Collaborative Research Practices and Shared Infrastructures for Humanities Computing

2nd AIUCD Annual Conference, AIUCD 2013
Padua, Italy, 11-12 December 2013

Proceedings of Revised Papers

Maristella Agosti and Francesca Tomasi (Eds)

cleup

http://www.umanisticadigitale.it

Prima edizione: settembre 2014

ISBN 978 88 6787 260 2

Cleup sc
"Coop. Libraria Editrice Università di Padova"
via G. Belzoni 118/3 – Padova (t. 049 8753496)
www.cleup.it - www.facebook.com/cleup

In copertina:
Graphic Design: Massimo Malaguti – Scuola Italiana Design (elaborazione del logo di Aiucd).

Organization

AIUCD 2013 was organized by the Information Management Group of the Department of Information Engineering of the University of Padua, Italy.

Committees

General Chair
Dino Buzzetti, Presidente AIUCD

Program Chairs
Maristella Agosti, Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Padova
Anna Maria Tammaro, Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Parma

Program Committee
Fabio Ciotti, Dipartimento Studi Umanistici, Università di Roma Tor Vergata
Giorgio Maria Di Nunzio, Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Padova
Maurizio Lana, Dipartimento di Studi Umanistici, Università del Piemonte Orientale
Federico Meschini, Dipartimento di Scienze Umanistiche, della Comunicazione e del Turismo, Università degli Studi della Tuscia
Nicola Orio, Dipartimento di Beni Culturali, Università degli Studi di Padova
Nicola Palazzolo, già ordinario nell'Università di Perugia

Roberto Rosselli Del Turco, Dipartimento di Studi Umanistici, Università di Torino
Marco Rufino, Fondazione Rinascimento Digitale, Firenze
Francesca Tomasi, Dipartimento di Filologia Classica e Italianistica, Università di Bologna

Award Chair
Francesca Tomasi, Dipartimento di Filologia Classica e Italianistica, Università di Bologna

Local Committee
Debora Leoncini, Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Padova
Marta Manfioletti, Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Padova
Chiara Ponchia, Dipartimento di Beni Culturali, Università degli Studi di Padova
Gianmaria Silvello, Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Padova

Table of contents

PANELS

PAPERS

Preface

The Italian Association for Digital Humanities[1] (AIUCD) was launched in 2011 to promote and disseminate methods and facilitate scientific collaboration and development of useful resources in the field of digital humanities in Italy. AIUCD is an associate organization of the European Association for Digital Humanities[2] (EADH), which brings together and represents the Digital Humanities in Europe across the entire spectrum of disciplines that research, develop, and apply digital humanities methods and technology. AIUCD is thereby represented in the Alliance of Digital Humanities Organizations[3] (ADHO) which promotes and supports digital research and teaching across all arts and humanities disciplines, acting as a community-based advisory force, and supporting world-wide excellence in research, publication, collaboration and training.

We are very pleased to present the volume of the proceedings of the 2nd Annual Conference of the Italian Association for Digital Humanities (AIUCD 2013) on "Collaborative Research Practices and Shared Infrastructures for Humanities Computing"[4], which took place at the Department of Information Engineering of the University of Padua, 11-12 December 2013. The first conference was held in Florence, 13-14 December 2012, for "An Agenda for Humanities Computing and Digital Culture"[5]. The third conference

[1] Associazione per l'Informatica Umanistica e la Cultura Digitale http://www.umanisticadigitale.it/
[2] http://www.eadh.org/
[3] http://adho.org/
[4] http://aiucd2013.dei.unipd.it/
[5] http://www.umanisticadigitale.it/2012/12/

will be held in Bologna, 18-19 September 2014 on "Humanities and Their Methods in the Digital Ecosystem"[6]. This means that AIUCD is working hard towards building a strong interdisciplinary community of researchers in the field of Digital Humanities, and it is working to form a substantial body of scholarly publications contained in the conference proceedings which contribute to representing the different facets of the sector that are active in Italy. Like all areas with a strong interdisciplinary basis, the scientific area of interest of the association and of the conference is struggling to be recognized. We hope that this volume contributes not only to the presentation of the significant aspects of the area but also to the spread of knowledge and broadening of its recognition.

The general theme of AIUCD 2013 was "Collaborative Research Practices and Shared Infrastructures for Humanities Computing" so we particularly welcomed submissions on interdisciplinary work and new developments in the field, encouraging proposals relating to the theme of the conference, or more specifically: interdisciplinarity and multidisciplinarity, legal and economic issues, tools and collaborative methodologies, measurement and impact of collaborative methodologies, sharing and collaboration methods and approaches, cultural institutions and collaborative facilities, infrastructures and digital libraries as collaborative environments, data resources and technologies sharing.

The call showed a large interest on these topics from both the computer science and the humanistic communities: 29 abstracts were submitted and subsequently evaluated through a peer-review process. Members of the Program Committee as well as of the Local Committee were given the task to evaluate them. Eventually 14 submissions were accepted, and organized in 3 sessions during the second day of the conference. About 70 people physically attended the conference.

The first day the conference started with the keynote address entitled "Toward a Computational Narratology" by Jan Christoph Meister, University of Hamburg, Germany. The keynote address argues for a methodological positioning of Digital Humanities research as one that necessarily combines phenomenological with formal conceptualizations of its object domain. After the keynote a session devoted to the presentation of invited contributions from Italian research groups and centers was held and chaired by Fabio Ciotti. The first day was completed by a panel on "Digital Resources and

[6] http://aiucd2014.unibo.it/

Network Services for the Digital Humanities Research" chaired by Dino Buzzetti.

The accepted scientific submissions were presented on the second day and they were organized into three sessions respectively chaired by Maurizio Lana, Roberto Rosselli Del Turco and Francesca Tomasi. All submitted abstracts for which the primary author was a Master or PhD student were eligible for the "AIUCD 2013 Best Student Abstract Award"; this award was introduced for the first time at this conference. The prize was awarded to Chiara Ponchia – a PhD student at the University of Padua at the time of the conference – for the contribution "Dante. A web-application for the History of Art". The day was topped off by a panel on "The Digital Library to Support the Computer Humanist" chaired by Anna Maria Tammaro. The AIUCD business meeting was held immediately after lunch and before the scientific sessions of the afternoon. Summing up a very dense second day of conference!

The proceedings reflect the structure of the conference with some reorganizational components. After the keynote contribution the session devoted to invited speakers is presented. The aim of the works in this session is to describe Italian projects and activities managed by some centers and groups involved in the digital humanities domain. Two panels follow on digital resources in the humanities and digital libraries. The papers session – i.e. the selection of submitted papers after the peer review process – is then organized on three macro-topics: digital philology, digital cultural heritage and education. Finally, the language of the proceedings is English, although some contributions are in Italian.

The conference took place in the Aula Magna "Antonio Lepschy" of the Department of Information Engineering of the University of Padua. Thanks to the financial support of the Department the participation and mutual knowledge of the participants were encouraged, offering them the opportunity of sharing the moments of rest between work sessions.

The success of AIUCD 2013 would not have been possible without the invaluable contributions of all members of the Program Committee, Local Committee and the technical and administrative staff of the Department of Information Engineering that supported the conference in its various stages.

July 2014

Maristella Agosti, *Francesca Tomasi*

Keynote

Keynote Address / Intervento invitato

Toward a Computational Narratology*

Jan Christoph Meister
Institut für Germanistik, Universität Hamburg, Germany
jan-c-meister@uni-hamburg.de

> The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. (…).
> Alan Turing, *Computing machinery and intelligence* (1950, 458)

Abstract: This paper argues for a methodological positioning of Digital Humanities research as one that necessarily combines phenomenological with formal conceptualizations of its object domain. One of the pre-computational humanities theories based on a similar methodological mix is that of the formalist approach in narrative analysis initiated by Propp in 1928. His model of so-called 'narrative functions' remains highly influential in present day narratological as well as in Artificial Intelligence research into narrative logic, but has often been misinterpreted as a (failed) attempt to approach narrative in a purely formal fashion. Against this backdrop, *heureCLÉA* – a current research project in Computational Narratology – presents a concrete example for how these two approaches can today complement one another by combining manual (computer aided) text annotation with a machine learning based investigation of annotation regularities. Once identified by this 'digital heuristic', such regularities can then be fed back as probabilistic annotation suggestions presented to the human annotator.

Keywords: Digital Humanities, digital heuristic, text annotation.

1. Introduction

The formation of a new academic discipline as well as the establishing of a new scientific methodology has occasionally been traced back to the

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. Cleup, Padova, 2014.

discovery of a single new tool or instrument. Early 17th century invention of the telescope and its importance for the development of a heliocentric and rationalist world view is often used to illustrate the unplanned progression from technological feat to epistemological shift in paradigm. In a historical perspective the technology inspired formation of these and other epistemological and philosophical transformations prove to have been contingent developments – hardly anybody has ever invented an instrument or a tool with the intention to establish a new conceptual paradigm. Tools get invented to do a job and not in order to change how we think about the world.

By contrast, the mystique which some concerned humanists ascribe to "the computer" as something that apparently threatens to obliterate our traditional hermeneutic approaches altogether is oddly ahistorical. Rather than engaging in a critical appraisal of the instrument's development, its increasing capabilities and its conceptual limitations, the polemic tries to capitalize on rhetorical blunder. For talk of "the computer" and the dangers of "the digital" does not refer to a particular type of machine; rather it is a personification: "the computer" represents the marauding *Golem* of mathematical formalization let loose on the human emotional and cognitive realm; in short: on the traditional object domain of the humanities.

But what exactly *could* be wrong with formalisation from a traditionalist's point of view; in which sense might this choice of methodology indeed threaten "the human", i.e. our sense of individuality, creativity, emotional, ethical, intellectual and aesthetic engagement with the world and our fellow beings? Indeed, the success of present day statistical algorithms provides for some creepy experiences that go much further than recommender systems (Amazon) and data mining (Google). Take for example the recent undermining of so-called CAPTCHA code as a way to distinguish between the human user of a web form, and a bot. By training their algorithms on vast amounts of human generated decodings of nearly illegibly wavy, distorted

> eCAPTCHA sequences of characters and numbers, Google's researchers have just claimed to have developed a neural-network based algorithm that can automatically 'decipher' such an image with over 99% accuracy. As Goodfellow *et al.* (2014) report, eCAPTCHA is one of the most secure reverse turing tests that uses distorted text as one of the cues to distinguish humans from bots. With the proposed approach we report a 99.8% accuracy on transcribing the hardest category of reCAPTCHA puzzles. Our evaluations on both tasks, the street number recognition as well as reCAPTCHA puzzle transcription, indicate that at specific

> operating thresholds, the performance of the proposed system is comparable to, and in some cases exceeds, that of human operators.

Note that 'decipher' is of course also a metaphor – after all, this is pure statistics and number crunching, not intelligence at work. The algorithm isn't intelligent, but the researchers who developed it certainly are – and so is the amused blogger who commented on the related report in the German newspaper *Spiegel Online* with the suggestion of a *reverse* test for identifying a true human user: surely, he argued, if an agent takes *more* than three attempts to identify an illegible CAPTCHA code image correctly, that dumb agent must be a human![1] In an intellectual and economic climate dominated by technology and the natural sciences such irony sometimes seems to be the humanities last resolve. As humanists we somehow missed the boat, and it must have happened to us sometime around mid-20th century, when the first modern computers were developed, if not even before…

The fundamental decoupling of the two intellectual paradigms, that of the humanities and that of natural sciences and technology, which had gone hand-in-hand since the Renaissance was prominently criticized by C.P. Snow in his often cited 1959 "Reed Lecture". For the sake of polemic effect Snow laid the blame squarely onto the humanist intellectuals of his time, the "men of letters", who would not just openly admit, but indeed proudly brag about their disinterest in the foundational theorems of the natural sciences. To quote:

> A good many times I have been present at gatherings of people who, by the standards of the traditional culture, are thought highly educated and who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the Second Law of Thermodynamics. The response was cold: it was also negative. Yet I was asking something which is the scientific equivalent of: *Have you read a work of Shakespeare's?*
>
> I now believe that if I had asked an even simpler question – such as, *What do you mean by mass, or acceleration*, which is the scientific equivalent of saying, *Can you read?* – not more than one in ten of the highly educated would have felt that I was speaking the same language. So the great edifice of modern physics goes up, and the majority of the cleverest people in the western world have about as

[1] See "Wo ist das Problem?" by humbahumba http://www.spiegel.de/netzwelt/web/google-knackt-den-captcha-code-a-964955.html

> much insight into it as their Neolithic ancestors would have had. (Snow 1993 [1959], 14-15)

Today our common short hand for C.P. Snow's famous 1959 *Reed lecture* is the "Two Cultures" thesis. Indeed, dialogue between the human and the natural sciences leaves much to wish for, and in many instances it is still at best a one-way communication. Heisenberg for example, we may assume, will no doubt have read his Goethe – but Goethe, although keenly interested in natural phenomena, would never have made his peace with the analytical and highly abstract mathematical model of the physical world that has become the foundation of theoretical physics. For this was precisely what Goethe detested in Newtonian physics: the radical analytical abstraction from phenomenal reality (Goethe 1810).

And this is probably the real issue at stake in many traditional humanists' reluctance toward the computational study of cultural symbolic artefacts, including, of course, the computational modelling of narratives. However, in order to understand this reluctance and appreciate its philosophical dimension, we must go further than what by now has become cultural studies folklore: the lament over the "great divide" and the splitting of the "two cultures". One possible approach is to re-think the opposition in methodological terms, namely as that of *phenomenology versus formalization*.

2. Phenomenology vs. Formalization

Phenomenological approaches conceptualize phenomena in terms of what they are to the human experience. Phenomenology is therefore intrinsically indexical – it always points back to the position of an observer who, by design, is an integral methodological point of reference in its models and theories.

By contrast, what does formalization point to? First, of course, its object domain – i.e. that which is being formalized, the entities or processes or ideas that we want to understand better. Second, every formalization, unless it is faulty and incoherent, implicitly confirms the cohesiveness of its own rule set – every "one plus one equals two" also confirms the validity of how, subject of course to an axiomatic logic, the operation of addition is defined. But there is no systemic self-awareness built into this approach. Indeed, formalization hinges on *avoiding* indexality in order to arrive at a truly *context free* model. The phenomenological triple of object, observer and model is

consciously transformed into that of object, theory and model. But the model, it seems, is then no longer pragmatically motivated; the intellectual construct has become self-referential, and in its ideal form may even be elevated to logical or mathematical abstraction.

However, if we dig long and deep enough, even mathematical abstractions can sometimes be traced back to an initial real-world observation. Here is an example:

$$A \approx \left(\frac{8}{9}d\right)^2 = \frac{256}{812}\,r^2 = 3{,}16049\ldots r^2$$

This formula is the abstract representation of a message contained in the so-called *Rhind papyrus* which dates back to around 1800 BC.[2] The papyrus instructs its readers how to calculate the area circumscribed by a circle. Historians of mathematics believe that this insight was arrived at by practical observation, not by mathematical abstraction. The papyrus can be read as a kind of operational sequence: measure a circle's diameter d, then construct a square with a side a that is 8/9th the length of d. If you now map this square – whose coverage is simple to calculate: a^2 – onto your circle you will have covered the same area: the segments of the circle that protrude on each of the four sides will fit perfectly into the four corners of the square which lie outside the circle.

Of course, today we no longer use the Egyptian method, we use the constant ϖ which was identified by Archimedes some 1600 years later. Archimedes conceptualized covering the circle with a polygon, rather than with a simple square. Then he gradually increased the number of sides in the polygon until he had constructed one with 96 sides. On the basis of this mathematical abstraction he then calculated the ratio between diameter and circumference as being 3,14159… Today we have added a couple of billion further sides to Archimedes' polygon without finding any regularity in the numbers behind the comma – and so the business of approximation continues. Be this as it may, for most practical purposes 3,14159 will certainly be sufficiently precise – the real irony being that the same might be said about the method offered by the Rhind papyrus: it already produced a 99,4% ac-

[2] See Chace A.B. 1979 [1927-1929]. *The Rhind Mathematical Papyrus. Free Translation and Commentary with Selected Photographs, Translations, Transliterations and Literal Translations*. Classics in Mathematics Education 8, 2 vols, National Council of Teachers of Mathematics.

curate result. And so an almost 4000 year old observation based practice in solving a complicated mathematical problem turns out to be only 0.4 % less precise than Google's recent, *big data* based statistical and neural network solution to the eCAPTCHA task. One thing is for sure then: the phenomenological approach to reality is anything but obsolete.

What does this comparison between a phenomenological and a mathematical approach in problem solving tell us? We can use ϖ to calculate the numerical relation between a circle's diameter and circumference, or between its radius and coverage – but useful as it is, ϖ remains an irrational number. We can think ϖ and attempt to formalize the correlation numerically, but we cannot relate that formalization without loss to the phenomenal experience of the area contained within a circle. One might say that ϖ is not "real" (meant figuratively, not mathematically) per se, although 'it works'. The circle on the other hand is a real phenomenon that escapes formalization – whether we use our fingers or our computers, in abstraction we can still only approximate it.

3. Toward the ϖ of narrative

Narrative and its formalization present us with a similar conundrum: we experience the phenomenon, the story and yet in formalization, we can only approximate it. In fact, it seems that the closest we can approximate the full phenomenal richness of a real story is not even a formula, a pure relational abstraction, but perhaps a "map" – an iconic rather than a numeric symbolic representation. If that is indeed the case then the best we can hope for would seem to be a 1:1 map of narrative – a whimsical idea entertained by Umberto Eco in *Dell'impossibilità di costruire la carta dell'impero 1 a 1* (Eco 1994). Obviously, there is no point in following that approach. Maps like formulae have to be reductive and selective, it is their *raison d'etre* to abstract in order to facilitate understanding and orientation of the primary object of investigation.

But there is still an important difference between a map and a formula: maps do not aim to generalize; they still relate to a specific, unique object of observation. And that is where the humanist's dilemma arises: many of our theories try to be both, a "map" preserving the phenomenal uniqueness of the human symbolic artefact, and a "formula" that allows us to generalize and abstract from the phenomena in spite of their contingent qualities. For that reason some formal and computational approaches toward

narrative amount to a border-line experience where the researcher lingers and oscillates between the phenomenological and the formal. On the one hand we are indeed after some "ϖ of narrative"; on the other hand it has become obvious that we require a stronger base in the empirical and phenomenological dimension. Neither the traditional study of exemplary texts, nor the structuralist models of narrative that were so *en vogue* in the 1960s and 1970s have been able to answer both demands. The question is: can Digital Humanities?

The methodology of combining the formal with the empirical approach in narrative studies is implied in, among other, the *Call for papers* for the 2012 *Computational Models of Narrative Workshop* in Istanbul. Its organizers, the AI specialists Pablo Gervas and Mark Finlayson state that their own

> ... field has yet to address key needs with regard to shared resources and corpora that could smooth and hasten the way forward. The vast majority of work on narrative uses fewer than four stories to perform their experiments, and rarely re-uses narratives from previous studies. Because NLP (= natural language processing; JCM) technology cannot yet take us all the way to the highly-accurate formal representations of language semantics, this implies significant amounts of repeated work in annotation. The way forward could be catalyzed by carefully constructed shared resources.[3]

Though in this instance "the field" refers to computational modelling of narratives, and not to philological narratology, we can assume that many traditional scholars would agree: shared resources and corpus based approaches in narrative studies are indeed lacking. Digital Humanities and the paradigm of *distant reading* (Moretti 2000) hold a promise in this regard, and in the following I will present an example for a method and a system which we are currently developing in order to support and harness collaborative work in our narratological analysis and modelling of narrative. But let us first clarify what a corpus based approach can achieve, and what it cannot deliver.

One of the foundational myths of formalist studies into narrative concerns Propp's *Morphology of the Folk Tale* (Propp 1928). Propp, it is often claimed, analysed 100 Russian fairy tales in order to extract his famous 31 element chain of so-called "functions." And so many who refer to his *Morphology* believe that not only did he contribute the ground breaking idea to

[3] http://narrative.csail.mit.edu/ws12/

formalize the surface structure of a narrative's action in terms of a stringent functional deep structure, but that he also introduced a corpus based, inductive approach into the study of narrative.

However, attempts to re-run the Proppian experiment have raised doubt as far as its methodological rigour is concerned. Claude Bremond and Jean Verrier have shown that a number of tales in the Proppian corpus will in fact not match his sequential formula unless rewritten in a specific way (Bremond & Verrier 1984, 177-195). Did Propp perhaps 'massage the data'? Did he really analyse the entire corpus of 100 tales in a bottom-up mode before he presented his conclusion? It seems more likely that his 31 function formula was initially retrieved from a sub-set of about ten tales, and only then put to the test with the remainder.

Indeed, Propp would have had good reason not to apply stringent empiricism. From the German Romantic period onward and throughout the 19th century, philologists had already compiled various corpora of folk tales, of which the collection of the Grimm brothers is still the most famous. This tradition of research culminated in 1910 when the Finnish folklore scholar Annti Aarne merged the Grimm's and two Scandinavian collections and extracted from them a taxonomy of 2000 elementary motifs (Aarne 1910). His follower Stith Thompson expanded the taxonomy further, complementing it with 500 so-called 'tale-types' (Aarne & Thompson 1961), to which Hans-Jörg Uther finally added another 250 before he re-published the catalogue for a third time, now under the title *The Types of International Folktales* (Uther 2004).[4]

Over the past 100 years empirical folklore studies have progressed from a regional corpus to an international one, and have built a taxonomy of some three thousand distinct motifs and tale-types. However, the vast majority of these entries were already catalogued when Propp published his work in 1928. And so in terms of research methodology it certainly made more sense for him to complement Aarne and Thompson's painstakingly detailed inventory with a model that was at least in part based on extrapolation, rather than to start bottom-up – and this all the more since the two approaches could very well coexist, as Cesare Segre observes:

> The extremely empirical methodology of Stith Thompson and the Finnish school does not conflict at all with that of Propp, given that their aims and objectives

[4] For a critical assessment of the methodology established in the tradition of Aarne and Thompson, see Dundes 1977.

> have nothing in common. Instead of tackling a finite corpus, it aims at an exhaustive inventory; it does not confine itself to homogeneous tales but considers the old together with the new, the literary as well as the popular. (Segre 1995, 28)

Propp, whose theory counts among the most influential in narratological and computational research of narrative alike, was in the end somewhat less of a formalist, and at heart still more of a traditional philologist than one would expect. He did look at his corpus and abstracted from it – but he is also likely to have jumped back and forth between induction and deduction. Being a folklore researcher, his formalist approach was still strongly rooted in phenomenological experience – and the title of his book, which refers to Goethe's morphology (from which Propp took a motto for his book) indicates his intellectual and philosophical kinship. Like the "Urpflanze" from which, according to Goethe, all other plants in their manyfold mutations had developed over biological time, Propp conceptualized something of an "Urnarrative". It is thefore important to understand that the methodological focus in his approach is not purely on the generic, but also on the genetic and historical dimension.

