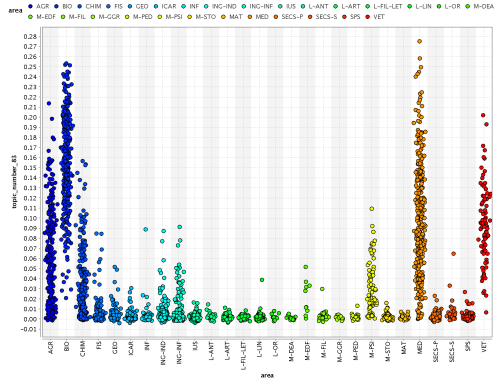


Topic words: cell gene expression protein treatment role effect increase human

# Results

This is how the theses behave according to one of the most important topics in Biology:
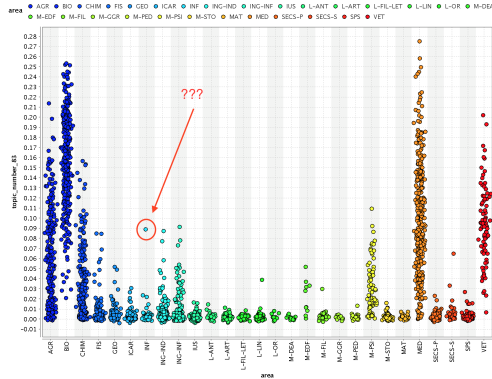


Topic words: cell gene expression protein treatment role effect increase human

# Results

**Investigating the role of single point mutations in the human proteome: a computational study**

Tiwari, Shalinee (2011) *Investigating the role of single point mutations in the human proteome: a computational study*, [Dissertation thesis], Alma Mater Studiorum Università di Bologna. Dottorato di ricerca in Informatica, 23 Ciclo. DOI 10.6092/unibo/amsdottorato/3363.

Documenti full-text disponibili:

Documento PDF (English) – Richiede un lettore di PDF come Xpdf o Adobe Acrobat Reader
2734Kb

### Abstract

In the post genomic era with the massive production of biological data the understanding of factors affecting protein stability is one of the most important and challenging tasks for highlighting the role of mutations in relation to human maladies. The problem is at the basis of what is referred to as molecular medicine with the underlying idea that pathologies can be detailed at a molecular level. To this purpose scientific efforts focus on characterising mutations that hamper protein functions and by these affect biological processes at the basis of cell physiology. New techniques have been developed with the aim of detailing single nucleotide polymorphisms (SNPs) at large in all the human chromosomes and by this information in specific databases are exponentially increasing. Eventually mutations that can be found at the DNA level, when occurring in transcribed regions may then lead to mutated proteins and this can be a serious medical problem, largely affecting the phenotype. Bioinformatics tools are urgently needed to cope with the flood of genomic data stored in database and in order to analyse the role of SNPs at the protein level. In principle several experimental and theoretical observations are suggesting that protein stability in the solvent–protein space is responsible of the correct protein functioning. Then mutations that are found disease related during DNA analysis are often assumed to perturb protein stability as well. However so far no extensive analysis at the proteome level has investigated whether this is the case. Also computationally methods have been developed to infer whether a mutation is disease related and independently whether it affects protein stability. Therefore whether the perturbation of protein stability is related to what it is routinely referred to as a disease is still a big question mark. In this work 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. To this aim in the first part of the thesis for each mutation type we have derived two probabilistic indices (for 141 out of 150 possible SNPs): the perturbing index (Pp), which indicates the probability that a given mutation effects protein stability considering all the "in vitro" thermodynamic data available and the disease index (Pd), which indicates the probability of a mutation to be disease related, given all the mutations that have been clinically associated so far. We find with a robust statistics that the two indexes correlate with the exception of all the mutations that are somatic cancer related. By this each mutation of the 150 can be coded by two values that allow a direct comparison with data base information. Furthermore we also implement computational methods that starting from the protein structure is suited to predict the effect of a mutation on protein stability and find that overpasses a set of other predictors performing the same task. The predictor is based on support vector machines and takes as input protein tertiary structures. We show that the predicted data will correlate with the data from the databases. All our efforts therefore add to the SNP annotation process and more importantly found the relationship among protein stability perturbation and the human variome leading to the diseasome.

| | |
|---|---|
| Tipologia del documento: | Tesi di dottorato |
| Autore: | Tiwari, Shalinee |
| Relatore: | **Casadio, Rita** |
| Dottorato di ricerca: | Informatica |
| Scuola di dottorato: | Scienze e ingegneria dell'informazione |
| Ciclo: | 23 |
| Coordinatore: | **Martini, Simone** |
| Settore disciplinare: | Area 01 – Scienze matematiche e informatiche > INF/01 Informatica |
| Settore concorsuale: | Area 01 – Scienze matematiche e informatiche > 01/8 – Informatica > 01/B1 Informatica |
| Parole chiave: | perturbing index disease index disease class index disease related mutations correlation coefficients protein function SNP annotation |

# Interdisciplinarity detection

1) Evaluate our method for interdisciplinarity detection.

2) Create a graphic interface for the tool.

3) Use it to answer specific research questions.

# The historian today

We are dealing with **a new kind of primary sources**.

Working with digital archivists and computer scientists is **not an option anymore**: it's the only suitable way for studying our past.

# The historian tomorrow

Will the combination of these different backgrounds be the key to
**imagine new research questions**?

## **Thanks.**

Federico Nanni
PhD Student in Science, Technology and Society
University of Bologna
federico.nanni8@unibo.it