This methodological bifurcation can still be traced in the structuralist narratology that built on Propp, notably in Genette's seminal 1972 *Discours du récit* in which he derived an entire narratological taxonomy from an in-depth analysis of one exemplary text (Genette 1972). Of course, Proust's À la *recherche du temps perdu* is a heavy-weight from which many phenomena can be gathered, and Genette of course also brought to fruit his considerable knowledge of other works of world literature in as much as his readings of earlier theories of narrative, such as those developed by Foster or Lämmert. Nevertheless, the examples of Propp and Genette both prove that highly useful theories of narrative have thus far never emanated from stringent empirical analysis and by way of perfectly transparent inductive reasoning. While empirical approaches and shared resources certainly have a role to play in narrative research, be it in traditional literary history, be it in formalist and structuralist narratology or in computational modelling of narrative, we should not expect them to be able to replace speculative intelligence altogether. We will need both, and we must find out how to combine these two research strategies in a fruitful way. But how can one arrange for a meeting of minds between number-crunching algorithms and human intelligence, between computation and speculation?

4. Toward an applied Computational Narratology

Humanists as well as Computational Scientists measure the efficiency of machine intelligence by comparing it to the output of the human brain. The so-called *Turing Test* – Turing himself referred to it as "the imitation game" – is one example for this: a machine passes the test when a human being cannot distinguish its computationally produced output from that produced by the natural intelligence of another human being (Turing 1950, 433-460). But as the demise of AI has shown, human intelligence has a knack for escaping formalization – it is very hard to model in a computational system. Of course one can always blame the dynamics and complexity of real life context for that, but in a philosophical perspective there is more at stake. A simple hypothetical experiment might serve to illustrate the point:

Let us assume we could feed a computer all the narratives that mankind has ever produced, every single one of them in all its variations, and that we could also feed it with all the contextual data that define every potential reading of every single narrative in history. Then we will run some statistical algorithms on this huge amount of data and its multi-dimensional contexts, including algorithms that can detect the meta-logic of variation, genre formation, shifts in cultural conventions etc. The computer will identify patterns in this data: co-occurrences, correlations, dependencies, etc. However, when asked to *produce an original narrative*, none of this analytical power will help: at the very best the machine might be able to mimic human creativity and reconstruct a variant on a known narrative, using what is called case-based reasoning. But will we regard its output as equivalent to that of a human narrator? Most probably not; we are more likely to criticise it for merely imitating human narrative competence, just as we would criticize an epigonal author for his lack of originality.

But what if a human decides nevertheless to *react* to this imitative narrative *as if it were original*? This question, again, was also raised by Turing, who rightly pointed out that part of our scepticism toward machine creativity stems from our definition of the concept of originality. Originality is mainly in the eye of the beholder; it does not make sense to conceptualize it purely from the generative perspective of the originator. Focusing on the effect of 'surprise' as an indicator for originality, Turing argued that

> … it is perhaps worth remarking that the appreciation of something as surprising requires as much of a "creative mental act" whether the surprising event originates from a man, a book, a machine or anything else.
>
> The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it. It is a very useful assumption under many circumstances, but one too easily forgets that it is false. A natural consequence of doing so is that one then assumes that there is no virtue in the mere working out of consequences from data and general principles. (Turing 1950, 451)

If we take Turing's interjection seriously then we cannot only look at the machine when it comes to the computational investigation of narrative logic – we have to *look at the machine and the human in interaction*. The machine in isolation will never be able to come up with anything that a human expert would deem original unless we invest our reading of what it has generated with an interpretation that reflects its automatically produced narrative in the light of our human concerns.

This means that rather than handing over a huge corpus of narratives and their holistic interpretations to the computer, we must work on a more elementary level: Let human readers provide intelligent semantic annotation of the sentences that make up narratives, and then forward the annotations to the computer for a combined statistical analysis that correlates the markup with the original object data – the narratives annotated and explicated by the human reader.

This research design is the main idea in a new variant of computational narratology, one that I suggest to term *applied computational narratology.* It is represented, among other, by the project heureCLÉA, the third stage in a development that began in 2009 with the development of CATMA: short for "Computer Aided Textual Markup and Analysis" (Gius *et al.* 2012).[5] CATMA started out as an attempt to re-engineer TACT (Textual Analysis Computing Tools)[6], a ground breaking DOS-application developed in the mid-1980s by John Bradley, Ian Lancashire and others at Toronto University. However, CATMA soon grew into a far more complex stand-alone desktop application

[5] For further details on the open source software CATMA see http://www.catma.de
[6] See http://projects.chass.utoronto.ca/tact/

for TEI/XML-compliant markup of digital text, as well as for the subsequent analysis of markup and original text. CATMA thus supports the two quintessential philological operations of

(1) text annotation, i.e. markup, and
(2) combined analysis of the source text and its associated annotation.

Of course one can also perform these two tasks without computers – but **with** a computer, we can handle much larger corpora, and we will also be more consistent in our method because machines do not tolerate inconsistency. Only sufficiently well-defined concepts can be operationalized for a computational approach, whereas fuzzy definitions or inconsistent use of them will render the approach useless.[7]

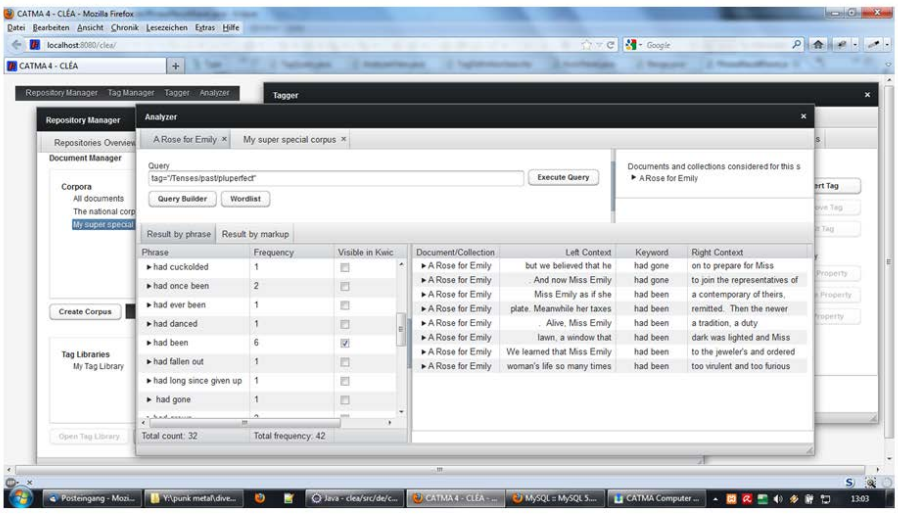


Fig. 1. CATMA 4.0 interface.

To illustrate these two methodological advantages – broader empirical basis and more rigorous application of method – let us look at a second hypothetical research project. Let us assume we want to study the historical development of a particular narratological phenomenon throughout a given period – say, of *analepsis* in French non-fiction narrative between 1815 and

[7] Note that this type of 'fuzzyness' goes well beyond what can be currently formalized with fuzzy logic based approaches.

1871. Maybe this research is motivated by a qualitative research hypothesis, such as "I believe that *analepsis* in non-fiction narratives between 1815 and 1871 co-occurs with a reference to pre-Napoleonic political events in French and European history". Note that this is not the type of question one can answer via a Google nGram-query as *analepsis* is not an explicitly (or sufficiently implicit) marked surface feature of texts – rather, it is an interpretative category that presupposes narratological interpretation.

To validate or reject this hypothesis we will first aggregate a relevant literary corpus. Then human annotators will be asked to annotate every individual text in CATMA by highlighting and tagging all occurrences of *analepses* and historical *events* in terms of a narratological taxonomy. In a third step we can now use CATMA to analyse the frequency and distribution patterns of *analepses* and *events*. When looking at the sections of text where co-occurrences of both phenomena can be found, we might realize that a second narratological phenomenon seems to become more and more prominent as we approach the 1871 end of our corpus – perhaps some aspect of focalization or perspective? Interesting; so let's go back to the corpus and annotate these as well and then re-run the analysis. And so forth until we have eventually exhausted the data, lost interest, or spent all our research funds…

From a conceptual and methodological point of view this workflow is of course standard philological practice, just that one uses a computational tool. But is it trivial? Not as far as the computational aspect is concerned, because the conceptual model of the hermeneutic and philological workflow that underlies CATMA is no longer that of a mechanistic input-output-pipeline. Rather, it is cyclical and open ended. Analysis, annotation and interpretation go hand in hand and one can enter, exit, terminate or re-start this procedural sequence at any point.

The importance of acknowledging the principal open-endedness of philological problem-solving is crucial to CATMA. But there is more to it than recursion: Unlike other annotation tools, CATMA also allows you to do exactly what human narratologists and philologists do all the time: disagree, debate, contradict one another, and even contradict ourselves. One annotator may have good reasons to highlight a certain passage in a text as an *analepsis*, whereas a second person thinks the very same passage should rather be labelled a *representation of inner speech and thought*. CATMA is designed to accept and memorize both variants, and it will also accept ***n*** others, if need should be. One particular text can have an infinite number of annotations; some will overlap, others will contradict each other – and what will eventual-

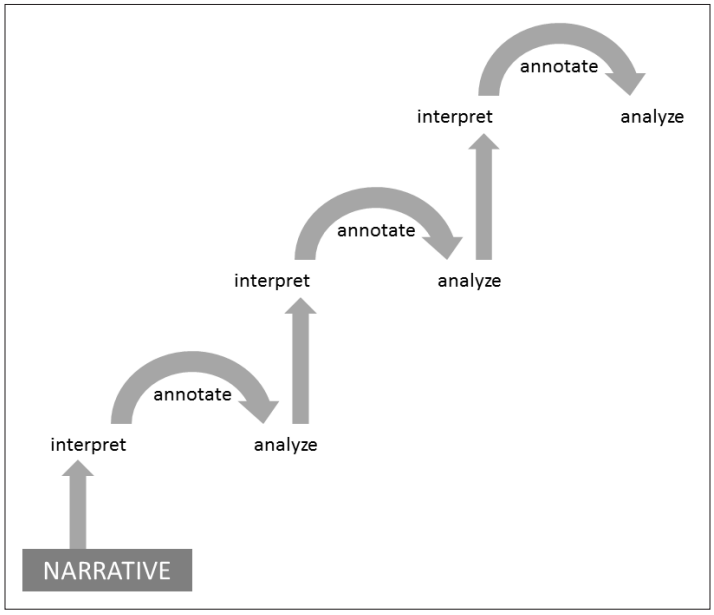


Fig. 2. Recursive workflow model implemented in CATMA.

ly come out as the *communis opinion* is a question of discourse and contexts, not one of dogma or an automated reasoning pipeline. CATMA's conceptual model invites us to consider dissent as productive, rather than supporting the illusion of a deterministic labelling of, among other, narratological phenomena.

At this point a second methodological principle comes into play: collaboration. The business of philology has always been a communal effort – we exchange and discuss our findings, and we build on each other's results. Computational narratology has to take cognisance of this well-established routine, and it must accommodate and support this social practice of our scholarship as best as possible.

With this desideratum in mind we progressed to the second phase of development and enhanced the software in order to support collaboration among scholars and annotators. CATMA's collaborative version no longer 'lives' on an individual researcher's computer, but is completely web based. It provides a platform for narratologists and other text scholars to work in

cooperation, and to share their results *ad hoc* via a seamlessly integrated web repository. This CATMA version is called CLÉA[8], short for *Collaborative Literature Exploration and Annotation.* Its philosophy is unflinchingly *open source* and *open data*, meaning that anybody can use the platform for free – on condition that the user is willing to share

- primary data: the object texts and corpora
- meta data: the markup produced in the course of a text analysis project
- tag libraries: the taxonomies and descriptive terms that were used to annotate primary data.[9]

While crowd sourcing may help us to proceed with annotation more quickly it nevertheless still necessitates a lot of manual tagging. More importantly, crowd sourcing per se does not really make use of the computer's abilities either; it merely emulates what humanists and narratologists have always done in their traditional research practice: share the work load, exchange results, engage in critical discourse, revisit and revise previous work, etc. – just that all of this can now be done in real time and with more material.

Against this backdrop the next logical step is: automation. Wouldn't it be helpful if the CLÉA's system would also include some background function that will learn how to detect, say, a potential *analepsis* and present the users of the platform with automatically generated suggestions for its narratological annotation – just as Google's or Amazon's recommender algorithms have over time learned to generate surprisingly relevant suggestions for where to click or what to buy? In other words, can one progress from a mere narratological and collaboratory workbench toward an automated narratological heuristics?

---

[8] The development of CLÉA was generously supported by two subsequent Google Digital Humanities awards. Our particular thanks go to Will Brockman at Google Inc. for help with providing access to a sub-set of Google Books documents and for technological advice. As for the accent on éxploration, we are aware that it is wrong: it was added intentionally to highlight the 'diacritical concerns' of the non-English speaking world.

[9] The single biggest hindrance for the implementation of this research philosophy in a computational system is not technological, but sociological and economical. Apart from copyright issues relating to primary texts in one's corpus it also raises the need to credit and document intellectual property and the collaborating individuals' rights to annotations and tag libraries. The latter is an aspect of research politics and legal frameworks affecting computational approaches toward text that has hitherto received comparatively little attention in Digital Humanities. At the same time, it serves to illustrate that collaboration centred computational approaches in the humanities have a significant potential for transforming institutionalized research practices in the traditional humanities.

This is the goal of heureCLÉA. In this third project phase the aim is to exploit the entire corpus of individually as well as collaboratively generated narratological annotations collected by CATMA using *machine learning* (ML) procedures. A ML approach is statistical and probabilistic, not intelligent. A pattern recognition algorithm has no idea what *analepsis* means, and what semantic and aesthetic purpose it serves in, say, a specific paragraph of *A la recherché du temps perdu*. But it is precisely its ignorance toward content and semantics that enables a ML algorithm to identify correlations which, albeit hermeneutically absurd, might prove statistically valid nevertheless. For example, we might find out that there is, say, a 98% chance for the occurrence of an *analepsis* in *A la récherche* if the 382nd character to the left of a new sentence is an exclamation mark, and the 3rd narratological phenomenon to the right is either a switch from *zero* to *internal focalization*, or to an *extradiegetic* perspective. Once sufficiently tested for formal validity in terms of precision and recall we can use such a finding as a heuristic and analyse a new text or corpus for similar distribution patterns. These findings can then be brought to the attention of a human user – just as a typical recommender system on a web merchant platform would do, only that the aim is not to trigger a commercial transaction, but rather an intellectual one: "Other users have annotated this text feature as an analepsis." When it comes to high level concepts it still remains for the human user to decide whether he will "buy" this suggestion and authorize CLÉA to annotate this, as well as similar findings, accordingly. On the other hand, where comparatively low-level attributions are concerned, such as in POS-tagging, such a complicated heuristics would be dysfunctional – there is no point in taking a detour via crowed sourced annotations and the statistical black box of ML if we already have sufficiently explicit grammars or glossaries available that are considered normative in a given research community.

This, then, is the complete heuristic workflow schema of the approach toward a computational narratology which is currently being developed and tested in heureCLÉA[10]:

[10] The heureCLÉA corpus consists of 21 German short stories from around 1900. Five annotators are analyzing the corpus, using a tagset that consists of roughly 100 narratologically relevant tags. To date, about 19,600 annotations have been captured.

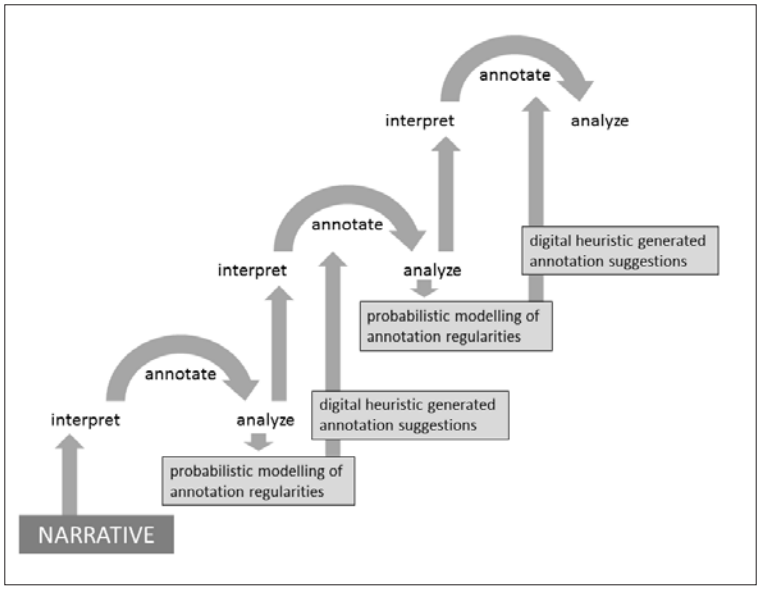


Fig. 3. Digital heuristic enhanced annotation workflow in heureCLÉA.

5. Conclusion

If the approach outlined thus far strikes you as absurd, then you are probably not really a hard core narratologist: for conceptually it is analogous to the approach taken by Propp when he abstracted from narrative content and focused on what he termed 'narrative functions'. Admittedly, his concept of *function* was not as removed from the phenomenological and experiential dimension of narrative as a purely statistical model in its radically mathematical reduction of a narratological phenomenon to the prevailing numerical correlation among n features will be. But the two approaches are nevertheless based on the same methodological principle: Formalism in its narratological as much as in its mathematical variant is an appeal to abstract from surface phenomena, and from content, and to focus on an object's structural logic instead. To repeat: an automated heuristic functionality will not make intelligent, manual annotation obsolete, nor can it automate narratological text description altogether. Rather, we want to contribute to a computational *heuristics of narrative* by building a system that can analyse a given object text or corpus using purely statistical means, and then

propose to the human expert a probable narratological categorization for qualitative validation. Also note that such a system does not implement a specific pre-defined narratological theory – it merely 'learns' dynamically from analysing the (hopefully) intelligent annotation output of human users and relating it mathematically to the object texts or corpora that these users dealt with. The choice of annotation schema and underlying narratological theory remains with the human user, and the automated heuristics will be bound by its conceptual constraints. But users can of course also decide to contrast competing heuristics based on different theoretical underpinnings – for example, will a Genette inspired narratological annotation of Proust's *A la récherche* yield more interesting heuristic output than one that is based on the model of Schmid or Stanzel?

In this respect heureCLÉA's approach toward a computational narratology goes beyond that of a traditional DH research tool: rather than just using the tool to "do the job", we use it to reflect upon, to test and evaluate the underlying methodological and theoretical primitives of our research practice. In his recent article in the web-based *living handbook for narratology*, Inderjeet Mani has defined *Computational Narratology* as

> (…) the study of narrative from the point of view of computation and information processing. It focuses on the algorithmic processes involved in creating and interpreting narratives, modeling narrative structure in terms of formal, computable representations.

The scope of computational narratology, he continues,

> … includes the approaches to storytelling in artificial intelligence systems and computer (and video) games, the automatic interpretation and generation of stories, and the exploration and testing of literary hypotheses through mining of narrative structure from corpora. (Mani 2013, par. 1, emphasis added)

CATMA and CLÉA are obviously not "automatic storytelling systems"; they are workbenches and analytical platforms for philologists and narratologists. And while the latest stage in CLÉA's development aims at designing a heuristic functionality, this is a far cry from the second goal, i.e. that of an "automatic interpretation of stories". In terms of Mani's definition, our project rather falls into the third conceptual category, namely that of the "exploration and testing of literary hypotheses through mining of narrative structure from corpora."

Applied computational narratology, as I would propose to label Mani's third category, then, combines the methods of computational modelling with philology's traditional hermeneutic interest in explicating the meaning of narrative as man's privileged symbolic practice. Its principal credo is that hermeneutic and philological interest can go hand-in-hand with formal modelling, provided that formalization can be related back to the experiential realm of phenomena. This is, I believe, what Vladimir Propp really had in mind when he presented his formula of 31 narrative functions: not an impartial abstraction that bears no trace of human interpretive engagement with narratives, but rather a high-level interpretive annotation of narratives that will facilitate new inter-subjective insights into the functioning and functions of narratives.

Looking at the Digital Humanities at large, a remark by the English novelist Tim Parks comes to mind which helps us to understand why the tension between phenomenology and formalism is so fascinating, and why we ought to accept it as a productive premise in all humanities research. Commenting on the necessary subjectivity of interpretation, Parks concludes: "To be impartial about narrative would be to come from nowhere, to be no one" (Parks 2014).

6. References

Aarne A. (1910). *Verzeichnis der Märchentypen mit Hülfe von Fachgenossen*, Suomalainen Tiedeakatemia.

Aarne A., Thompson S. (1961). *The types of the folktale. A classification and bibliography*, Academia Scientarum Fennica.

Bremond C., Verrier J. (1984). *Afanasiev and Propp*. «Style», vol. 18, no 2, pp. 177-195.

Chace A.B. (1979) [1927-1929]. *The Rhind Mathematical Papyrus. Free Translation and Commentary with Selected Photographs, Translations, Transliterations and Literal Translations*. Classics in Mathematics Education 8, 2 vols, National Council of Teachers of Mathematics.

Dundes A. (1997). *The Motif-Index and the Tale Type Index: A Critique*. «Journal of Folklore Research», 34, 3, pp. 195-202.

Eco U. (1994). *Dell'impossibilità di costruire la carta dell'impero 1 a 1*. In: *Il secondo diario minimo*, Bompiani.

Genette G. (1972). *Discours du récit*, Éditions du Seuil.

Gius E., Meister J. C., Petris M., Schüch L. (2012). *Crowdsourcing Meaning. A Hands-On Introduction to CLÉA, the Collaborative Literature Éxploration and Annotation Environment*. In Digital Humanities 2012 Conference Abstracts, Hamburg University Press, pp. 24-25. URL=http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/crowdsourcing-meaning-a-hands-on-introduction-to-clea-the-collaborative-literature-exploration-and-annotation-environment/ [22.04.2014].

Goethe, J.W.v. (1810). *Zur Farbenlehre*, Cotta: *Enthüllung der Theorie Newtons. Des Ersten Bandes Zweyter, polemischer Theil*. URL=http://www.deutschestextarchiv.de/book/format/html/goethe_farbenlehre01_1810?hyphenation=1&normalize=1&fw=1&marginals=1&footnotes=1&textwidth=0&format=html [23.04.2014]

Goodfellow I. J., Bulatov Y., Ibarz J., Arnoud S., Shet V. (2014). *Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks.*
URL=http://arxiv.org/pdf/1312.6082v4.pdf [21.04.2014].

Mani I. (2013). *Computational Narratology*. In P. Hühn *et al.*, eds., *the living handbook of narratology*, Hamburg University Press. URL=http://www.lhn.uni-hamburg.de/article/computational-narratology [15.09.2013].

McCarty W. (2005). *Humanities Computing*, Palgrave Macmillan.

Moretti F. (2000). *Conjectures on World Literature*. «New Left Review», 1, January-February 2000. URL=http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature [1.05.2014]

Segre C. (1995). *From Motif to Function an Back Again*. In C. Bremond, J. Landy, T. G. Pavel, eds., *Thematics: New Approaches*, SUNY Press, pp. 21-32.

Snow C.P. (1993) [1959]. *The Two Cultures*, Cambridge University Press.

Terras M., Nyhan M., Vanhoutte E., eds. (2013). *Defining Digital Humanities. A Reader*, Ashgate.

Turing A. (1950). *Computing Machinery and Intelligence*. «Mind», vol. 59, no 236, pp. 433-460. URL=http://www.loebner.net/Prizef/TuringArticle.html [22.04.2014].

Parks T. (2014). *Where I'm Reading From*. In *The New York Review of Books*, 19 March 2014, URL=http://www.nybooks.com/blogs/nyrblog/2014/mar/19/where-im-reading/ [1.05.2014].

Propp V. (1958) [1928]. *Morphology of the Folktale*, Research Center, Indiana Univ.

Uther H.-J. (2004). *The Types of International Folktales. A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*, Suomalainen Tiedeakatemia.

Invited

Contributions from Research Groups and Centers /
Contributi di centri e gruppi di ricerca

Nuovi scenari per la ricerca in filosofia: i testi e gli strumenti del portale Daphnet*

Michela Tardella, Cristina Marras
Istituto per il Lessico Intellettuale Europeo e Storia delle Idee-Cnr, Roma, Italia
michela.tardella@iliesi.cnr.it, cristina.marras@cnr.it

Abstract. Questo articolo presenta l'organizzazione della piattaforma *Daphnet*: *Digital Archives of PHilosophical text on the Net* pubblicata dall'Istituto per il Lessico Intellettuale Europeo e Storia delle Idee del Cnr di Roma. Descrive i linguaggi di codifica dei testi e gli strumenti utilizzati per l'annotazione semantica, segnatamente il software *Pundit* e The*o*fPhilo - *Thesaurus of Philosophy*, una collezione terminologica finalizzata all'arricchimento semantico, alla soggettazione dei testi e all'*information retrieval*. In conclusione si tracciano alcune linee di riflessione legate alle prospettive e agli sviluppi futuri del lavoro, con particolare riguardo ai risultati dell'interazione tra filosofi e informatici nell'elaborazione di strategie per la ricerca, la rappresentazione e la condivisione dei risultati. L'obiettivo è infatti quello di riflettere sulla creazione di ambienti integrati per la ricerca utili non solo per l'accesso alle fonti e ai documenti, ma anche per lo scambio e la creazione di nuova conoscenza.

Parole chiave: web semantico, Xml-Tei, filosofia digitale, tesauri, lessicografia.

1. Introduzione

Questo lavoro è suddiviso in quattro parti: nella prima, testi, si presenta il corpus di opere pubblicate in formato digitale dall'Istituto per il Lessico Intellettuale Europeo e Storia delle Idee del Cnr di Roma[1] nel porta-

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. Cleup, Padova, 2014.

[1] L'Iliesi nasce nel 1964 come gruppo di studio del Consiglio Nazionale delle Ricerche (Cnr) presso l'Istituto di Filosofia dell'Università di Roma. Nel 2001 diviene Istituto del Cnr assumendo il nome di *Lessico Intellettuale Europeo e Storia delle Idee* (Iliesi): nell'Istituto confluì allora, come sua sezione, il Centro di Studio per il Pensiero Antico (http://www.iliesi.cnr.it). Attualmente l'Iliesi da accesso ad una ampia collezione di risorse digitali di testi, dall'Archivio dei filosofi del Rinascimento ai Lessici filosofici dell'età moderna (http://www.iliesi.cnr.it/

le *Daphnet*: *Digital Archives of PHilosophical text on the* NET (http://www.daphnet.org/). Nella seconda parte, linguaggi, vengono sinteticamente descritte le modalità di codifica dei testi e affrontati alcuni dei problemi emersi durante questo specifico lavoro. Nella terza parte, strumenti, viene descritta sia l'applicazione utilizzata per l'annotazione semantica dei testi nelle piattaforme, il software Pundit, sia The*of*Philo - *Thesaurus of Philosophy*, una collezione terminologica finalizzata all'arricchimento semantico, alla soggettazione e all'*Information Retrieval* (IR). Nella quarta e ultima parte, vengono proposti alcuni temi legati alle prospettive e agli sviluppi futuri del lavoro; si discutono, in particolare, le diverse modalità di interazione tra filosofi e informatici nell'elaborazione di strategie per la ricerca, la rappresentazione e la condivisione dei risultati con l'obiettivo di riflettere su strumenti e processi di standardizzazione atti a favorire la creazione di ambienti integrati, utili non solo per l'accesso alle fonti e ai documenti, ma anche per lo scambio e la creazione di nuova conoscenza[2].

Alcuni temi, ormai condivisi con gran parte di coloro che affrontano la pubblicazione digitale di testi e la necessità dell'interoperabilità degli strumenti (Gold 2012), hanno fatto da sfondo e da guida alle riflessioni e descrizioni condotte in questo articolo: l'integrazione tra gli archivi e l'accessibilità dei materiali, l'attenzione alle modalità innovative di analisi di corpora di grandi dimensioni, le diverse forme di rappresentazione dei dati e le connessioni tra differenti tipi di dati, le architetture che favoriscono la strutturazione di ambienti integrati di lavoro, gli strumenti di supporto per la collaborazione e la costruzione di comunità scientifiche in filosofia.

---

attivita.php?tp=a_d), per i quali è disponibile un motore di ricerca per termini, autori e opere che consente l'interrogabilità di tutti gli archivi.

[2] Questo lavoro nasce dalla riflessione maturata nell'ambito delle attività dell'ILIESI condotte su due recenti progetti europei: Discovery (http://www.discovery-project.eu) e Agora (http://www.project-agora.org); per una descrizione del progetto si vedano Marras, Lamarra 2013 e Tardella 2013. Descrive dunque alcuni dei risultati ottenuti dai gruppi di ricerca dedicati. Oltre alle autrici del presente contributo, hanno lavorato ai progetti: Antonio Lamarra che ha anche coordinato il progetto Agora, Roberto Palaia, Marco Veneziani, Ada Russo, Simona Lampidecchia, Giancarlo Fedeli, Emidio Spinelli, Francesco Verde, Andrea Costa, Angela Taraborrelli, Francesco Giampietri, Alberto Manchi, Daniela Papitto, Giuseppe Iannotta, Giuseppe Lucchesini, Miriam Bianco.

2. Testi

Daphnet è un portale multilingue costituito da 4 piattaforme testuali: *Ancient Source* e *Modern Source* dedicate alla pubblicazione di fonti primarie; la *Daphnet Digital Library* dedicata alla pubblicazione di fonti secondarie collegate, per lo più, ai testi della letteratura primaria; la rivista internazionale *Lexicon Philosophicum International Journal for the History of Texts and Ideas* (fig. 1). Le piattaforme si caratterizzano per alcuni aspetti comuni quali l'utilizzo di programmi e applicazioni *open source*, l'accesso libero e aperto, la codifica standard dei testi, un identificativo stabile per ogni documento pubblicato e un *permalink* per la loro citazione, la presenza di un comitato scientifico di esperti per il controllo e la validazione dei contenuti. Le fonti primarie sono organizzate per autore, ogni autore e testo è corredato da opportuna documentazione. La barra di navigazione è organizzata in modo tale da consentire l'apertura di più documenti contemporaneamente per facilitare il lavoro; ogni documento selezionato può essere annotato semanticamente (vedi sotto §4); di ogni volume pubblicato sono stati inoltre compilati i metadati.

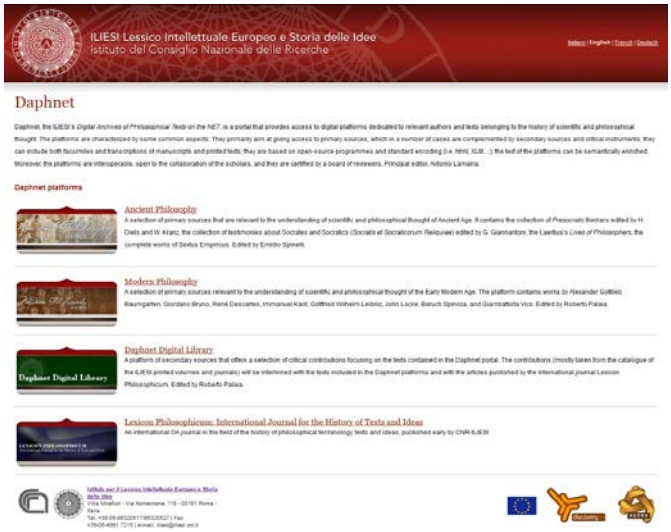


Fig. 1. Homepage del portale *Daphnet*.

L'architettura delle piattaforme è fornita da MURUCA (http://www.muruca.org)[3] un framework per la costruzione di biblioteche digitali che consente non solo di pubblicare e editare oggetti digitali, ma anche di essere integrato con strumenti di ricerca. MURUCA è interoperabile con le grandi biblioteche digitali e con progetti come Europeana e si avvale, inoltre, di tecnologie *linked open data*.

Modern Source (http://modernsource.daphnet.org)

La piattaforma *Modern Source* consente di accedere ad un vasto corpus di testi in lingua latina, italiana, francese e tedesca, rappresentativi della storia del pensiero filosofico e scientifico dei secoli XVI, XVII e XVIII. Include opere di Giordano Bruno, René Descartes, Gottfried Wilhelm Leibniz, John Locke, Baruch Spinoza, Immanuel Kant, Alexander Gottlieb Baumgarten, l'*Opera Omnia* di Giambattista Vico in undici volumi curata da Fausto Nicolini, per un totale di 108 volumi. Alcune di queste opere, come ad esempio i testi di Giambattista Vico, sono state pubblicate non solo in trascrizione, ma anche in facsimile.

I testi sono stati selezionati seguendo alcuni criteri: la loro rilevanza scientifica e culturale, il fatto che attestano l'affermarsi di una moderna terminologia filosofica e scientifica e che siano le edizioni effettivamente circolanti al tempo e dunque tra le più rappresentative del panorama culturale della prima modernità.

Ancient Source (http://ancientsource.daphnet.org)

È un originale archivio, relativo alla filosofia antica, che include diverse sezioni: la sezione *Pre-Socratics Source,* che offre la trascrizione dei frammenti dei filosofi presocratici (*Die Fragmente der Vorsokratiker*) edita da Hermann Diels e Walther Kranz e pubblicata in 3 volumi per la Weidmann di Berlino nel 1958, accompagnata dalla traduzione italiana curata da Gabriele Giannantoni, *I Presocratici. Testimonianze e frammenti* pubblicata per i tipi della Laterza nel 1983; la sezione *Socratics Source,* che dà accesso alla trascrizione della collezione integrale delle *Socratis et Socraticorum Reliquiae* originalmente raccolte e edite da Gabriele Giannantoni (Napoli, 1990); la

[3] La tecnologia a supporto del lavoro dell'ILIESI per la costruzione del portale *Daphnet* e delle singole piattaforme è stata fornita da Net7 (http://www.netseven.it).

sezione *Diogenes Laertius Source*, nella quale è possibile consultare il volume *Vita e opinioni dei filosofi*, trascrizione di una sinossi delle ultime tre edizioni cartacee della *Vita e opinioni dei filosofi* in dieci libri, collazione delle edizioni di Robert Drew Hicks, Herbert Strainge Long e Miroslav Marcovic, con la traduzione italiana di Marcello Gigante e la ricostruzione a fronte del testo greco secondo le note filologiche del traduttore; la sezione *Sextus Empiricus Source*, che contiene l'*Opera Omnia* di Sesto Empirico nell'edizione curata da Hermannus Mutschmann e Jurgen Mau (*Sexti Empirici Opera* recensuit, coll. *BT*, Leipzig 1912-1954, 4 voll).

Per la loro collocazione nell'ambito della piattaforma si è tenuto conto della specifica tipologia delle edizioni di origine, in particolare, per quanto riguarda i Presocratici e Socrate e i Socratici, si tratta di sillogi contenenti frammenti e testimonianze estratte da un migliaio di opere della letteratura antica.

Daphnet Digital Library (http://scholarlysource.daphnet.org/index.php/Ddl)

Si tratta di una importante piattaforma, costruita utilizzando l'*Open Journal System* (Ojs), dedicata alla letteratura secondaria e suddivisa in diverse sezioni. La biblioteca digitale di *Daphnet* contiene, per quanto concerne il pensiero moderno, nella sezione "Lexicon philosophicum", una ampia selezione di articoli già pubblicati nei volumi a stampa di *Lexicon Philosophicum. Quaderni di Terminologia filosofica e storia delle idee*; una serie di studi, nella sezione "Iliesi Proceedings", pubblicati nei volumi degli atti dei Colloqui Internazionali organizzati dall'Iliesi dedicati alla storia della terminologia. Entrambe le serie sono edite della casa editrice Olschki di Firenze. Per quanto riguarda i contributi relativi ai testi antichi, la sezione "Elenchos" presenta una scelta di lavori critici già pubblicati dalla rivista *Elenchos. Rivista di studi sul pensiero antico* (Bibliopolis, Napoli). È inoltre possibile consultare dei volumi monografici nella sezione "Book collection": *Socratis et Socraticorum reliquiae*, volume IV curato da Gabriele Giannantoni, e *Questioni Scettiche: letture introduttive al pirronismo antico* di Emidio Spinelli (Lithos, Roma, 2005). Completa la raccolta la sezione "Lexica", dedicata a strumenti e risorse lessicografiche. Al momento è disponibile il *Glossarium Epicureum* di Hermann Usener, con la prefazione di Marcello Gigante e Wolfgang Schmid, pubblicato a stampa per le Edizioni dell'Ateneo di Roma nel 1977.

Tutti gli studi inclusi nella biblioteca digitale di *Daphnet* sono stati scelti per costituire il nucleo iniziale dell'apparato di letteratura critica delle fonti primarie, in alcuni casi rendendo accessibili agli studiosi lavori fondamentali di non facile reperimento. Il corpus delle fonti secondarie è inoltre arricchito da nuovi contributi critici, dal 2013 è online la rivista internazionale *Lexicon Philosophicum International Journal for the History of Texts and Ideas* (http://lexicon.cnr.it), che pubblica annualmente, dopo un rigoroso processo di valutazione[4], studi, note critiche, saggi, e in generale contributi relativi alla storia della filosofia e della scienza, con particolare riguardo alla questioni terminologiche e lessicali.

3. Linguaggi

Per recuperare la consistente quantità di testi già digitalizzati e presenti nelle banche dati dell'Iliesi per la pubblicazione in *Daphnet*, si è proceduto alla conversione della codifica proprietaria Iliesi nella codifica standard Xml-Tei[5]. Ciò ha sollevato alcune questioni non solo operative ma anche teoriche. Il problema è stato, innanzitutto, nella fase di arricchimento strutturale e semantico, quello di preservare, valorizzare e usare il più possibile la ricchezza delle informazioni offerte dalla originaria codifica Iliesi. I documenti contengono infatti tutte le informazioni necessarie per un diretto e 'facile' accesso ai testi (volume, pagina, riga, differenze nei font, etc.). L'obiettivo sotteso al lavoro sui testi della banca dati era quello di dare informazioni utili alla lemmatizzazione, in quanto il progetto originario dell'Istituto era la realizzazione di un lessico intellettuale europeo. I testi venivano pertanto interrogati e analizzati, al fine di una loro lematizzazione, con un approccio che può essere definito metodologicamente 'top down', in ragione di un orientamento all'uso piuttosto che alla struttura. I documenti venivano dunque divisi in zona-testo, zona-lemmi e riferimenti. Per la pubblicazione in *Daphnet* l'obiettivo e l'approccio al testo sono stati differenti. Oltre all'esigenza di trasformare la codifica proprietaria in una codifica standard, si è

[4] La rivista sperimenta un complesso processo di peer review sia doppio cieco, sia aperto. Per maggiori dettagli si può vedere http://lexicon.cnr.it/index.php/LP/about/editorialPolicies#peerReviewProcess

[5] Per interessanti approfondimenti sulla codifica Xml-Tei in relazione ad ambienti integrati di ricerca si vedano Eide 2013 e Reaner 2013.

trattato innanzitutto di pensare il testo non in una banca dati con funzioni di interrrogazione principalmente terminologica, quanto di un vero e proprio archivio digitale che consentisse una interrogabilità trasversale, la cui codifica doveva rappresentare la struttura macro del testo stesso restituendone l'integralità (anche se segmentato semanticamente) e la struttura (divisione in righe, pagine, paragrafi, etc.).

Per la trasformazione delle codifiche sono stati elaborati degli script utilizzando delle tabelle di corrispondenza; facendo ricorso poi all'editor Oxygen sono state controllate le conversioni e si è aumentato il livello di granularità delle informazioni. Per esempio, nella codifica dei nomi propri, dalla indicazione 'np', si è passati ad una maggiore granularità, specificando che si tratta di un antroponimo. È stata inoltre aggiunta come *key* la normalizzazione del nome (fig. 2).

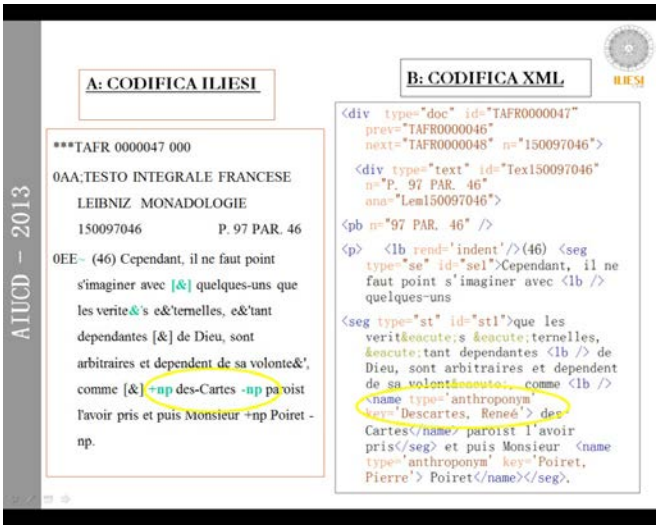


Fig. 2. Dalla codifica proprietaria a XML-TEI.

La codifica dei testi e la loro disposizione in rubriche ha trovato una sistemazione appropriata con le "div" (*fragment division*) e gli "id" (*identification code*) che identificano e tipizzano senza ambiguità ogni singolo documento. Il lavoro ha affrontato diversi problemi; la codifica dei papiri è stata, per esempio, particolarmente laboriosa perché la restituzione del

testo doveva in qualche modo conservare traccia dell'aspetto fisico del papiro stesso. Come codificare con precisione, ad esempio, una parte di testo mancante? Normalmente si è fatto ricorso a dei puntini di sospensione al posto di un codice. Oppure, come codificare i caratteri 'incerti'? È il codice "*unclear*" corretto per indicare questo caso particolare?

Il processo che ha accompagnato la trasformazione delle codifiche ha aperto un dialogo tra ricercatori, filosofi e linguisti, e coloro che hanno offerto la consulenza e il supporto tecnico al lavoro, un dialogo che si è rivelato interessante non solo per gli aspetti di carattere operativo (la codifica del testo) ma, soprattutto, per le riflessioni che ha stimolato. È emerso, infatti, il carattere interpretativo delle scelte dei criteri di codifica e le conseguenti implicazioni di carattere scientifico, non potendo la codifica stessa prescindere dagli obiettivi di ricerca e dalle finalità legate alla fruizione e all'analisi dei testi da pubblicare. L'impostazione tecnica è infatti inscindibile dall'approccio scientifico al testo che si va a codificare[6].

4. Strumenti

Gli strumenti sviluppati per la ricerca e per l'analisi dei testi pubblicati in *Daphnet* rispondono all'esigenza di testare nuove modalità di interrogazione e navigazione delle piattaforme e di proporre modelli innovativi per l'organizzazione della conoscenza inerente alle opere e agli autori di competenza. L'insieme di testi, strumenti e procedure di analisi e critica delle fonti è stato concepito in primo luogo per favorire lo sviluppo della collaborazione e della condivisione tra studiosi, ma ha l'importante valore aggiunto di costituire un apparato utile ed efficace per l'insegnamento, sia nelle università che nelle scuole superiori.

4.1 Annotazioni semantiche 'Text-to-Text'

Centrale, nel summenzionato quadro di lavoro, è stata l'attività di annotazione semantica (Andrews, Zaihrayeu and Pane 2011) per mezzo del software Pundit (http://thepund.it), un *semantic web annotator* sviluppato

[6] A questo proposito si veda il seminario di Ada Russo http://www.iliesi.cnr.it/iniziative/XML-TEI.pdf

dalla società Net7 di Pisa[7]. Concepito nel più ampio contesto del Web Semantico, il software è liberamente scaricabile e permette di compiere varie attività di annotazione su ogni tipologia di oggetto digitale presente nella rete (Grassi *et al.* 2013).

Dal punto di vista operativo Pundit utilizza un data model fondato sulla tecnologia RDF, 'Resource Description Framework'. Costituiti da un 'Soggetto', un 'Oggetto' e da una 'Relazione' (o Predicato) che li connette (come ad es. "G. W. Leibniz [S]/ *isAuthor* [P]/*Monadologie* [O]"), *statements* così configurati presentano la fondamentale caratteristica di consentire "l'interoperabilità tra applicazioni che si scambiano sul Web informazioni machine-understandable" (Signore 2002, 37), di condividere cioè, nella rete, informazioni strutturate sintatticamente secondo criteri universali. Per questo sono utilizzabili da altri studiosi e dalle macchine per la produzione di ulteriore nuova conoscenza.

Tra le molteplici annotazioni praticabili utilizzando Pundit, ci si è concentrati in particolar modo su due tipologie: la prima, definita 'Text-to-Text interlinking', consiste nella creazione di link tra le fonti primarie e la letteratura critica di riferimento disponibile nella *Daphnet Digital Library* o, più in generale, nella interconnessione di due testi (è possibile cioè mettere in relazione due fonti primarie o due fonti secondarie); la seconda, 'Text-to-Subject interlinking' (cfr. § 4.2), è finalizzata invece a stabilire relazioni tra i testi e un insieme di soggetti filosofici rilevanti, che costituiscono un vocabolario controllato di dominio preliminarmente definito dal gruppo di lavoro.

L'ambiente annotativo di Pundit è infatti molto duttile, plasmabile in relazione alle esigenze di coloro che lo utilizzano. Per quanto concerne il lavoro di ricerca condotto sui testi di *Daphnet*, è stato necessario introdurre nel sistema non soltanto il vocabolario controllato di cui si è appena accennato, ma anche una serie di specifiche relazioni attraverso le quali connettere tra loro le entità appartenenti alle classi dell'ontologia di riferimento[8] (persone - filosofi, editori e studiosi -, testi, luoghi e soggetti filosofici). Le relazioni sono dunque state articolate in quattro categorie: 1. 'Peoples' Pro-

[7] L'uso di Pundit è stato condotto nell'ambito dell'esperimento di *semantic linking* del progetto Agora (http://www.project-agora.org/experiments/semantic-linking/). Per l'ILIESI hanno lavorato all'esperimento Antonio Lamarra, Michela Tardella, Francesco Verde, Andrea Costa.

[8] Sul tema delle ontologie, oltre ai già citati Eide 2013 e Reaner 2013, si vedano Garshol 2004, Grenon e Smith 2009 e Gruber 1993.

perties'; 2. 'Properties of the Documents'; 3. 'Properties of the Texts', a loro volta suddivise in: 3a. 'Relations Text-to-Text': Direct quote (*QuotesDirectly, IsDirectlyQuotedBy*); Indirect Quote (*QuotesIndirectly, IsIndirectlyQuotedBy*); Refutation (*Refutes*, *IsRefutedBy*); Arguing in favour (*ArguesFor, IsArguedForBy*); Explanation (*Explains, IsExplainedBy*); Criticism (*Criticizes, IsCriticizedBy*); Agreement (*AgreesWith, IsAgreedWithBy*); Interpretation (*Interprets, IsInterpretedBy*); Similarity (*IsSimilarTo*); External reference (*MakesAReferenceTo, IsReferencedBy*); Internal reference (*MakesAnInternalReferenceTo, IsInternallyReferencedBy*); 3b. 'Relations Text-to-Subject'; 4. 'Relations Subject-to-Subject' (per le categorie 3b e 4 cfr. § 4.2).

Questo schema di lavoro può essere meglio presentato attraverso alcuni esempi. Nelle annotazioni 'Text-to-Text' sia il soggetto che l'oggetto della tripla RDF sono costituiti da testi o porzioni di testo. La fig. 3 mostra un'annotazione relativa a *Lineamenti Pirroniani*, II, 84 in cui si afferma: "Il frammento testuale evidenziato a sinistra [S] è simile (*IsSimilarTo*) [P] al frammento testuale evidenziato a destra [O]", quest'ultimo appartenente ad un'altra opera sestana, l'*Adversus Mathematicos*. L'annotazione stabilisce dunque una relazione tra le due porzioni testuali in base al principio di similarità concettuale. Un altro genere di annotazioni 'Text-toText' consiste nel connettere le fonti primarie alla bibliografia secondaria disponibile nella *Daphnet Digital Library* affinché gli studiosi possano compiere il passaggio diretto ai contributi critici.

Fig. 3. Annotazione con predicato RDF '*IsSimilarTo*'.

Nel lavoro di *semantic enrichment* si è fatto inoltre ricorso ad altre due funzioni disponibili in Pundit. La prima ha riguardato la connessione dei testi a database esterni alle piattaforme, in particolare DBPedia e Freebase, datasets appartenenti al *Web of Data*, veicolanti quindi informazioni semanticamente strutturate. La seconda funzione consente invece di commentare liberamente i testi, mettendo a disposizione della comunità scientifica informazioni e conoscenze elaborati dagli autori delle annotazioni e/o non reperibili in rete. Sono state generate triple che, ad esempio, suggeriscono link a pagine web della *Stanford Encyclopedia of Philosophy*, utile per approfondire lo studio di concetti, autori, scuole e temi filosofici o anche triple che specificano notizie e informazioni bio-bibliografiche relative a filosofi o, più in generale, a studiosi meno noti.

4.2 Annotazioni semantiche 'Text-to-Subject': TheofPhilo

Per quanto concerne le annotazioni semantiche 'Text-to-Subject', si è lavorato alla definizione di predicati che permettessero di fare delle soggettazioni piuttosto articolate, fondate non soltanto sull'occorrenza dei termini nel corpo del testo, ma anche sull'interpretazione del testo stesso in relazione al concetto di pertinenza. Le relazioni sono le seguenti: Definition (*Defines*, *IsDefinedBy*); Indirect Definition (*IndirectlyDefines*, *IsIndirectlyDefinedBy)*; Extensional Instantiation (*IsAnExtensionalInstanceOf*, *IsExtensionallyInstantiatedBy*); Intensional Instantiation (*IsAnIntensionalInstanceOf, IsIntensionallyInstantiatedBy*); Dealings (*DealsWith*, *IsDealtWithBy*). Un esempio di questo tipo di annotazioni è il seguente: "Il frammento testuale selezionato [S] È rilevante per (*DealsWith*) [P] il Subject *to alethes* [O]" o ancora "Il frammento testuale selezionato [S] Definisce (*Defines*) [P] il Subject *ataraxia* [O]" (cfr. fig. 4).

L'attività scientifica di annotazione semantica 'Text-to-Subject' ha implicato un'ampia riflessione sulla terminologia filosofica e sulla relativa lessicalizzazione nelle lingue di cultura presenti nel portale. Il plurilinguismo che caratterizza i testi di *Daphnet* ha inevitabilmente imposto la strutturazione di un thesaurus di dominio filosofico multilingue in Italiano, Latino, Greco, Francese e Inglese. Si è creato su questi presupposti un prototipo, The*of*Philo_*Thesaurus of Philosophy*, che al momento offre circa 4500 termini/entrate nelle cinque lingue sopra citate. Con il termine 'thesaurus' si è intesa una collezione terminologica concept-based (Magris *et al*. 2002), priva di

Fig. 4. Annotazione con predicato RDF '*Defines*'.

definizioni perché elaborata con lo specifico fine di soggettare le opere e di garantire il recupero delle informazioni relative ai concetti variamente lessicalizzati[9]. Una volta ultimato il lavoro di annotazione e di strutturazione del thesaurus sarà possibile interrogare ed accedere ai testi utilizzando la lingua in cui questi sono stati scritti, anche partendo da una lingua diversa da quella originale. I termini sono stati infatti implementati in un database relazionale MySQL all'interno del quale sono interrelati secondo la relazione di equivalenza interlinguistica (cfr. fig. 5).

Visto il carattere pilota di The*of*Philo[10], se gli equivalenti interlinguistici sono già disponibili, le relazioni terminologiche intralinguistiche saranno sviluppate in successive fasi di lavoro. Verranno dunque generate triple RDF considerando i soggetti filosofici sia come soggetto, sia come oggetto di ciascuno *statement*, connessi dai seguenti predicati: Sinonimy (*IsSynonymOf, HasSynonym*); Homonymy (*IsHomonymOf*, *HasHomonym*); Hyperonymy

[9] Le principali fonti lessicografiche utilizzate per la selezione terminologica (in quanto provviste di un sistema di entrate multilingue) sono: N. Abbagnano, *Dizionario di Filosofia* (Torino 1998); A. Lalande, *Vocabulaire technique et critique de la philosophie* (Paris 1983); S. Maso, *Lingua Philosophica Graeca* (Milano-Udine 2010). A queste si aggiungono: A. Bailly, *Dictionnaire Grec- Français*, Édition revue par L. Séchan et P. Chantraine, Paris 1950; J. M. Baldwin, *Dictionary of Philosophy and Psychology*, Gloucester, Mass. 1960; B. Cassin, *Vocabulaire Européen des Philosophie*, Tours 2004; *Enciclopedia filosofica*, Roma 1979; L. Rocci, *Vocabolario Greco-Italiano*, Perugia 1993; Liddell-Scott, *Greek- English Lexicon*, Rev. by H. S. Jones, Oxford 1968; T. Sanesi, *Vocabolario Italiano-Greco*, Pistoia-Siena 1916.

[10] Il thesaurus è il risultato di uno specifico progetto di ricerca al quale lavorano al momento Antonio Lamarra, Michela Tardella e Ada Russo. L'accesso al database non è ancora pubblico, ma ristretto al personale ILIESI. Può essere tuttavia visionato su richiesta.

Fig. 5. Output di una query lanciata in The*of*Philo.

(*IsHyperonymOf, HasHyperonym*); *Hyponymy (IsHyponymOf*, *HasHyponym*); Co-Hyponymy (*IsCo-HyponymOf, HasCo-Hyponym*); Antonymy (*IsAntonymOf*, *HasAntonym*). Una volta ultimata l'attività di annotazione semantica, The*of*Philo consentirà agli studiosi che lanciano una ricerca (query) relativa ad un Subject l'accesso diretto ai testi inerenti, secondo le relazioni stabilite, al Subject stesso. Inoltre si potrà approfondire la ricerca e lo studio della terminologia e dei testi secondo le relazioni interlinguistiche ed intralinguistiche ad essi correlate.

È opportuno sottolineare che TheofPhilo, oltre ad essere uno strumento di IR, costituisce di per sé un interessante oggetto di ricerca: può essere infatti considerato come un consistente corpus di termini filosofici, interrelati a livello inter ed intralinguistico; si presta dunque ed essere studiato ed analizzato dal punto di vista tanto linguistico quanto storico-filosofico.

5. Conclusioni

A conclusione di questa breve presentazione riteniamo opportuno proporre alla riflessione alcuni temi nati dal lavoro integrato di teoria e pratica condotto in questi anni. Ci soffermiamo innanzitutto sul tema dell'*interazione/relazione*, intesa come collaborazione tra umanisti e tra umanisti ed informatici. A questo riguardo menzioniamo almeno due casi: l'interessante collaborazione sviluppata tra linguisti e filosofi durante il lavoro condotto sui testi pubblicati in Daphnet, collaborazione che si è rivelata particolarmente costruttiva nell'attività di annotazione semantica, e l'interazione tra umanisti e informatici, che ha accompagnato l'intero processo di costruzione delle piattaforme e dei loro strumenti, in un continuo dialogo e scambio di competenze e obiettivi.

Nel primo caso è stata particolarmente fruttuosa la riflessione condotta sulla scelta delle relazioni utili a stabilire connessioni tra testi e tra testi e concetti. Si è trattato infatti di entrare nel processo di elaborazione di nuova conoscenza fruibile da altri studiosi o anche da studenti che utilizzano le piattaforme. Nel secondo caso è aumentata la consapevolezza che l'uso degli strumenti informatici retroagisce sulla riflessione teorica e viceversa in un rapporto di reciproca influenza. Ciò avviene in relazione a due aspetti: in primo luogo la modifica degli strumenti, fatta per rispondere ad esigenze che si presentano nel corso del lavoro di ricerca (per esempio l'aggiunta di funzioni nel caso in cui ne manchino). In secondo luogo la riflessione critica sul testo stesso: la costruzione di una tripla impone, per esempio, un ragionamento intorno alle relazioni inter ed intratestuali, che induce alla formulazione di ipotesi critiche sul testo e alla conseguente necessità di poter rappresentare queste relazioni attraverso la costruzione di reti i cui nodi sono costituiti dalle relazioni RDF, attivando diverse tipologie di grafi.

Questo insieme di considerazioni è quanto emerge da un cambiamento nell'approccio culturale (Ciula 2013) e strumentale sia nell'uso delle tecnologie sia nelle modalità di approccio ai testi e alle funzionalità di interrogazione e di accesso che si è attuato nell'articolato lavoro di ricerca condotto sulle piattaforme del portale Daphnet.

Sebbene questa non sia la sede più appropriata per discutere di questo cambiamento "antropologico", di cui probabilmente tutti coloro che sono coinvolti in un lavoro nell'ambito dell'"umanistica digitale" sono già consapevoli (e che richiede, a nostro parere, ancora più approfondite tematizzazioni), vorremmo tuttavia sottolineare, in conclusione, l'importanza della

collaborazione intesa come efficienza, armonia e dinamicità *versus* produzione di mere infrastrutture e servizi statici. Vorremmo inoltre affermare la convinzione che il lavoro congiunto e dialettico tra teorie e pratiche di umanisti e informatici può rivelarsi un matrimonio felice (sebbene complesso come tutti i matrimoni) che apre alla possibilità di intervenire positivamente e consapevolmente nella dinamica dell'*effetto looping* (McCarty 2013).

6. Bibliografia

Andrews P., Zaihrayeu I., Pane J. (2011). *A classification of semantic annotation systems*. «Semantic Web Journal», vol. 0, pp. 1-27. URL=http://www.semantic-web-journal.net/sites/default/files/swj123_0.pdf [ultima visita 16/04/2014].

Ciula A. (2013). *Which Changes are Currently Taking Place in our Research and Academic Culture?*. In *International colloquium: Research Conditions and Digital Humanities: What are the Prospects for the Next Generation?*, German Historical Institute, Paris, June 2013. URL=http://www.slideshare.net/arimare/presentation-ciulaparis2013 [ultima visita 16/04/2014].

Eide Ø. (2013). *Ontologies, data modelling, and TEI*. In *TEI Conference*. URL=http://digilab2.let.uniroma1.it/teiconf2013/program/papers/abstracts-paper#C107 [utima visita 16/04/2014].

Garshol L.M. (2004). *Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all*. URL=http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N773 [ultima visita: 16/04/2014].

Gold M.K., a c. di (2012). *Debates in Digital Humanities*. University of Minnesota Press. URL= http://dhdebates.gc.cuny.edu/ [ultima visita 16/04/2014].

Grenon P., Smith B. (2009). *Foundations of an ontology of philosophy*. «Synthese», pp. 185–204.

Grassi M. *et al.* (2013). *Pundit: augmenting web contents with semantics*. «Literary and Linguistic Computing», vol. 28, no 4, pp. 640-659.

Gruber T.R. (1993). *A translation approach to portable ontology specification*. «Knowledge Acquisition», vol. 5, no 2, pp. 199-220.

Magris, M. *et al.* (2002). *Manuale di terminologia*, Hoepli.

Marras C., Lamarra A. (2013). *Scholarly Open Access Research in Philosophy: Limits and Horizons of a European Innovative Project*. In *Digital Humanities International Conference*, University of Nebraska-Lincoln. URL=http://dh2013.unl.edu/abstracts/ab-316.html [ultima visita 16/04/2014].

McCarty W. (2013). *What does Turing have to do with Busa?*. In F. Mambrini, M. Passarotti, C. Sporleder, a c. di, *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities* (ACRH-3), Sofia, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, pp. 1-14.

URL=http://www.mccarty.org.uk/essays/McCarty,%20Turing%20and%20Busa.pdf [ultima visita 18/04/2014].

Reaner A. (2013). *Text encoding, ontologies, and the future*. In *TEI Conference* (Keynote). URL=http://digilab2.let.uniroma1.it/teiconf2013/program/keynotes/ [ultima visita 16/04/2014].

Signore O. (2002). *Rappresentare la conoscenza con Resource Description Framework.* In *Knowledge Management (V edizione)- Forum Proceedings on the knowledge management in the organizations,* Milano, 28 marzo 2003, pp. 35-46. URL=http://www.w3c.it/papers/RDF.pdf [ultima visita 21/04/14].

Tardella M. (2013). *Agora. Scholarly Open Access Research in European Philosophy*. «Blityri. Studi di storia delle idee sui segni e le lingue», II/1, pp. 167-174.

Acquisizione e Creazione di Risorse Plurilingui per gli Studi di Filologia Classica in Ambienti Collaborativi*

Federico Boschetti
Istituto di Linguistica Computazionale "A. Zampolli", Cnr, Pisa, Italia
federico.boschetti@ilc.cnr.it

Abstract. Questo articolo illustra metodi e strumenti per l'acquisizione e l'estensione di risorse digitali plurilingui per gli studi classici, sviluppati in collaborazione tra il CoPhiLab dell'Ilc-Cnr e il Perseus Project della Tufts University. Si descrivono tre linee di intervento: a) la progettazione e l'implementazione di un sistema di correzione dell'output dell'Ocr applicato al Greco antico; b) la creazione e la valutazione di un nucleo di *synsets* per Ancient Greek WordNet e c) l'allineamento di un campione di testi greci e latini con le relative traduzioni italiane.

Keywords: Greco antico, Ocr, WordNet, Allineamento.

1. Introduzione

Lo sviluppo delle attività illustrate in questo articolo è cominciato durante la visita di un semestre alla Tufts University,[1] allo scopo di potenziare i corpora greci e latini, le risorse linguistiche e gli strumenti messi a disposizione dal Perseus Project tramite strumenti e risorse sviluppate all'Istituto di Linguistica Computazionale "A. Zampolli" (Ilc-Cnr di Pisa).

Le attività si possono dividere in tre linee d'intervento:

- sviluppo di un proof-reader on-line per la correzione dell'Ocr applicato a testi greci;
- creazione del nucleo iniziale della WordNet di Greco antico;
- allineamento di testi greci e latini alle relative traduzioni in lingua italiana.

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. Cleup, Padova, 2014.

[1] La visita al Perseus Project - Tufts University, Medford, Ma (Dicembre 2012 - Giugno 2013) è stata possibile grazie ad un cofinanziamento Usa-Italia del Neh e del Cnr.

2. Aggregatore di informazioni per la correzione dell'Ocr e Proof-reader on-line

La prima linea d'intervento è focalizzata sull'acquisizione di testi greci. Come discusso in (Crane *et al*. 2006), (Stewart *et al*. 2007) e in (Boschetti *et al*. 2009), l'applicazione dell'Optical Character Recognition (Ocr) a edizioni critiche di testi greci è difficoltosa sotto molteplici punti di vista. I problemi principali riguardano il possibile danneggiamento delle pagine in edizioni datate, il largo numero di glifi che deve essere riconosciuto per il Greco politonico, la complessità dell'impaginazione e il contenuto plurilingue dell'apparato critico. Correntemente, molte migliaia di pagine relative ad edizioni critiche greche e latine sono state processate da B. Robertson presso la Mount Allison University (Nb, Canada)[2].

La stessa pagina è processata più volte con parametri di luminosità e contrasto diversi. Rigaudon, il software sviluppato da B. Robertson[3], incorpora un componente sviluppato presso il CoPhiLab (Ilc-Cnr) da F. Boschetti che seleziona il risultato più soddisfacente attraverso la valutazione delle parole riconosciute dallo spell-checker o, in caso di fallimento, riconosciute come sequenze sillabiche ben formate. Durante la visita al Perseus Project sono state sviluppate due applicazioni di post-processing: l'aggregatore di informazioni per la correzione dell'Ocr e il proof-reader on-line.

L'aggregatore

L'output dell'Ocr ottenuto sulla grid canadese da B. Robertson è passato all'aggregatore, che mette insieme i dati necessari a facilitare la correzione manuale. Sia l'output originale che il risultato dell'aggregazione sono codificati nel microformato hocr[4], che incorpora informazioni relative alla mappatura del testo sull'immagine, al grado di confidenza della lettura dell'Ocr e alle possibili letture alternative fornite dallo spell-checker. Le risorse linguistiche e testuali messe a disposizione dell'aggregatore sono costituite da repertori di parole flesse, dalla lista delle sillabe greche e da una collezione di testi precedentemente digitalizzati. Il reportorio delle forme flesse relative

[2] I risultati sono disponibili all'indirizzo http://heml.mta.ca/rigaudon/catalog. Tutti i link sono stati verificati il 22 Aprile 2014.
[3] Disponibile all'indirizzo https://github.com/brobertson/rigaudon
[4] Si veda http://code.google.com/p/hocr-tools

alle lingue classiche è basato su Morpheus, l'analizzatore morfologico sviluppato presso la Tufts University (Crane 1991), e sulla lista completa delle forme flesse che occorrono nel corpus di testi della Perseus Digital Library: il database delle analisi morfologiche di Morpheus contiene alcune varianti morfologiche non presenti nel corpus testuale e la collezione di testi contiene nomi propri non presenti nel database morfologico, quindi le risorse sono complementari.

Quando l'output dell'Ocr è valutato, sequenze testuali riconosciute come forme flesse presenti nei repertori appena indicati sono considerate sequenze riconosciute correttamente, anche se la probabilità che un errore di Ocr corrisponda casualmente ad una parola flessa varia da lingua a lingua. Le parole non individuate nel database sono testate per valutare la natura del possibile guasto. Se una sequenza di testo, trasformata da caratteri minuscoli in caratteri maiuscoli, si trova nel repertorio delle forme maiuscole, allora l'errore di Ocr ha un'alta probabilità di essere dovuto a un inadeguato riconoscimento di spiriti e accenti, cioè ad un tipo di errore che è neutralizzato nella versione in maiuscolo della sequenza analizzata. Se anche questo test fallisce, la sequenza testuale è divisa in sillabe usando il sillabatore (hyphenator) del Greco antico e il sistema valuta se la sequenza sillabica è compatibile con la lingua in oggetto. Ad esempio in Greco αὐ-κα-τϙι-μός è una sequenza sillabica ben formata (anche se è soltanto una pseudo-parola), mentre κα-τϙι-αὐ-μός è una sequenza sillabica malformata, perché la sillaba αὐ può apparire soltanto all'inizio di una parola (secondo il repertorio delle sillabe greche, basato sul corpus dei testi a disposizione). Una sequenza sillabica ben formata nella maggior parte dei casi è una parola riconosciuta correttamente dall'Ocr ma non ancora presente nei repertori delle forme flesse, ad esempio perché è una variante morfologica rara o un nome proprio poco frequente. Sequenze di caratteri casuali (cioè che non superano alcuno dei precedenti test) sono considerate meri errori. L'aggregatore, dopo l'identificazione e la classificazione dei possibili errori, associa a ciascuno di essi la lista dei suggerimenti prodotta dallo spell-checker.

Quando una differente edizione della stessa opera è disponibile, l'output dell'Ocr è allineato ad essa parola per parola, usando l'algoritmo di Needleman-Wusch. Le parole dell'edizione precedentemente digitalizzata, quando sono allineate a possibili errori di Ocr, sono aggiunte all'inizio della lista dei suggerimenti forniti dallo spell-checker. Secondo la tipologia dei possibili errori, se l'*edit distance* (cioè il numero di operazioni di inserimento, sostituzione e cancellazione necessarie a trasformare una stringa in un'altra)

tra il suggerimento e il possibile errore di OCR è di una sola operazione e la tipologia del possibile errore è relativa a spirito o ad accento, si procede alla sostituzione automatica. Nella grande maggioranza dei casi, una differente edizione della stessa opera fornisce la parola adeguata a correggere il testo, ma in alcuni casi l'output dell'OCR contiene la soluzione corretta non riconosciuta dallo spell-checker (ad es. perché è una variante rara) e l'altra edizione contiene una *lectio facilior* riconosciuta dallo spell-checker. Questi casi possono creare contaminazioni indesiderate tra edizioni.

L'applicazione web di supporto alla correzione manuale

Il proof-reader on-line fornisce agli utenti un ambiente collaborativo per la correzione basato sull'output dell'aggregatore appena illustrato. L'output dell'OCR e i relativi testi corretti sono contenuti in un database centralizzato per facilitare la gestione dell'intero processo di correzione. L'interfaccia web del proof-reader è ispirata ad *hocr editor*, un plug-in per Firefox sviluppato da J. Garrison[5], che attualmente non è più aggiornato. L'interfaccia web fornisce all'utente una lista di coppie costituite dall'immagine di una linea di testo e dal relativo testo digitalizzato prodotto dall'OCR che deve essere corretto, come si vede in fig. 1. Il sistema usa le coordinate di ciascuna linea di testo che sono incorporate nel file html, contenente il microformat hocr.

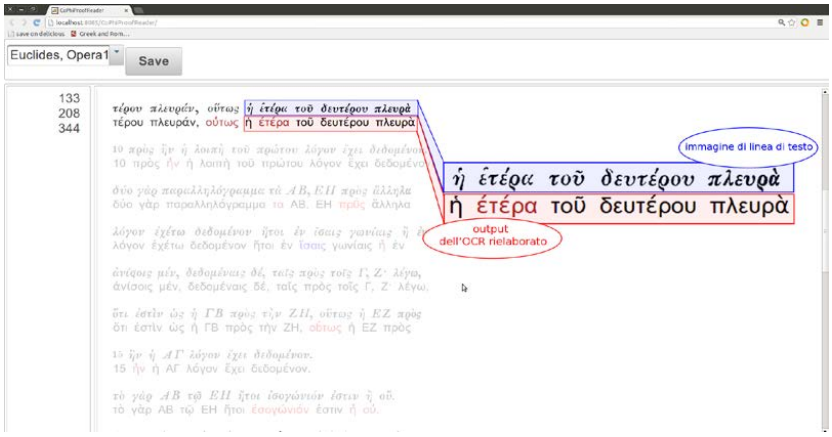


Fig. 1. Coppie costituite dall'immagine di una linea di testo e dal relativo output dell'OCR.

[5] Disponibile all'indirizzo https://addons.mozilla.org/it/firefox/addon/hocr-editor

Errori ed auto-correzioni (secondo la strategia illustrata nella sezione precedente) sono evidenziati con colori diversi, come mostrato in fig. 2, al fine di catturare l'attenzione dell'utente su diversi tipi di intervento. In particolare, le autocorrezioni devono essere controllate accuratamente per evitare il rischio di contaminazione fra edizioni.

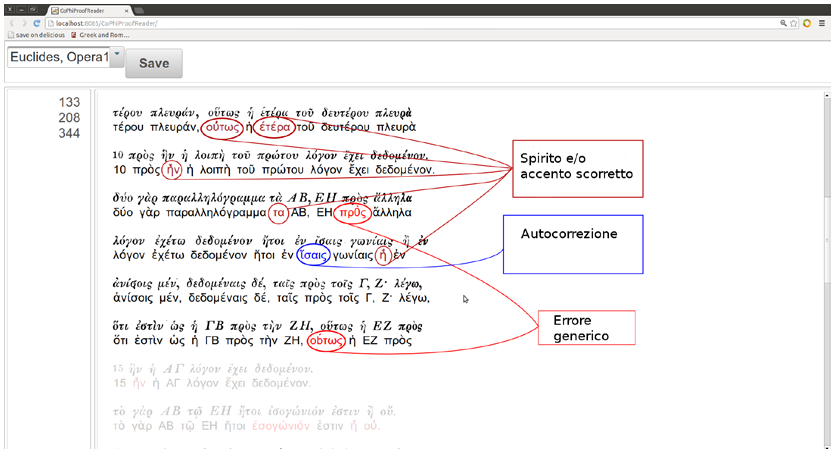


Fig. 2. Colorazione dei possibili errori (o auto-correzioni).

L'applicazione web per il proof-reading e stata testata sia da volontari che da professionisti di ditte di data entry con risultati positivi[6].

3. AncientGreekWordNet

La seconda linea di ricerca è focalizzata sulla creazione di strumenti lessicali per lo studio di testi greci e latini, le loro traduzioni italiane e possibili relazioni con documenti in altre lingue (in particolare Inglese e Arabo). Negli ultimi decenni, seguendo il modello dell'originale WordNet per la lingua inglese sviluppata presso l'Università di Princeton (Fellbaum 1998), lessici strutturati semanticamente sono stati sviluppati per altre lingue. Secondo

[6] Il codice sorgente è disponibile all'indirizzo https://github.com/CoPhi

il modello di WordNet, le relazioni semantiche (come la relazione genere/specie) e le relazioni lessicali (come la relazione di antonimia) sono strutturalmente distinte. I nodi concettuali sono associati alla glossa e sono interconnessi tramite molteplici relazioni, come l'iperonimia e l'iponimia o l'olonimia e la meronimia, in modo da formare una rete concettuale. Le parole sono organizzate in synsets, cioè in liste di sinonimi (ad es., in Inglese, [tool, instrument]) associate al nodo concettuale di pertinenza. In caso di polisemia, la stessa parola appartiene a diversi synsets (ad es. *horse* come animale e *horse* come strumento ginnico, cavallina).

In questo modo, la stessa rete concettuale può essere condivisa da lingue differenti, anche se sono necessarie strategie di adeguamento delle relazioni. In Italia, tre risorse rilevanti per i nostri scopi sono state sviluppate nel recente passato:

- LatinWordNet è stata sviluppata presso l'Università di Verona (Minozzi 2008) con più di 8.000 synsets;
- ItalWordNet (Roventini e*t al.* 2000) è stata sviluppata da M. Monachini e dal suo gruppo di lavoro presso l'ILC-CNR;
- MultiWordNet (Pianta *et al.* 2002) è stata sviluppata presso la Fondazione Bruno Kessler (FBK) di Trento;

In stretta collaborazione con l'Alpheios Project[7], partner principale del Perseus Project, synsets greci e latini sono stati estratti da dizionari bilingui disponibili in formato digitale e i risultati sono stati collegati alle altre wordnets a disposizione. L'algoritmo è basato sul principio che termini sinonimi nella lingua d'origine sono per lo più tradotti con gli stessi termini nella lingua di destinazione; in questo modo molti sinonimi greci (o latini) possono essere raggruppati perché condividono le stesse traduzioni. Ad esempio, πόντος, θάλασσα, ἅλς e πέλαγος sono tradotti con "sea" e per questa ragione sono automaticamente raggruppati nello stesso synset. Il problema principale sorge quando il termine usato nella traduzione della lingua di destinazione è polisemico, perché questo fa raggruppare nello stesso synset termini che non hanno relazione semantica nella lingua d'origine. Ad esempio, in Inglese "sound" è un termine altamente polisemico e significa, fra gli altri sensi, "auditory sensation", "strait" e "healthy"; per questa ragione parole greche come ψόφος (noise), φωνή (sound, tone); σόος (safe), ἄνοσος (without sickness); στενόχωρος (strait) sono collassate nello stesso synset. H. Diakoff ha estratto i synsets greci dal Liddell-Scott Jones e dal Middle

[7] Si veda http://alpheios.net

Liddell. Usando questo approccio, 34.925 lemmi di Greco su 130.000 sono stati distribuiti in 33.910 synsets[8].

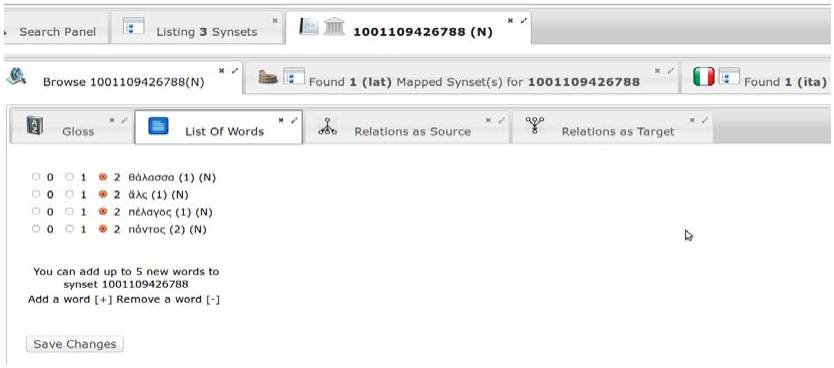


Fig. 3. L'interfaccia web di AncientWordNet.

Durante uno stage all'ILC-CNR, lo studente Y. Bizzoni, supervisionato da M. Monachini e F. Boschetti, ha validato un campione di synsets greci (1.013 su 33.910) al fine di valutare le prestazioni del sistema e iniziare la correzione degli errori. Come mostrato in tab. 1., 84 synsets su 1.013 (8.3%) sono stati disattivati a causa di un'erronea associazione a concetti moderni alieni all'antichità, come ad esempio "a series of linked atoms (generally in an organic molecule)". Tale glossa era stata associata dal sistema automatico a ὅρμος, ἅλυσις, σύσφιγμα, ὁρμαθός (termini riconducibili al concetto di catena).

14 synsets su 1.013 (1,4%) sono stati marcati come "near to the concept expressed by a definition that needs adjustments". Questi casi sono molto interessanti perché mostrano chiaramente la distanza fra *Sinn* e *Bedeutung*, nell'accezione di Frege. Ad esempio, il concetto associato a γῆ, γαῖα, è glossato in Princeton WordNet nel modo seguente: "the 3rd planet from the sun". La *Bedeutung* di γαῖα è chiaramente il nostro pianeta, ma il *Sinn* che

[8] Attualmente è in avanzato stadio di sviluppo da parte di R. Del Gratta un'interfaccia web per l'interrogazione e l'editing della risorsa (si veda fig. 3), accessibile all'indirizzo http://www.languagelibrary.eu/new_ewnui. Maggiori dettagli si possono trovare in (Bizzoni *et al.* 2014).

definisce il concetto è legato ad uno specifico paradigma scientifico[9] (tolemaico o copernicano).

I 1.013 synsets validati corrispondono a 6.457 sensi, vale a dire parole eventualmente ripetute in più synsets, con significati diversi. Come mostrato in tab. 2., 3.479 sensi su 6.457 (53,9%) sono stati accettati dal valutatore, cioè dallo studente stagista, mentre 2.101 sensi (32,5%) sono stati rifiutati e 877 (13,6%) sensi sono stati considerati incerti. I sensi incerti nella maggior parte dei casi hanno una relazione semantica con il concetto a cui sono stati associati automaticamente, ma una relazione diversa dalla sinonimia (ad es. iperonimia, meronimia, etc.).

Al fine di confrontare questo nuovo approccio agli studi tradizionali sulla sinonimia del Greco antico, l'indice della *Synonymik der Griechischen Sprache* è stato digitalizzato, assegnando 4.123 termini a 150 campi semantici analizzati da (Schmidt 1876). Un caso significativo merita di essere discusso. Il synset n#02818832, glossato con "a piece of furniture that provides a place to sleep", ha quattro corrispondenze nel cluster di Schmidt #25: κοίτη, λέχος, εὐνή e λέκτρον, con il significato di "bed", ma tale cluster contiene anche altri termini, giudicati dal valutatore come semanticamente correlati tramite una relazione diversa dalla sinonimia: θαλάμη (lair), φυλλάς (leafy bed), ὑπόστρωμα (litter) e κράββατος (small bed for poor people). φυλλάς, κράββατος e ὑπόστρωμα sono iponimi di "bed", perché hanno un significato più specifico. Ma θαλάμη (lair) non può essere subordinato o sovraordinato al concetto di "bed" nella struttura gerarchica della WordNet di Princeton, perché "bed" è un oggetto e "lair", "lurkingplace" sono luoghi. Schmidt raggruppa insieme a sinonimi anche termini solo etimologicamente correlati, come λέχος (bed) e ἄλοχος (bride), che sono totalmente distinti in AncientWordNet. Verbi collegati al concetto di "letto" in Schmidt si trovano nello stesso cluster: ἀωτέω, βαυβάω, δαρθάνω, εὕδω, ἰαύω, καθεύδω, καταδαρθάνω, κνώσσω, νυστάζω, mentre in AncientGreekWordNet questi termini sono associati al synset n#14024882, glossato come "a natural [...] state of rest [...]", che contieneanche nomi quali κοίμημα, ὕπνος, ὦρος (sleep), κῶμα (deep sleep), etc. Si è deciso di ammettere verbi nei synsets nominali, se il concetto è riconducibile ad un *nomen actionis*: infatti l'infinito di un verbo è equivalente ad un *nomen*

[9] Maggiori dettagli relativi alla costruzione di risorse lessico-semantiche in prospettiva diacronica si possono trovare in (Khan *et al.* 2014).

actionis. Gli stessi verbi appartengono anche al synset verbale v#00014742, glossato come "to be asleep".

Tramite il collegamento dei synsets greci e latini estratti dai dizionari bilingui a ItalWordNet (integrata con la sezione italiana di MultiWordNet), si è in grado di fornire una nuova risorsa bilingue, che associa a parole greche o latine la lista di parole italiane che si suppone siano adeguate per la traduzione. La lista dei sinonimi è seguita dalla lista degli iperonimi diretti, perché in molti casi un termine specifico, per ragioni stilistiche, può essere tradotto con un termine più generale. Ad esempio, πατήρ può essere tradotto sia con "padre" che con "genitore".

4. Allineamento di testi (originale e traduzione)

La terza linea di ricerca è focalizzata sull'allineamento tra gli originali in lingua greca o latina e le relative traduzioni in lingua italiana, seguendo la procedura adottata per l'allineamento delle opere originali con le traduzioni in lingua inglese, come illustrato in (Bamman *et al.* 2008). Le traduzioni in Italiano provengono da risorse eterogenee disponibili on-line, ma in particolare da WikiSource[10].

L'allineamento è eseguito a vari livelli di granularità: sezione per sezione, enunciato per enunciato e parola per parola.

Se tanto l'originale quanto la traduzione sono strutturati gerarchicamente in libri, parti, capitoli etc., l'allineamento sezione per sezione è basato su semplici euristiche che identificano elementi di ancoraggio (*milestones*), forniti come informazione paratestuale nelle edizioni digitali. Se sono strutturati in forma dialogica, come i testi drammatici o le opere di Platone, l'allineamento delle sezioni deve tener conto di possibili incoerenze fra l'originale e la traduzione. Infatti, le traduzioni italiane a nostra disposizione abitualmente non sono basate sulle stesse edizioni dei testi originali digitalizzati presenti nella collezione del Perseus Project: per questa ragione, ad esempio, un lungo discorso nell'edizione del testo greco si può trovare diviso in tre diverse battute nella traduzione. In casi come questo, l'algoritmo usato per allineare i discorsi valuta non soltanto la sequenza degli interlocutori ma anche la lunghezza dei discorsi, al fine di ottenere la corrispondenza ottimale.

[10] Si veda http://www.wikisource.org

L'allineamento sezione per sezione è utile soprattutto per l'annotazione libera di testi paralleli. La fig. 4. mostra l'*incipit* dell'*Agamemnon* di Eschilo fruibile tramite *Aporia*, l'applicazione web sviluppata presso l'ILC-CNR[11].

Una o più parole possono essere selezionate in una delle due lingue (originale o traduzione) o in entrambe. Sequenze testuali di una o più parole possono essere selezionate e annotate con marcatori predefiniti (ad es. relativi all'analisi retorica) oppure commentate liberamente, in modo non strutturato.

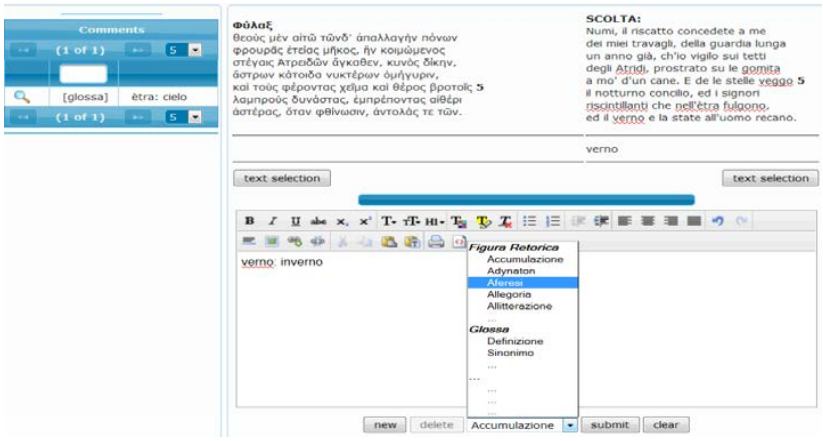


Fig. 4. Il sistema di annotazione *Aporia*.

Il sistema implementato da D. Bamman ed esteso da B. Almas presso il Perseus Project riceve in input testi divisi in sezioni con riferimenti incrociati, grazie agli elementi di ancoraggio. Per migliorare le prestazioni del sistema di allineamento, il testo deve essere lemmatizzato, anche se, limitatamente a tale scopo, non è necessaria un'elevata accuratezza della lemmatizzazione. Per il Greco e il Latino la lemmatizzazione è stata eseguita con Morpheus, mentre per l'Italiano la lemmatizzazione è stata eseguita con il lemmatizzatore (oltre a PoS tagger e analizzatore sintattico) sviluppato da F. Dell'Orletta presso l'ILC-CNR, come illustrato in (Dell'Orletta *et al.* 2007) e (Dell'Orletta

[11] *Aporia* è stata sviluppata da A.M. Del Grosso e F. Boschetti ed è accessibile on-line con autenticazione all'indirizzo http://cophidev.ilc.cnr.it:8080/Aporia_Wapp

2009). Testo e relativa traduzione sono processati da una sequenza (*pipe*) di scripts che incorpora l'allineatore di enunciati illustrato in (Moore 2002) e l'allineatore di parole/sintagmi MGIZA++, illustrato in (Gao-Vogel 2008).

Anche se il sistema aumenta le sue prestazioni con quantità crescenti di testi paralleli, ha già prodotto risultati apprezzabili su una piccola quantità di opere usate per testarlo. In particolare, il sistema è stato testato sulle *Historiae* di Erodoto per il Greco e sul *De divinatione* di Cicerone per il Latino. Come mostrato in fig. 5, il sistema non solo è in grado di allineare correttamente parole singole, come *medicis* – *medici*, *herbarum* – *erbe*, *oculorum* – *occhi*, *morbos* – *malattie*, ma anche singole parole corrispondenti a sintagmi complessi, come nel caso di *mirari* – *constatare con lieta meraviglia*.

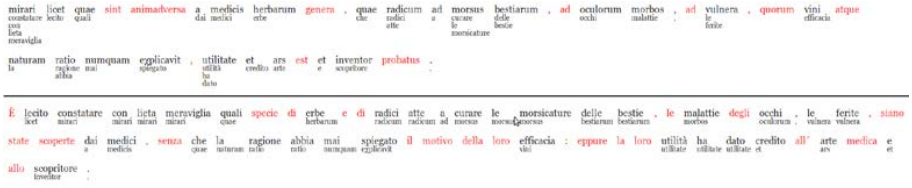


Fig. 5. Allineamento Latino/Italiano.

5. Conclusione

Lo scopo principale della visita alla Tufts University è stato il rafforzamento della collaborazione tra l'ILC-CNR e il Perseus Project, condividendo metodi e risorse necessarie allo sviluppo di nuovi strumenti per lo studio dei classici e la localizzazione in lingua italiana. I prodotti delle tre linee di ricerca illustrate in questo contributo necessitano miglioramenti, estensioni e correzioni ma sono mutualmente perfettibili grazie al fatto di essere strettamente correlate.

6. Bibliografia

Bamman D., Crane G. (2008). *Building a Dynamic Lexicon from a Digital Library*. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2008), Pittsburgh, Pennsylvania, ACM Digital Library, 2008-06.

Bizzoni Y., Boschetti F., Del Gratta R. Diakoff H., Monachini M., Crane G. (2014). *The Making of Ancient Greek WordNet*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14).

Boschetti F., Romanello M., Babeu A., Bamman D., Crane G. (2009). *Improving OCR Accuracy for Classical Critical Editions*. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas, a c. di, *Research and Advanced Technology for Digital Libraries*, Proceedings, Springer, pp. 156-167.

Crane G. (1991). *Generating and Parsing Classical Greek*. «Literary and Linguistic Computing», vol. 6 , no 4.

Crane G., Jones A., Bamman D., Cerrato L., Mimno D., Packel D., Sculley D., Weaver G. (2006). *Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries*. In Proceedings of Research and Advanced Technology for Digital Libraries: 10th European conference, ECDL 2006, Alicante, Spain, September 17-22, pp. 353-366.

Dell'Orletta F., Federico M., Montemagni S., Pirrelli V. (2007). *Maximum Entropy for Italian POS Tagging*. In Proceedings of Workshop Evalita 2007. «Intelligenza Artificiale» vol. 4, no 2.

Dell'Orletta F. (2009). *Ensemble system for Part-of-Speech tagging*. In Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, December.

Fellbaum C. (1998). *WordNet: An Electronical Lexical Database,* The MIT Press.

Gao Q., Vogel S. (2008). *Parallel implementations of word alignment tool*. In Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing (Setqa-Nlp '08), Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 49-57.

Khan A.F., Boschetti F., Frontini F. (2014). *Using* lemon *to Model Lexical Semantic Shift in Diachronic Lexical Resources*. In Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014).

Minozzi S. (2008). *La costruzione di una base di conoscenza lessicale per la lingua latina: LatinWordnet*. In G. Sandrini, a c. di, *Studi in onore di Gilberto Lonardi*, Fiorini Editore, pp. 243-258.

Moore R.C. (2002). *Fast and Accurate Sentence Alignment of Bilingual Corpora*. In S.D. Richardson, a c. di, AMTA 2002, LNAI 2499, Springer-Verlag, pp. 135-144.

Pianta E., Bentivogli L., Girardi C. (2002). *MultiWordNet: developing an aligned multilingual database*. In Proceedings of the First International Conference on Global WordNet, Mysore, India, January 21-25.

Roventini A., Alonge A., Calzolari N., Magnini B., Bertagna F. (2000). *ItalWordNet: a Large Semantic Database for Italian*. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, 31 May – 2 June 2000, Volume II, Paris, The European Language Resources Association (ELRA), pp. 783-790.

Schmidt J.H.H. (1876). *Synonymik der griechischen Sprache*, Teubner.
Stewart G., Crane G., Babeu A. (2007). *A New Generation of Textual Corpora: Mining Corpora from Very Large Collections*. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2007), Vancouver, British Columbia: ACM Digital Library, pp. 356-365.

Da *Musisque Deoque* a *Memorata Poetis*
Le vie della ricerca intertestuale*

Paolo Mastandrea, Luigi Tessarolo
Dipartimento di Studi Umanistici, Università Ca' Foscari, Venezia, Italia
mast@unive.it, luigi.tessarolo@fastwebnet.it

Abstract. Le concordanze, gli onomastici, i volumi di indici tematici e le raccolte di luoghi simili a stampa sono ormai da qualche decennio sostituiti dall'interrogazione automatica dei grandi *corpora* digitalizzati. L'applicazione delle tecnologie informatiche in campo filologico non può limitarsi ad ampliare gli archivi, ma deve aprire nuove vie per rispondere a domande sempre più esigenti sulla storia, la circolazione e il riuso dei testi, antichi e moderni. Si dà qui una panoramica delle funzionalità di ricerca metrico-verbale in *Musisque Deoque* (http://www.mqdq.it) e delle componenti in sviluppo nella nuova risorsa dedicata alla ricerca tematica-semantica *Memorata poetis* (http://www.memoratapoetis.it).

Parole chiave: intertestualità formale e concettuale, memoria poetica, archivi di testi letterari.

1. Introduzione

Da quando – ormai una quindicina di anni orsono – iniziammo a caricare testi di poesia latina in rete, il fuoco della nostra attenzione era rivolto al raffinamento della 'ricerca intertestuale' – un affare che ha preso questo nome grazie a Julia Kristeva (1969)[1], ma può dirsi costituisse oggetto di pratica quotidiana sin dai primordi della disciplina filologica, nata negli ambienti della Biblioteca d'Alessandria, durante il regno di Tolomeo II Filadelfo e dunque verso la metà del III secolo a.C.

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. Cleup, Padova, 2014.

[1] La letteratura sulla metodologia e la pratica della intertestualità sta diventando smisurata, e quasi incontrollabile, giorno dopo giorno; rimando per questo alla recente sintesi di Neil Coffee, un latinista statunitense che si muove allineato sulla direzione e verso gli scopi che cerchiamo di perseguire coi nostri studi.

Il nostro interesse al problema corre su due piani complementari. Al primo livello ci si pone in linea di continuità con le vecchie concordanze a stampa, che agevolavano le ricerche dei singoli vocaboli seguendo elenchi alfabetici iniziali; ovviamente, sarebbe superfluo spiegare quale vantaggi ulteriori aggiunga l'interrogazione automatica nel reperimento di legami che coinvolgono gli abbinamenti, o le associazioni di più parole, in sequenza o a distanza, nell'ordine stesso o in quello inverso, oppure limitatamente a parti di esse: rimane pur sempre principio fondamentale dell'esplorazione l'esame di una stringa di caratteri alfabetici.

Al secondo livello si prosegue piuttosto l'antica pratica dell'indicizzazione tematica, a caccia di quella che Cesare Segre chiamava 'interdiscorsività'[2] – utile ad uno studio dei sintagmi diegetici: trame e intrecci del racconto, per lo più ma non esclusivamente in prosa. In un caso e nell'altro, si tratta di una continuazione della obsoleta *Quellenforschung*; la critica di impianto positivistico vi faceva corrispondere *grosso modo* due diverse tipologie di studi, che ad esempio rispetto alle fonti del capolavoro ariostesco produsse con Pio Rajna (*Le fonti dell''Orlando Furioso'*, Firenze 1876[1]) un capostipite dell'indagine condotta sulle genealogie tematiche, mentre Antonio Romizi (*Le fonti latine dell''Orlando Furioso'*, Torino 1896) si concentrava sugli apparati dei riferimenti puntuali alla tradizione letteraria, classica e italiana; o ancora, contrapporrei i saggi ricostruttivi di Eduard Norden (il grande commentatore di *Vergils Aeneis VI*, Leipzig 1903[1], ma soprattutto restauratore di parti degli *Annales* col suo *Ennius und Vergilius*, Leipzig 1915), a fronte delle tante raccolte di *loci similes*, così accuratamente redatte in quei decenni non solo da giovani aspiranti filologi nelle loro dissertazioni, ma anche da accademici maturi e affermati in tutti i campi – e vorrei citare il repertorio di *Scripture and Classical Authors in Dante*, curato da Edward Moore e uscito per la prima volta ad Oxford nel 1896. C'è da restare stupiti di fronte all'ampiezza della documentazione, in rapporto ai mezzi allora disponibili: eppure la reazione di matrice idealistica ebbe allora la capacità di stroncare un intero filone di studi[3].

Ma torniamo all'oggi, magari guardando al domani. Attualmente si procede con pari interesse – se non uguale distribuzione di forze – su due linee

[2] Segre 1982.

[3] Per un inquadramento storico e bibliografico del dibattito avviato in Italia dal saggio del Croce su *La ricerca delle fonti* (in *Problemi di Estetica*, Bari, Laterza, 1910) si veda Pasini 1988, pp. 7-30.

direttrici: non parallele, anzi abbastanza divergenti fra loro, ma promettenti entrambe di buoni risultati

2. La ricerca metrico-verbale estesa alle varianti dei testi latini in *Musisque Deoque*[4]

Alle caratteristiche iniziali – e più volte illustrate, anche in sede di convegni e relativi atti a stampa[5] – altre non secondarie funzioni si sono aggiunte negli ultimi mesi. Alludiamo innanzi tutto alla integrazione con http://www.pedecerto.eu, programma di metrica latina digitale, le cui risorse sono ancora da sfruttare sia al fine di condurre ricerche automatiche di forme / lemmi con analoga collocazione nel verso, sia per disambiguare su base prosodica nei testi le occorrenze di forme che possono appartenere a più lemmi.

Introdotta più di recente, ancora sotto collaudo, è la funzione http://www.mqdq.it/cooccorrenze.jsp, che riesce a confrontare un testo sorgente, a scelta del ricercatore, con l'intero corpus di *Musisque Deoque*, oppure con una sua parte; al momento le restrizioni del target consentite sono le seguenti: 1) solo autori precedenti; 2) solo autori successivi; 3) solo metri dattilici (default); però sarà presto possibile limitare il target ad uno o più autori / opere.

Il confronto consiste nella ricerca delle cooccorrenze di coppie di parole del testo sorgente nei testi target, con le seguenti opzioni selezionabili dall'utente:

- distanza tra le due parole, misurata come numero di parole interposte: da 0 a 5 (default 1);
- ricercare le parole solo nella stessa sequenza o anche in ordine inverso (default);
- escludere o meno termini molto frequenti (sono contenuti in una lista di 121 voci, al momento non modificabile, ma lo si potrebbe fare);
- ricercare o meno (default) anche le varianti in apparato;
- ricercare solo le forme esatte oppure cercare per lemma (default).

[4] Allocati negli archivi e interrogabili attraverso motore di ricerca al sito: http://www.mqdq.it
[5] Da ultimo, in breve: *Text Retrieval on Critical Editions*, a cura di M. Manca, Linda Spinazzè, F. Boschetti oltre a chi scrive, in *Workshop on Annotation of Corpora for Research in the Humanities* (Heidelberg, Jan. 5, 2012), pp. 1-12 = «Journal for Language Technology and Computational Linguistics» 26, 2011, pp. 129-140.

Per quanto riguarda lo score, l'assegnazione di un punteggio a ciascuna occorrenza non ha lo scopo di individuare i migliori risultati, in maniera da farli affiorare e presentarli allo studioso per primi (un po' come fa Google), quanto piuttosto di eliminare i peggiori. Il fatto è che anche un breve componimento produce, se messo a confronto con l'intero corpus poetico (diverso sarebbe il confronto tra due singoli autori o opere), un numero molto elevato di risultati (mediamente assai più di 100 per singolo verso, con i parametri di default) e i criteri adottati per giudicarli non hanno una grana abbastanza fine da riuscire a portarne in superficie una quantità sufficientemente ristretta da costituire una guida per lo studioso. In altri termini, se il testo sorgente non è proprio di dimensioni minime, le occorrenze a punteggio massimo sono comunque molte centinaia o migliaia, e il nostro algoritmo non ha la pretesa di effettuare tra esse una più fine discriminazione. Perciò, in attesa di realizzare uno strumento di maggiore capacità risolutiva (e intravediamo la possibilità di raggiungere un tale risultato ricorrendo a confronti metrici fini, mettendo a frutto, come si diceva, le opportunità di *Pede certo*), abbiamo ritenuto più vantaggiosa per lo studioso una esposizione dei risultati di tipo tradizionale, cioè nell'ordine dei versi del testo sorgente, sfruttando comunque lo scoring per rendergli la vita più facile eliminando le occorrenze meno significative.

Lo strumento è pensato in modo speciale per i testi poetici; questo vale particolarmente per i criteri della ricerca per lemmi e per quelli dello scoring. Nei dettagli, la ricerca per lemmi non è estesa a tutte le forme flesse che fanno capo al lemma, ma solo a quelle con lo stesso numero di sillabe della forma sorgente. Per converso la query può estendersi pure ad altri lemmi, nel caso di voci composte che differiscano per il solo prefisso.

I criteri per l'assegnazione di un punteggio all'occorrenza sono i seguenti:

1) identità delle forme: 2 punti per parola;
2) stessa sequenza delle due parole: 2 punti;
3) stessa distanza tra le parole nel testo sorgente e nel target: 2 punti
4) posizione nel verso: è assegnato 1 punto per ciascuna parola che occupa, nei testi confrontati, la stessa posizione significativa (prima, seconda, penultima, ultima); sono dati complessivamente 2 punti anche nel caso in cui una sola parola condivida la posizione, ma sia uguale la distanza tra le parole.

Si trovano ancora in fase sperimentale, ma non proprio aurorale, le ricerche da effettuare mediante http://www.mqdq.it/cometri.jsp, predisposte alle funzioni:

1) Cerca lo schema del verso (tutte le pause, sinalefe, iato)
2) Cerca lo schema del verso (solo pause principali)
3) Cerca le parole nella posizione metrica
4) Cerca sequenze di 4/5 sillabe

Le prime due cercano i versi che riproducono per intero lo schema (quantità e pause) del verso sorgente, senza alcun riguardo alle parole o sillabe interessate.

Le funzioni 3 e 4 sono invece intese ad individuare riprese di sequenze metrico-verbali parziali, anche e soprattutto in assenza di una identità morfologica, basate perciò prevalentemente su una somiglianza di suono. Questi ultimi due approcci producono di fatto risultati parzialmente diversi e complementari, ma stiamo lavorando a combinare le due ricerche in un'unica operazione.

Intanto, con 3) si cercano le parole nella posizione metrica: da ciascun termine del verso sorgente è estratta una chiave di ricerca costituita dalla sua posizione metrica e dalle sole vocali, che viene poi reperita nel corpus; sono accolte solo le occorrenze che presentano la coincidenza di almeno quattro sillabe, in una o più parole; ma poi le occorrenze sono ordinate per rilevanza, sulla base del numero delle sillabe coincidenti, della contiguità delle parole trovate e della corrispondenza di consonanti.

Con 4) si cercano sequenze di quattro/cinque sillabe, in tal modo: si estraggono dal verso tutte le possibili sequenze di 4 o 5 sillabe, senza riguardo ai confini di parola; di ogni sillaba si considerano la posizione metrica e la vocale; dalla grande massa dei risultati si estraggono quelli che presentano un certo numero di corrispondenze anche nella parte consonantica delle sillabe.

3. La ricerca dei rapporti intertestuali di carattere semantico e tematico

Sin dalle origini della Filologia, cui si accennava, il transito e il reimpiego di elementi da un testo all'altro fu sempre visto come uno dei principali obiettivi di questa disciplina. Scopo fondamentale del progetto *Memorata Poetis*, che coinvolge studiosi di varie sedi universitarie e di altre istituzioni di ricerca nazionali (lo presentiamo qui, ma si vada anche direttamente al sito http://www.memoratapoetis.it) è la riunione, l'adattamento, il perfezionamento e per certa parte la creazione ex novo, di archivi digitali cui applicare strumenti elettronici di indagine capaci di analizzare testi plurilingui

con il fine di scorgerne le mutue riprese e relazioni. Si potrebbe così individuare ogni presenza (cosciente o incosciente) della memoria di un poeta all'atto della riscrittura di un altro, sicché quanto di solito è soltanto postulato, oppure anche documentato ma in modi occasionali ed estemporanei, trovi conferme oggettive – anzi, se è lecito usare il termine, – scientifiche'.

Il lavoro di esegesi, più o meno sistematica e completa dei testi, andrà così fiancheggiato, sostenuto e in più di un caso presumibilmente anticipato da una analisi fine dei dati, ottenuta incrociando due procedimenti: per un verso, nel tentativo di creare all'interno di una stessa lingua un thesaurus che permetta di eccedere i limiti della pura ricerca lessicale, esplorando i *corpora* testuali lemmatizzati con procedure (semi-)automatiche, comunque adeguate alle caratteristiche morfologiche, stilistiche, insomma linguistiche dei diversi testi (fatta salva una preventiva scelta di genere che garantisca livelli certi di ripetitività se non di formularità); basate sulla osservazione empirica per cui termini allocati in contesti simili hanno molto probabilmente significati simili o comunque correlati; nonché di stabilire relazioni reciproche dei termini fra lingue diverse, sfruttando dizionari bilingui greco-inglese, latino-inglese e arabo-inglese e grazie ad algoritmi probabilistici applicati ai contesti.

Dall'altro verso, la classificazione dei singoli documenti sulla base di tematiche, situazioni, ruoli e personaggi che popolano la produzione letteraria caratterizzata dalla maggiore brevità e concisione: l'epigrammatica, l'ecloga, l'elegia, la lirica e le corrispettive eredità (ad esempio in volgare italiano il sonetto di epoca tardomedievale, umanistica e seriore).

Si sta effettuando un censimento il più completo possibile di iscrizioni poetiche, antiche, medievali, umanistiche, moderne, e un trattamento di marcatura in vista della loro connessione a banche dati confluite già entro archivi elettronici di poesia latina, consultabili in rete http://www.mqdq.it e http://www.poetiditalia.it o su CD-Rom (*PoetriaNova*, 2da edizione, Firenze, SISMEL, 2010). Vi confluiscono materiali di tradizione diretta (da supporti materiali duri: pietra, bronzo, muro, ecc.) e indiretta (da manoscritti antologici e miscellanee, dall'interno di opere letterarie, ecc.), con relativi apparati critici, contesti di immagini, corredi di bibliografia. In tal modo il motore di interrogazione abbraccia una tradizione culturale più che millenaria con lo sguardo rivolto alla sopravvivenza delle tecniche espressive e stilistiche, curioso verso i meccanismi con cui agirono dall'interno i processi di riuso e attualizzazione; mirando ad una analisi combinata e raffinata del dettato e del portato del messaggio, offrendo ad una pluralità di studiosi i mezzi in-

vestigativi più opportuni – mezzi che divengono insuperabili o insostituibili quando siano applicati alla casistica dei prelievi non-consci degli autori dal deposito della memoria individuale.

Senza attendere tempi troppo lunghi, e men che meno la conclusione di un lavoro così imponente, sarà messo a disposizione della comunità scientifica, direttamente fruibile su web, un *corpus* di testi, per lo più ma non soltanto epigrafici, articolato per lingua (Latino, Greco, Italiano, Arabo) ed esteso in un arco cronologico dai primi documenti scritti dell'antichità ellenica ai secoli moderni; interrogabile sulla base di criteri storici, filologici, linguistici, stilistici; ed ancora, archeologici, paleografici, iconografici; sarà dunque offerta agli utenti della rete, attraverso appositi strumenti, la possibilità di individuare le correlazioni più latenti tra testi epigrafici e testi letterari. La marcatura previa dei testi darà luogo alla redazione di un lessico della letteratura epigrammatica, organizzato per aree semantiche, in sottoinsiemi che mettano in evidenza le relazioni analogiche esistenti fra gli elementi delle aree stesse; verrà così quasi automaticamente tracciata una mappa della circolazione o della migrazione, cronologica e topografica, dei temi e dei motivi individuati.

La componente più nettamente sperimentale del progetto comprende la messa a punto di strumenti e procedure informatiche in grado di rilevare in modo automatico (o semi-automatico) la presenza di temi e motivi all'interno delle composizioni poetiche, l'esplorazione degli spazi semantici grazie all'applicazione di metodi statistici al fine di identificare relazioni semantiche che emergono dalla cooccorrenza di termini in contesti simili, il confronto delle relazioni reciproche dei termini fra lingue diverse. I risultati attesi vanno dalla possibilità di ottenere una rappresentazione grafica bidimensionale delle distanze tra documenti epigrafici e documenti letterari, da cui poter inferire la probabilità che gli uni siano derivati dagli altri, o viceversa ne siano stati d'ispirazione (nel caso del confronto plurilingue è lecito attendersi analogamente che i documenti epigrafici processati nella lingua veicolare si dispongano sul piano cartesiano in modo da mostrare le relazioni intertestuali interlinguistiche), alla realizzazione di un metamotore di ricerca che travalichi sia i limiti lessicali dell'interrogazione verbale, sia i confini tra le culture (latina, greca, italiana) e i generi (letterario, epigrafico), ed eventualmente, oltre che operare su chiavi scelte di volta in volta dall'utente, possa essere impiegato per una scansione su larga scala dei testi al fine di riconoscere 'similitudini' fra i componimenti dei *corpora* epigrafici da un lato e quelli letterari dall'altro.

Le tecnologie e i formalismi che sono attualmente rubricati sotto la etichetta di Web Semantico mirano ad associare alle risorse informative variamente disponibili in rete una descrizione formale del loro significato, sicché un programma possa elaborare tale informazione in modo significativo (cioè tenendo conto di che cosa essa significhi), dedurne conseguenze, e generare automaticamente nuove informazioni. Questo nuovo orizzonte apre prospettive interessanti anche alle ricerche di tipo letterario, soprattutto nell'ambito della "tematologia" comparata e nello studio sui fenomeni della intertestualità, come testimoniato dalla letteratura internazionale più recente nel campo delle *Digital Humanities*. Lo stesso si può dire delle tecnologie di *knowledge extraction* e *text mining* applicate a vasti corpora testuali di tipo letterario, che sono state sperimentate con risultati interessanti in almeno tre progetti di ricerca in area statunitense, tra cui possiamo segnalare il *Tesserae Project: Intertextual Analysis of Latin Poetry* http://tesserae.caset.buffalo.edu di Neil Coffee *et al.*, il *Monk Project* http://monkproject.org diretto da John Unsworth e i progetti diretti da Franco Moretti presso lo Stanford Literary Lab.

La lemmatizzazione delle lingue classiche e l'attribuzione della categoria grammaticale si possono eseguire manualmente, ottenendo un grado di precisione molto elevato, oppure con l'applicazione di algoritmi probabilistici. Fin dall'origine della linguistica computazionale, il primo sistema fu applicato da p. Roberto Busa al *corpus Thomisticum* e dal CIPL-LASLA di Liegi a testi di autori della grecità e della latinità classica. *Morpheus*, il sistema di analisi morfologica realizzato da Gregory Crane ed usato presso il *Perseus Project* http://www.perseus.tufts.edu, è stato esteso da qualche anno con un sistema in grado di mettere in ordine decrescente di probabilità le possibili analisi proposte. Quanto al latino, il sistema di analisi morfologica realizzato a Pisa presso l'ILC-CNR – che si può dire rappresenti lo stato attuale dell'arte nel suo ambito – è disponibile online all'indirizzo http://www.ilc.cnr.it/lemlat/lemlat/index.html. Per l'arabo, si sta lavorando all'interno dello stesso Istituto, anche ai fini specifici della nostra ricerca.

L'analisi metrica automatica delle diverse realizzazioni dell'esametro greco e di altri metri non lirici può essere condotta facilmente con varie metodologie, mentre per i versi dattilici latini è già pronto uno strumento efficace e raffinato come *Pede certo* – di cui abbiamo parlato poco sopra.

L'esplorazione degli spazi semantici si effettua applicando metodologie statistiche a corpora testuali. I thesauri vengono generalmente definiti come collezioni ordinate di termini legati da relazioni di tipo gerarchico, asso-

ciativo e sinonimico, che costituiscono il lessico specialistico di un dominio della conoscenza e vengono usati per l'indicizzazione e il reperimento dell'informazione all'interno di tale dominio. Quelli basati su un sistema di classificazione multidimensionale dimostrano maggior flessibilità rispetto a quelli costruiti in base ad una classificazione monodimensionale, poiché forniscono una molteplicità di punti di accesso all'informazione e quindi maggior possibilità di indagini sui dati lessicali. Anche questi aspetti teorici e le varie applicazioni sono oggetto di lavoro per le due unità di Roma Tor Vergata e di Pisa, ILC-CNR.

4. Bibliografia

Coffee N. (2012). *Intertextuality in Latin Poetry*. In D. Clayman, ed., *Oxford Bibliographies in Classics*,OUP.

Pasini G.F. (1988). *Dossier sulla critica delle fonti (1896-1909)*, Pàtron.

Segre C. (1982). *Intertestuale / interdiscorsivo. Appunti per una fenomenologia delle fonti*. In C. Di Girolamo, I. Paccagnella, a c. di, *La parola ritrovata. Fonti e analisi letteraria*, Sellerio, pp. 15-28.

Panels

Digital Resources and Network Services for Digital Humanities Research / Risorse digitali e servizi di rete per la ricerca in campo umanistico

Digital humanities: difficoltà istituzionali e risposte infrastrutturali*

Dino Buzzetti
olim Università di Bologna, Italia
Fondazione per le Scienze Religiose 'Giovanni XXIII', Bologna, Italia
dino.buzzetti@gmail.com

Abstract. Il riconoscimento accademico delle ricerche prodotte nel campo delle *digital humanities* incontra particolari difficoltà dovute al peculiare assetto istituzionale dell'università italiana e alla rigida definizione dei settori scientifico-disciplinari previsti nell'ordinamento accademico nazionale. La creazione di un'infrastruttura scientificamente accreditata che ospiti i prodotti della ricerca realizzati in forma digitale può sensibilmente contribuire al superamento di tali difficoltà. L'esempio del progetto NINES, ora ampliatosi nel consorzio ARC, promosso negli Stati Uniti per risolvere analoghe difficoltà di natura istituzionale, può istruttivamente essere preso ad esempio e costituisce un modello al quale può giovare ispirarsi anche per le soluzioni tecnologiche e per la progettazione dell'ambiente infrastrutturale.

Parole chiave: *digital humanities*, riconoscimento istituzionale, infrastrutture.

1. Carenze strutturali e impegno strategico dell'AIUCD

Quest'anno la nostra Associazione (AIUCD)[1] ha scelto di dedicare una giornata del suo secondo convegno annuale all'attività degli istituti e dei centri di ricerca che operano nel campo delle *digital humanities*, consapevole del contributo essenziale che essi recano all'avanzamento degli studi in questo settore. Uno stretto rapporto tra l'intera comunità di ricerca e gli istituti e i centri che ne costituiscono eminenti punti di riferimento è dunque manifestamente fondamentale per affrontare le difficoltà strutturali che ostacolano il pieno sviluppo della formazione e della ricerca nel settore dell'informatica umanistica.

---

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

[1] Associazione per l'Informatica Umanistica e la Cultura Digitale http://www.umanisticadigitale.it/

Tali difficoltà investono non solo il processo di formazione delle competenze, ma anche il riconoscimento e la valutazione dei prodotti della ricerca. La rigida definizione dei settori scientifico-disciplinari su cui si fonda l'intero ordinamento accademico nel nostro paese pone un ostacolo strutturale enorme all'accertamento delle competenze specifiche del corpo docente nel campo delle *digital humanities*. Le effettive competenze interdisciplinari necessarie all'attività di ricerca non vengono adeguatamente riconosciute e accertate e gli scompensi strutturali che mettono a rischio il regolare percorso di formazione dei formatori si riflettono sulla formazione dei discenti e pregiudicano la piena adeguatezza didattica dei corsi variamente introdotti e dei pochi percorsi formativi organicamente istituiti.

Uno degli impegni strategici dell'Associazione è quindi rivolto a promuovere soluzioni che contribuiscano a porre rimedio a questa palese disfunzione di carattere strutturale, che investe in modo diretto non solo i processi di formazione e l'esercizio stesso della didattica, ma anche l'intera attività di ricerca. Infatti, il nesso tra formazione e ricerca è diretto e evidente, poiché è dalla valutazione dei risultati della ricerca che dipende il giudizio sull'acquisizione delle competenze richieste per il conseguimento della piena qualificazione all'attività didattica. Di qui nasce l'esigenza di un secondo e fondamentale impegno strategico per l'attività dell'Associazione, quello rivolto alla promozione di iniziative che incidano in modo sostanziale sui ritardi che ancora impediscono il pieno riconoscimento della validità scientifica dei risultati della ricerca prodotti e diffusi in forma digitale.

È sul riconoscimento di queste esigenze fondamentali e decisive per le condizioni attuali e per lo sviluppo futuro dell'attività di formazione e di ricerca nel campo delle *digital humanities* che si fonda l'iniziativa che l'Associazione intende proporre e che mira a creare un aggregatore di risorse digitali, valutate attraverso un processo di *peer review*, ai fini del pieno riconoscimento del loro valore scientifico e del loro accoglimento in un ambiente infrastrutturale istituzionalmente accreditato. Come le pubblicazioni a stampa vengono accreditate e valutate sulla base dell'autorevolezza delle riviste e delle sedi editoriali che le ospitano, in modo analogo i contributi di ricerca prodotti in forma digitale possono essere accreditati in base alle sedi infrastrutturali che le ospitano e attraverso le quali ne viene controllato il processo di revisione e di successiva riedizione e ne viene assicurato l'accesso stabile e permanente. Non tutto ciò che si produce in forma stampata può per ciò stesso avere validità scientifica, ma persiste il pregiudizio che non giudica in nessun modo sufficienti le garanzie che as-

sicurano validità scientifica a ciò che si produce in forma digitale e si rende accessibile in rete.

Per tutte queste ragioni, l'iniziativa dell'Associazione si ispira a un modello che ha proposto soluzioni concrete, che si sono rivelate idonee a superare analoghe difficoltà di natura istituzionale, ottenendo al tempo stesso il riconoscimento delle comunità scientifiche e disciplinari di riferimento.

2. La proposta infrastrutturale

Il modello a cui si ispira la proposta dell'Associazione trae origine dal progetto NINES (Networked Infrastructure for Nineteenth-Century Electronic Scholarship)[2], un'iniziativa che si è affermata come risposta infrastrutturale alla "diffusa, profonda e affatto comprensibile resistenza istituzionale al radicale cambiamento dei comportamenti legati all'attività di ricerca" (McGann 2005, 76) dovuta al diffondersi in misura sempre crescente della *digital scholarship*, ossia della pratica di ricerca sviluppata con metodologie computazionali e dunque "prodotta, curata e disseminata in forme digitali" (77). Tuttavia

> il complicato processo di creare e di sottoporre a *peer review* ricerche svolte in forma digitale (*digital scholarship*) nelle discipline umanistiche è ostacolato perché tuttora non esiste nessun riconoscimento istituzionale per questo tipo di lavoro da parte delle associazioni disciplinari di ricerca accademicamente accreditate (78).

Per porre rimedio a difficoltà di questa natura è nato il progetto NINES e per gli stessi motivi la nostra Associazione ha deciso ora di avanzare una propria proposta infrastrutturale. Entrambe le iniziative muovono dalla consapevolezza che "i problemi fondamentali sono di natura istituzionale e politica, anziché di natura tecnica o addirittura, in senso stretto, economica" (78). Infatti, l'obiettivo principale è costituito dal riconoscimento istituzionale dell'attività di ricerca svolta nel campo delle *digital humanities* e del valore scientifico dei suoi risultati prodotti in forma digitale. La risposta a tali difficoltà, sostiene Jerome McGann, l'ideatore del progetto, non deve limitarsi a considerazioni di carattere culturale, ad una critica che viene mos-

[2] http://www.nines.org/

sa, per citare Dante Gabriel Rossetti, "da un punto di vista interno" (73) alla riflessione teorica e che finisce col risolversi in un puro e semplice "gesto retorico" (74). Essa deve piuttosto procedere sul piano pratico e istituzionale. Non è la ricerca di una soluzione tecnologica che crea le difficoltà, come dimostra la possibilità di accedere *online* ai contenuti delle riviste tradizionali. Per risolvere il problema, si deve trasferire *online* l'intero processo della produzione dei contenuti, del loro controllo e del loro accreditamento attraverso la *peer review* e si debbono proporre nuovi modelli di iniziativa e di "impresa editoriale" (78) in ambiente digitale. Al mancato riconoscimento della produzione dei contenuti della ricerca in forma digitale, si deve reagire organizzando e accreditando l'intero processo della produzione, della disseminazione e del reimpiego dei contenuti digitali. Di qui la necessità di progettare e introdurre infrastrutture efficienti e istituzionalmente riconosciute a sostegno dell'intero processo di creazione, circolazione e utilizzazione delle risorse.

Il progetto NINES ha tracciato un percorso che ha dimostrato nei fatti di poter raggiungere risultati concreti nel più generale processo di "migrazione dell'intero patrimonio culturale tradizionale verso un sistema digitale di archiviazione, di accesso e di riutilizzazione" delle risorse (McGann 2012, ¶ 5). NINES, infatti, si è costituito come un "meccanismo istituzionale" (McGann 2005, 77), funzionale ed operante, per "la pubblicazione *online* dei risultati della ricerca in area umanistica, integrati e aggregati fra loro, prodotti in forma digitale e valutati con un processo di *peer review*" (81). Sicché, da un lato, l'iniziativa promossa da NINES interviene nel campo delle biblioteche di ricerca e dell'editoria accademica e, dall'altro, agisce nel contesto delle tradizionali associazioni scientifiche disciplinari, guidata dalla radicata convinzione che "case editrici e studiosi operino necessariamente in modo interdipendente" (77).

Lo scopo del progetto è quindi quello di creare, da una parte, un'impresa editoriale pilota che tenga conto dei diversi interessi disciplinari e risponda alle esigenze dei diversi risultati prodotti, che richiedono, per ciascun tipo di contenuto, soluzioni tecnologiche specifiche e appropriate. A sua volta, sul versante disciplinare, il progetto muove dalla fondamentale consapevolezza che un simile progetto di archiviazione digitale dei contenuti e di politica editoriale *online* non possa affermarsi se non proponendo contenuti scientificamente validi, vagliati attraverso procedimenti tradizionali di *peer review*, che ne garantiscano il necessario riconoscimento istituzionale da parte delle associazioni disciplinari di ricerca accademicamente accreditate.

Infine, sul piano tecnologico, il progetto mette a disposizione strumenti efficaci per la produzione, l'analisi e il reimpiego dei materiali archiviati, creando un sistema completo per la gestione della ricerca e per ospitarne i risultati, che ne garantisce il riconoscimento da parte della comunità scientifica e l'accreditamento in forme istituzionalmente accettate.

Infatti, le ricerche nel campo delle *digital humanities* attraversano una grave "crisi" (77) di natura strutturale, a cui solo soluzioni come quelle qui descritte paiono essere in grado di porre rimedio. Le ricerche svolte applicando metodi computazionali nelle discipline umanistiche non si sono tuttora affrancate dal modello tradizionale delle ricerche condotte in ambiente cartaceo e dalle forme tradizionali e istituzionalizzate della loro pubblicazione e della loro valutazione. In genere, la ricerca praticata in ambiente digitale procede ancora in modo isolato e viene impostata secondo modalità idiosincratiche, incomunicabili tra loro. I progetti, prevalentemente di scala limitata, procedono solo grazie all'impegno e alla dedizione di singoli individui o di gruppi isolati. In questa situazione, le scarse risorse disponibili non permettono di assicurarne lo sviluppo, l'aggiornamento e la conservazione permanente e neppure un controllo oggettivo secondo "standard disciplinari comunemente riconosciuti" (77).

Ora, il successo raggiunto dal progetto NINES è accertabile con tutta evidenza osservando non solo l'insieme delle risorse digitali in esso aggregate, che riguardano l'area delle ricerche letterarie sul secolo diciannovesimo, ma anche e soprattutto considerando l'estendersi del modello alle ricerche riguardanti il secolo diciottesimo (18thConnect: Eighteenth Century Scholarship Online)[3] e il Medioevo (MESA: Medieval Electronic Scholarly Alliance)[4], nuove iniziative, queste, che si sono tutte consorziate, assieme a NINES, nell'Advanced Research Consortium (ARC)[5]. E il processo si sta ulteriormente ampliando con la prossima adesione della Renaissance English Knowledgebase (REKn) (cfr. Siemens 2011).

Di qui l'impegno dell'Associazione volto a promuovere la creazione di un aggregatore di risorse digitali ispirato al modello del progetto NINES. L'iniziativa prevede la creazione di un nucleo iniziale di risorse digitali prodotte nell'ambito delle discipline umanistiche aggregate tra loro e l'istituzione di comitati redazionali, che ne assicurino il vaglio e la valutazione attraverso

[3] http://www.18thconnect.org/
[4] http://www.mesa-medieval.org/
[5] http://idhmc.tamu.edu/arcgrant/

un processo di *peer review* riconosciuto dalle associazioni disciplinari di ricerca accademicamente accreditate, non solo per quanto riguarda l'aspetto contenutistico, ma anche per quanto riguarda l'adeguatezza delle soluzioni tecnologiche, con riferimento agli standard e alle procedure di ricerca comunemente accettate e riconosciute.

L'accreditamento istituzionale può infine essere raggiunto ospitando il progetto in una infrastruttura riconosciuta dallo European Research Infrastructure Consortium (ERIC)[6], il quadro normativo istituito il 28 agosto 2009 dalla commissione dell'Unione Europea per il riconoscimento delle infrastrutture di ricerca. Il nostro paese sta promovendo, attraverso il Dipartimento Scienze Umane e Sociali, Patrimonio Culturale (DSU) del CNR[7], la costituzione di consorzi nazionali per la partecipazione alle infrastrutture ERIC operanti nel campo delle discipline umanistiche - la Digital Research Infrastructure for the Arts and Humanities (DARIAH)[8] e la Common Language Resources and Technology Infrastructure (CLARIN)[9]. La costituzione di DARIAH-it e di CLARIN-it offrono la sede istituzionale adeguata per collegare le infrastrutture specifiche, già esistenti o in corso di realizzazione nel nostro paese, in un'unica infrastruttura nazionale, integrata con la rete europea.

Mentre la rete infrastrutturale CLARIN costituisce la sede più adeguata per accogliere le risorse e le tecnologie linguistiche frutto di ricerche specifiche nel campo della linguistica computazionale, per la più generale produzione di risorse digitali nell'ambito delle discipline umanistiche risulta manifestamente più appropriata l'allocazione nella più comprensiva rete infrastrutturale DARIAH. È quindi nel contesto di queste iniziative infrastrutturali, istituzionalizzate a livello nazionale e internazionale, che intende collocarsi il progetto dell'Associazione, ispirato al modello proposto e realizzato, per esigenze affatto analoghe, dal progetto NINES.

E questo è il tema centrale della discussione che qui si propone ai centri e agli istituti di eccellenza che operano autorevolmente nel nostro paese nel campo delle *digital humanities.*

[6] http:// ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric
[7] http://www.dsu.cnr.it/
[8] http://www.dariah.eu/
[9] http://clarin.eu/

3. Bibliografia

McGann J. J. (2005). *Culture and Technology: The Way We Live Now, What Is to Be Done?* «New Literary History», vol. 36, no 1, pp. 71-82.

McGann J. J. (2012) *Memory Now*. «4Humanities: Advocating for the Humanties». URL=http://4humanities.org/2012/08/jerome-j-mcgann-memory-now-2/ [ultima visita 09.06.2014].

Siemens R. *et al.* (2011). *Prototyping the Renaissance English Knowledgebase (REKn) and Professional Reading Environment (PReE), Past, Present, and Future Concerns: A Digital Humanities Project Narrative*. «Digital Studies / Le champ numérique», vol. 2, no 2. URL= http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/182/255 [ultima visita 11.06.2014].

Digital humanities e analisi dei testi[*]

Paolo Mastandrea
Dipartimento di Studi Umanistici, Università Ca' Foscari, Venezia, Italia
mast@unive.it

Abstract. Nell'attuale fase di sviluppo degli studi umanistici, compito fondamentale di chi opera nel campo della filologia è l'implementazione, l'organizzazione e la manutenzione costante di archivi elettronici contenenti grandi serbatoi di testi – che nel caso delle opere classiche, antiche e medievali, ma anche delle principali letterature moderne fino al Novecento, possono arrivare non lontano dalla completezza dei materiali. Entro tali *corpora* è però necessario svolgere un lavoro computazionale che crei strumenti di ricerca potenti e insieme sofisticati, flessibili, adatti alla molteplicità degli approcci, in vista di obiettivi ambiziosi e magari di conseguenze inattese: capaci di offrire basi solide al dibattito sul tema della critica intertestuale – dati 'scientifici', che sottraggono i risultati alla incertezza della soggettività, della tendenziosità, insomma dell'ideologia; in quell'ottica di revisione dei canoni che è in corso e coinvolge gli studiosi anche meno disposti alle novità tecnologiche.

Parole chiave: intertestualità formale e concettuale, memoria poetica, archivi di testi letterari.

Nella naturale simpatia con le attività della AIUCD, anzi nella stessa partecipazione al convegno di oggi, prevale l'istinto di una persona curiosa, sempre attratta da quanto si produce nell'universo espanso delle digital humanities; e che lo guarda però con prudenza: forse per una ormai 'antiquata' formazione disciplinare di base, forse per il fastidio caudato da certi frasari tecnici, esotici, astrusi, insomma esclusivi; soprattutto per un atteggiamento di cautela verso ogni novità che – almeno in prospettiva – non dimostri i chiari vantaggi della sua applicazione in campi d'indagine e di lavoro già collaudati. Senza false modestie, dunque ben lungi dalla pretesa di volare alle altezze di un teorizzatore sistematico o di un originale pensatore, mi muovo come un opportunista pratico, che mira direttamente allo studio del testo; o per meglio dire, dei testi, al loro interno e fra loro.

[*] M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

Fin dai primi contatti con un PC, avvenuti nel mio caso una trentina di anni orsono, ero colpito dai due aspetti che subito apparivano potenzialmente eversivi delle nostre abitudini e attitudini verso la comunicazione scritta: la facoltà di modificare continuamente e pressoché all'infinito la forma che abbiamo scelto di imporre all'espressione (di notizie, di concetti, di narrazioni, ecc.); ma soprattutto sbalorditiva, e comunque senza precedenti, si presentava la possibilità aperta a chiunque di operare in assoluta autonomia da ogni mezzo già predisposto da altri per lo studio del testo, senza l'obbligo di ricorrere a scelte o a mediazioni che non si condividono, rispetto a ricerche lessicali, ad indagini stilistiche, ad esami critici di varianti, collegamenti di idee e quant'altro era lecito operare sopra la carta stampata pagando costi insopportabili di fatica, di tempo, (e perché no?) di denaro.

Per consentire la massima efficacia a queste operazioni, appariva dunque indispensabile anzitutto disporre di grandi archivi elettronici che – magari sulla base di riconosciute omogeneità linguistiche e culturali – accorpassero autori ed opere delle grandi civiltà letterarie occidentali, moderne come antiche. Nei primi anni novanta fu l'avvento della tecnologia CD-ROM a permettere la collezione di grosse quantità di testi, cui si applicavano programmi di *word searching*, sistemi indicizzatori o concordatori, rimari eccetera; un decennio dopo quei caratteri, ormai volatili ma ancora fisicamente incisi nei solchi d'alluminio del compact disc, venivano trasferiti e quasi sospesi nella impalpabilità immateriale della Rete web; l'incremento dei dati procede senza sosta, con soddisfazione generale.

Non esiste forse campo scientifico dove si siano ottenuti dall'informazione automatica avanzamenti così notevoli come nella ricerca verbale sui grandi corpora letterari: e (cosa che forse alcuni ancora ignorano) a partire proprio dai classici antichi [ricordo che una volta la filologia latina fu chiamata da Gianfranco Contini "primogenita delle filologie moderne" [*Breviario di ecdotica*, Einaudi, 1990, p. 46]. Sin dalla metà dell'Ottocento, era proliferata la redazione di indici, repertori, vocabolari, soprattutto concordanze in grado di aprire accessi alla conoscenza stilistica – con riguardo speciale per l'idioletto dei singoli autori, o per le lingue tecniche, o per un determinato segmento temporale o genere storico-letterario; si tratta di strumenti affidabili, la cui sistematicità di base ci trasmette il sicuro ottimismo dove riposa la loro stessa concezione. Ma proprio quella 'stabilità' in apparenza oggettiva, che insieme all'apparato critico di matrice lachmanniana offriva ai filologi un senso di orgogliosa certezza sulle fondamenta della disciplina, costituiva il limite che la tecnologia permette ora di oltrepassare d'un balzo.

L'interrogazione dinamica, surclassando qualunque più sofisticato strumento a stampa oggi disponibile, ci mette in condizione di individuare nel contesto le presenze di abbinamenti o associazioni plurime di termini (dittologie nominali o verbali o miste, giunture tra aggettivi e sostantivi, predicati e verbi), rilevare le tipicità e i registri di stile, isolare gli idiotismi nella prosa come le strutture metrico-ritmiche nella versificazione; in secondo luogo favorisce una prassi di lettura irregolare e saltuaria, ma gratuita e veramente 'en-ciclo-pedica', cioè disobbligata dalla linea unidirezionale cui ci spinge l'abitudine, aperta in ogni snodo a rincorrere le cooccorenze, i parallelismi, le similitudini, quanto a proporre gratuiti confronti e collaudi sempre diversificati, libera e capace di sovvenire con rigore scientifico alle curiosità prevedibili dall'intelligenza umana.

È concessa in tal modo ai ricercatori una opportunità per raggiungere non soltanto i risultati già conseguibili un tempo attraverso gli strumenti cartacei tradizionali – anche i più progrediti, perché ottenuti negli ultimi decenni grazie ad elaborazione elettronica; ma anche per catturare i rapporti semantici, dunque concettuali, a prescindere dalla forma lessicale adottata; dall'altro lato, svelare le relazioni di pura assonanza, dunque alogiche o prelogiche: un'enorme massa di allitterazioni, rime, ritmi, scarti e alterazioni anche minime, riguardanti nella gran parte dei casi i fonemi consonantici.

Si allargano senza limiti le potenzialità individuali di sondaggio sulla parola (per qualsivoglia fine) dall'uomo, liberandoci da ogni remora e condizionamento, ma anche sottraendo alla soggettività, cioè al preconcetto critico, l'intuizione dei legami intertestuali; giungendo a livelli mai prima attinti di conoscenza, dove ogni rapporto tra significato e significante si scioglie, a favore di un'intima unione tra i suoni e il senso; dove la memoria dei poeti inganna se stessa, si tramuta in forme di ingenua spontaneità fanciullesca, magari suggestionata da cadenze e da echi; dove i fondamentali archetipi e i nobili modelli sono dall'artista plasmati secondo gusti estetici di pretesa originalità, in rielaborazioni che si vorrebero intatte da condizionamento grammaticale o (etimo-)logico.

A questo punto, è nostra responsabilità (oserei dire: nostro dovere) cooperare ai fini di un ulteriore allargamento dei grandi corpora testuali già formati; ma soprattutto è necessario si integrino e si rendano omogenei nella misura più ampia, dunque le risorse vanno messe in comune a disposizione di tutti, gli strumenti già creati o da creare vanno concepiti sin dall'inizio per offrire la massima collaboratività reciproca. Sono argomenti all'ordine del giorno, oggi qui come in tutte le occasioni di incontro fra quanti, in

ogni parte del mondo, si occupano di Digital Humanities: ma l'esigenza è sentita meglio di altri da chi sa che le letterature antiche e moderne costituiscono un 'sistema' che in Omero la propria origine, da sempre – cioè sin dall'introduzione della scrittura a fini artistici, nella Atene di Pericle. La 'tradizione classica', coi suoi generi e i suoi canoni, coi suoi autori e capolavori di riferimento, costituisce un repositorio vasto ma non immenso, un mar mediterraneo nel quale la navigazione che da Julia Kristeva in poi si chiama 'intertestuale' è pratica antica, diffusa sin dai filologi del Museo d'Alessandria. La funzione dell'informatica nello studio dei rapporti interni a questo sistema appare insostituibile non solo in chiave scientifica e tecnica, a beneficio di pochi esperti accademici, ma in generale per la comprensione dei significati basilari della nostra cultura e della nostra civiltà; permette di analizzare i fili della storia, di ricostruire nella loro continuità gli ambienti intellettuali e sociali, gli scopi per cui i libri si scrivevano, il pubblico delle persone cui erano destinati e in cui la eventuale scoperta di quanto esse già sapevano (reminiscenze, imitazioni, allusioni, insomma il processo di 'agnizione') costituisce il cuore del problema, il segreto fascino dell'estetica, il piacere di ogni fruizione artistica.

Infrastrutture e risorse digitali. L'esperienza dell'Iliesi*

Antonio Lamarra
Cnr-Iliesi
Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Roma, Italia
antonio.lamarra@cnr.it

Abstract. L'Iliesi vanta un'esperienza di diversi decenni nell'ambito della digitalizzazione dei testi, avendo coniugato fina dalla metà degli anni Sessanta ricerca storico-filosofica, lessicografia e metodologie informatiche per il trattamento dei dati linguisti e testuali. Con la partecipazione ai progetti europei *Discovery* e *Agora* l'Istituto ha poi ampliato la sua esperienza in direzione della creazione di archivi digitali funzionali alle esigenze della ricerca sui testi e degli studi sulla terminologia filosofica come pure alla pubblicazione online dei risultati. Attualmente, l'Iliesi guarda alla partecipazione al consorzio europeo Dariah, cui il nostro Paese ha recentemente aderito, come ad un'occasione particolarmente utile per la realizzazione di una rete infrastrutturale caratterizzata da procedure di validazione e di valutazione dei contenuti che, per rigore ed efficacia, trasferiscano alle pubblicazioni digitali la medesima credibilità e affidabilità che tradizionalmente si conferisce alle pubblicazioni su supporto cartaceo.

Parole chiave: digitalizzazione di testi, piattaforme testuali, procedure di validazione, validazione dei pari, infrastrutture di ricerca.

1. Dall'indicizzazione dei testi alle infrastrutture per la ricerca

Fin dalle origini, a metà degli anni Sessanta del secolo scorso, l'Iliesi – o, per meglio dire, l'allora Centro di Studio per il Lessico Intellettuale Europeo (Cslie) del Cnr[1] – ha caratterizzato il suo approccio alla storia delle idee filosofiche e scientifiche sotto un duplice profilo metodologico: (a) per una

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. Cleup, Padova, 2014.

[1] Nella sua configurazione attuale, l'Iliesi nasce nel 2001 dalla fusione del Cslie con un altro Centro romano del Cnr, il Centro di Studio del Pensiero Antico. Per una breve storia dell'Istituto, come dei due precedenti Centri, si vedano le pagine dedicate all'argomento sul sito: http://www.iliesi.cnr.it/storia.shtml

spiccata attenzione alla dimensione lessicale dei testi, considerata come una chiave d'accesso privilegiata per la loro comprensione storico-concettuale, e (b) per il ricorso a metodologie informatiche nel trattamento dei dati linguistici e testuali. Per questo secondo aspetto l'Istituto vanta, quindi, un ruolo sicuramente pionieristico, non solo a livello nazionale ma su scala europea. Se l'ammirevole opera del padre Busa aveva costituito una fonte d'ispirazione e offerto un primo esempio concreto di approccio computazionale all'analisi dei testi (e specialmente dei testi filosofici), l'esperienza dell'Istituto si è poi sviluppata in stretta connessione con quella delle maggiori iniziative internazionali che nel settore erano successivamente venute costituendosi in quegli stessi anni. Di particolare rilievo, in questo contesto, la collaborazione scientifica che venne ben presto a stabilirsi con l'Istituto di Linguistica Computazionale (ILC) di Pisa, che per diversi anni costituì per il CSLIE un riferimento costante per l'elaborazione dei testi.

Dapprima orientata allo sviluppo di indici e concordanze (lemmatizzate) e alla successiva redazione di lessici d'autore ovvero di singole opere, l'attività dell'istituto nel campo che oggi diciamo delle *digital humanities* si è poi rivolta, a partire dagli anni Ottanta, soprattutto alla costituzione di basi di dati e archivi digitali dedicati a testi filosofici fra i più emblematici della tradizione europea della prima età moderna: dalla *Banca dati di testi filosofici dell'età moderna*, ad archivi più specialistici e mirati, quali l'*Archivio dei filosofi del Rinascimento*[2] o l'*Archivio di testi per la storia dello spinozismo*[3], agli archivi dedicati alle *Opere complete* di Giambattista Vico[4] e ai *Lessici filosofici dell'età moderna*[5], per citare solo i più ampi e significativi. A questi si è venuto affiancando nell'ultimo decennio un portale, *Daphnet* – acronimo che sta per *Digital Archives of Philosophical Texts on the Net* – che per più versi segna un ulteriore momento di svolta nell'approccio dell'ILIESI alla digitalizzazione dei testi filosofici e alla conseguente elaborazione delle informazioni linguistiche e testuali[6].

In primo luogo, *Daphnet* non nasce come un archivio, ma come un portale strutturato in una pluralità di piattaforme che, pur dotate di una pro-

[2] http://www.iliesi.cnr.it/afr/index.shtml
[3] http://www.iliesi.cnr.it/perl/pagina_xhtml.pl?scelta=11
[4] http://151.100.146.63/DirVico/BookReader/html/application.html
[5] http://www.iliesi.cnr.it/Lessici/home_lessici.html
[6] http://www.daphnet.org/

pria autonomia, sono tuttavia strettamente interrelate. Attualmente esso dà accesso alle piattaforme seguenti: (a) *Ancient Philosophy*[7] e (b) *Modern Philosophy*[8], che contengono fonti primarie per lo studio del pensiero antico e della prima modernità europea, (c) *Daphnet Digital Library*[9], che invece raccoglie solo fonti secondarie relative ai testi presenti nelle prime due piattaforme, e infine (d) alla piattaforma dedicata alla rivista online *Lexicon Philosophicum. International Journal for the History of Texts and Ideas*, edita dall'Iliesi[10]. Il progetto del portale, per adesso realizzato solo su di un primo campione di testi, prevede la creazione di *link* intertestuali tra fonti primarie e secondarie, ivi compresi i contributi pubblicati sulle pagine digitali della rivista dell'Istituto. La realizzazione di questo portale fu resa possibile soprattutto dalla partecipazione dell'Istituto a due progetti europei, *Discovery*[11] e *Agora*[12], che ne determinarono alcune caratteristiche di fondo, per le quali esso si differenzia da tutti gli altri archivi precedentemente costituiti dall'Istituto: l'apertura alla filosofia antica, per quanto riguarda i contenuti testuali, e – per altri versi – l'abbandono di un sistema di codifica proprietario e il passaggio al sistema di codifica Tei, l'adozione senza riserve di un approccio *Open* tanto riguardo al software impiegato quanto alla consultabilità delle piattaforme, sono fra gli aspetti più notevoli che differenziano il portale *Daphnet* dagli altri archivi testuali dell'Iliesi. A questi vanno poi aggiunte quantomeno l'interconnessione delle piattaforme, la possibilità di annotare i testi e, soprattutto, la possibilità di una loro soggettazione linguistico-concettuale operata con riferimento ad una ontologia di dominio. In altri termini, il portale Daphnet risulta strutturalmente concepito non più solo come un archivio statico di testi, ma come un ambiente multifunzionale, dinamicamente indirizzato a favorire e supportare la ricerca.

Su questo sfondo si è appena aperta per l'Istituto una nuova prospettiva d'impegno nel campo delle *digital humanities*, con la sottoscrizione da parte italiana dei protocolli di adesione al consorzio europeo Dariah (*Digital Research Infrastructure for the Arts and the Humanities*)[13]. Infatti, mentre al Mibac competerà di rappresentare a livello politico gli interessi nazionali in

[7] http://ancientsource.daphnet.org/
[8] http://modernsource.daphnet.org/
[9] http://scholarlysource.daphnet.org/index.php/DDL
[10] http://lexicon.cnr.it/
[11] http://www.discovery-project.eu/home.html
[12] http://www.project-agora.org/
[13] https://www.dariah.eu/

seno al consorzio, il CNR si è vista riconosciuta la funzione di ente scientifico di riferimento per la comunità dei ricercatori italiani che si organizzerà all'interno di DARIAH. Di qui il ruolo di primo piano che, da un lato, svolgerà il *Dipartimento Scienze Umane e Sociali, Patrimonio Culturale* (DSU) del CNR e, d'altra parte, la concreta prospettiva che proprio all'ILIESI venga affidato il compito di coordinare la rete delle istituzioni italiane che vorranno collegarsi a livello nazionale (DARIAH-IT), per poter interagire efficacemente col consorzio DARIAH su scala europea. Un primissimo impegno in tal senso sarà costituito dall'organizzazione, cui l'Istituto si è già impegnato, del *4th General VCC Meeting* di DARIAH-EU, previsto a Roma nei giorni 17-19 settembre dell'anno in corso, e che vedrà riunito l'insieme degli organismi direttivi del consorzio[14]. Questa sarà un'ottima occasione per conoscere più da presso le forme organizzative che il consorzio ha già assunto nei diversi Paesi aderenti, ma anche per mettere a fuoco con qualche precisione la gamma di opportunità che per suo mezzo si offriranno alla comunità italiana dei ricercatori impegnati nel settore delle *digital humanities*. L'AIUCD in questo quadro costituirà senza dubbio un interlocutore privilegiato nella fase di realizzazione della rete italiana di DARIAH.

2. Infrastrutture e pratiche di validazione

Un'infrastruttura di ricerca, di per sé, non garantisce la natura scientifica dei contenuti che offre e che attraverso di essa vengono pubblicati e diffusi. L'organizzazione (e l'esistenza stessa) di una tale infrastruttura, tuttavia, favoriscono in misura decisiva la possibilità – per una specifica comunità scientifica – di dotarsi di strumenti condivisi di validazione e di garanzia. È esattamente quel che è avvenuto nell'universo gutemberghiano della carta stampata: di fronte all'evidenza che a stampa si potevano trovare tanto capolavori sublimi quanto perfette nullità letterarie, tanto testi rigorosamente scientifici quanto pubblicazioni prive di qualunque rigorosità metodologica, la comunità scientifica (nelle sue varie articolazioni) ha messo a punto nel tempo una serie più o meno ampia di strumenti condivisi (talora anche impliciti) che consentono in maniera relativamente semplice di qualificare l'appartenenza di uno scritto all'una o all'altra categoria, di riconoscere cioè se uno scritto appartiene alla letteratura scientifica oppure no. Pratiche

[14] https://www.dariah.eu/activities/general-vcc-meetings/4th-general-vcc-meeting.html

simili non esistevano prima dell'avvento della stampa a caratteri mobili su carta e in alcuni casi la comunità scientifica vi si è talmente assuefatta da non riuscire agevolmente a separare le strategie di certificazione che ha escogitato per le pubblicazioni cartacee dalla circostanza, di per sé contingente, che quei contenuti siano veicolati da un supporto cartaceo piuttosto che da un altro genere di supporto. Quasi che, in altri termini, i tratti caratteristici del rigore scientifico fossero riscontrabili esclusivamente in un sottoinsieme delle pubblicazioni cartacee, siano esse libri o periodici. Come ben sappiamo, è proprio ciò che troppo spesso ancora avviene nel settore delle scienze umane e sociali, che incontrano più difficoltà di altre discipline ad assumere fino in fondo le conseguenze della rivoluzione digitale, anche perché molto più di altre sono legate alla pubblicazione di libri piuttosto che di saggi o di articoli su riviste.

Una riflessione sulle pratiche da seguire per conferire alle pubblicazioni digitali una dignità in tutto comparabile a quella propria delle pubblicazioni tradizionali è in corso già da qualche tempo in diversi settori del mondo accademico più direttamente coinvolti nelle nuove metodologie digitali. Per quanto riguarda il nostro Istituto, questa riflessione data ad almeno una decina d'anni ed ha portato ad una serie di decisioni condivise con i partner dei progetti europei che ho sopra ricordato, già a partire da *Discovery*. Basti ricordare che tutte le piattaforme testuali cui si accede dal portale *Daphnet* sono certificate da un apposito Comitato Scientifico, così come non c'è contributo critico in esse pubblicato che non sia stato *peer-reviewed*; che tutti i contenuti delle piattaforme possiedono un indirizzo (URL) stabile e che, per altri versi, gli *editor* delle piattaforme garantiscono la stabilità dei testi delle pubblicazioni in esse contenute. In questo modo, oltre alla validazione scientifica fornita dalla preliminare valutazione dei pari, intendiamo garantire alle pubblicazioni digitali contenute nel portale *Daphnet* due caratteristiche che, proprie di ogni pubblicazione a stampa, non ineriscono necessariamente ai contenuti digitali: che siano sempre reperibili al medesimo indirizzo (così come un testo a stampa è sempre reperibile al medesimo riferimento bibliografico) e che, una volta pubblicati, non subiscano variazioni o alterazioni di sorta, a meno di revisioni in edizioni successive del medesimo testo.

È solo sula base di criteri editoriali di questo tipo che, ritengo, si potrà garantire ai contributi scientifici prodotti in forma digitale un credito in tutto analogo a quello di cui godono i risultati della ricerca pubblicati sui tradizionali media cartacei: sul piano strutturale, occorre che siano assolutamente stabili e sempre identificati dal medesimo riferimento – il che li rende citabili

e verificabili nel tempo – mentre, sotto il profilo della qualità, occorre che la comunità scientifica di riferimento possa riconoscerli come veri prodotti della ricerca, grazie all'accreditamento operato da comitati editoriali competenti o comunque in ragione della serietà scientifica dell'istituzione che provvede a pubblicarli. Il che, vorrei notare, non vuol dire certificare la qualità scientifica del contributo pubblicato, ma la sua natura di prodotto della ricerca e non ad esempio, di testo letterario o magari di pseudo-testo scientifico. Condivido in pieno, quindi, l'affermazione di Buzzetti, quando nel suo intervento a questa tavola rotonda nota che, per superare in maniera efficace la diffidenza che ancora circonda l'attività di ricerca svolta con metodologie digitali, si tratta di saper trasferire nell'universo digitale l'intero processo di produzione dei contenuti, del loro controllo, del loro accreditamento e perfino delle relative procedure di pubblicazione. In altri termini, solo se tutti gli aspetti qualificanti della produzione e della pubblicazione dei risultati della ricerca che sono riconosciuti dalla comunità scientifica trovano equivalenti efficaci e soddisfacenti anche in ambito digitale si potrà pretendere che quella stessa comunità valuti i prodotti digitali alla stessa stregua di quelli cartacei.

A questo fine, la diffusione di standard comportamentali e di buone pratiche condivise mi parrebbe davvero un impegno prioritario da parte di quanti si riconoscono nelle *digital humanities*, tuttavia un impulso molto importante potrà venire anche dalla costituzione di infrastrutture per la ricerca che, per le loro caratteristiche, sappiano accreditarsi come strumenti di elaborazione e di diffusione di contenuti di valore scientifico. Non v'è dubbio che il momento critico per l'accreditamento dei risultati della ricerca sia costituito dalla loro pubblicazione, cioè dal momento in cui per definizione il lavoro dello studioso viene reso di pubblico dominio, divenendo verificabile e controllabile. Pertanto, la possibilità di pubblicare contributi all'interno di un'infrastruttura accreditata e che sottoponga i contributi ricevuti per la pubblicazione a procedure condivise di validazione costituirebbe un indubbio vantaggio nel processo in corso di accreditamento del prodotto digitale presso una comunità scientifica che si mostra ancora molto diffidente. Il vantaggio evidente del puntare sulle infrastrutture di ricerca piuttosto che sulle pratiche che individualmente ciascun ricercatore dovrebbe seguire sta nella semplificazione di un processo altrimenti molto più lento e nella maggior facilità di giungere alla definizione di standard largamente condivisi. Infatti, una volta accreditata l'infrastruttura in ragione della serietà e dell'af-

fidabilità delle procedure cui sottopone i contributi destinati alla pubblicazione, verrebbero di conseguenza accreditati tutti i contributi che, avendo superato quel vaglio, venissero effettivamente pubblicati.

Come ho avuto modo di accennare, il nostro Istituto ha già cominciato a muoversi in questa direzione con la creazione del portale *Daphnet* e con il lancio di una nuova rivista internazionale interamente online ad accesso libero, cui ben presto seguirà il lancio di un collana di monografie e di studi in formato digitale. Per l'ILIESI la via della digitalizzazione delle risorse linguistiche e testuali, non meno che quella dell'editoria elettronica, rappresentano dunque una scelta metodologica assunta con determinazione e con convinzione. Senza inutili radicalismi (perché riteniamo che quantomeno nel nostro tempo vi sia ancora uno spazio molto significativo per libri e riviste a stampa), ma con pari convinzione ci sentiamo peraltro impegnati a favorire quel mutamento di mentalità necessario a riconoscere fino in fondo che, in ultima analisi, la validità scientifica di una pubblicazione non gli proviene dal suo sostegno materiale (o immateriale) ma dal contributo che offre all'avanzamento delle conoscenze e alla riflessione critica. La realizzazione a livello nazionale di un network di istituzioni e di imprese scientifiche all'interno del consorzio europeo DARIAH, la realizzazione cioè di una rete DARIAH-IT, costituirà indubbiamente un'occasione di grande rilievo in quella direzione alla quale ci sentiamo impegnati fin d'ora a dare il nostro contributo.

DH@ILC: linee di attività e ricerca*

Simonetta Montemagni
Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa, Italia
simonetta.montemagni@ilc.cnr.it

Abstract. Le principali linee di ricerca e sviluppo dell'ILC nel settore delle DH possono essere ricondotte ai seguenti filoni: acquisizione e conservazione di testi; progettazione e sviluppo di risorse e strumenti per il trattamento automatico di lingue classiche e varietà storiche della lingua; progettazione e sviluppo di strumenti per l'analisi del testo; costruzione di un'infrastruttura italiana per la ricerca nell'ambito delle scienze umane e sociali.

Parole chiave: archivi testuali digitali, risorse linguistiche, piattaforme Web per l'analisi del testo, trattamento automatico del linguaggio, CLARIN.

1. Introduzione

L'Istituto di Linguistica Computazionale, fondato da Antonio Zampolli e da lui diretto fino al 2003 e che oggi porta il suo nome, fin dalle origini ha svolto un ruolo centrale nel dibattito scientifico nazionale e internazionale sull'apporto dell'informatica agli studi umanistici. L'attività strategica svolta dall'ILC era basata sull'intuizione che la conservazione, lo spoglio e la fruizione dei testi del patrimonio culturale, condotti tipicamente attraverso la produzione di indici e concordanze, dovessero coniugarsi con la rappresentazione e l'elaborazione della loro struttura linguistica e del loro contenuto attraverso strumenti di analisi linguistica automatica del testo e l'uso di risorse linguistiche. In altre parole, i due filoni storici dello "Humanistic Text Processing" (HTP) da un lato e del "Natural Language Processing" (NLP, o "Trattamento Automatico del Linguaggio") dall'altro non dovevano essere visti come separati ma presentavano importanti sinergie con ricadute significative in entrambi i settori. All'interno di questo quadro, a partire dagli anni '90 l'ILC ha anche contribuito attivamente alla definizione e alla promozione di standard di rappresentazione dell'informazione testuale e

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

linguistica attraverso la partecipazione a comitati internazionali (TEI, EAGLES e le iniziative che da questi hanno avuto origine) per arrivare, più recentemente, alla definizione di infrastrutture di ricerca integrate finalizzate a stabilire funzionalità di accesso, interoperabilità e condivisione di risorse e strumenti linguistici per la comunità di ricerca nel settore delle scienze umane e sociali. Tra i prodotti storici dell'ILC rilevanti per il settore della ricerca umanistica vale la pena menzionare: il DBT (Data Base Testuale, di Eugenio Picchi, Brevetto CNR del 1988) che rappresenta uno dei primi sistemi di analisi testuale finalizzato alla gestione e navigazione di testi (singoli, corpora e strutturati) e che include componenti per la lemmatizzazione e l'annotazione del testo; LEMLAT, un componente software per la lemmatizzazione del latino (di Andrea Bozzi, Giuseppe Cappelli e Nino Marinone, Brevetto CNR – Università di Torino del 1992).

Oggi, la ricerca all'ILC nel settore delle discipline umanistiche continua a essere fortemente improntata alle linee strategiche delineate sopra con importanti innovazioni legate all'evoluzione della linguistica computazionale, delle discipline umanistiche che traggono beneficio dall'utilizzo di tecnologie del linguaggio fino ai più recenti sviluppi tecnologici nel settore dell'informatica: partendo dagli stessi ingredienti di base, la ricerca corrente sta sperimentando nuovi tipi di analisi ed elaborazioni, rese possibili da nuovi paradigmi di ricerca e dalle nuove possibilità tecnologiche offerte dai computer. In particolare, le principali linee di ricerca e sviluppo possono essere ricondotte ai seguenti filoni, brevemente illustrati di seguito attraverso una selezione delle attività più rappresentative in corso:

1. acquisizione e conservazione di testi;
2. progettazione e sviluppo di risorse e strumenti per il trattamento automatico di lingue classiche e varietà storiche della lingua;
3. progettazione e sviluppo di strumenti per l'analisi del testo;
4. costruzione di un'infrastruttura italiana per la ricerca nell'ambito delle scienze umane e sociali.

Linea 1 - Acquisizione e conservazione di testi

Qualsiasi analisi ed elaborazione automatica del testo presuppone la sua digitalizzazione, un processo che ancora oggi pone sfide non soltanto a livello tecnico ma anche scientifico. Le attività correnti dell'ILC in questa direzione toccano due aspetti complementari riguardanti l'acquisizione di testi non ancora disponibili in formato digitale da un lato, e la salvaguardia e la

conservazione di archivi testuali esistenti a rischio di perdita a causa dell'inevitabile obsolescenza tecnologica dall'altro.

Il primo aspetto, affrontato all'interno di una collaborazione tra l'ILC, l'Open Philology Project dell'Università di Lipsia, il Perseus Project della Tufts University (Medford, MA) e la Mount Allison University (NB, Canada), riguarda la messa a punto di metodi e tecniche avanzati per la correzione semi-automatica del risultato del processo di Optical Character Recognition (OCR). Al momento, la collaborazione è focalizzata sull'acquisizione di testi greci a partire da edizioni criti-che che presentano difficoltà a vari livelli (complessità dell'impaginazione, danneggiamento delle pagine in edizioni datate, ampio spettro di glifi da riconoscere e contenuto plurilingue dell'apparato critico). In particolare, l'ILC (Boschetti, questo volume) ha messo a punto tecniche avanzate per la correzione dell'output dell'OCR che fanno uso di risorse linguistiche (come reper-tori sillabici e di parole flesse) e testuali (testi precedentemente digitalizzati) guidandone la revi-sione manuale. A questo aspetto sono anche riconducibili le attività condotte sul versante dell'epigrafia digitale: l'ILC ha una collaborazione in atto con il "Visual Computing Laboratory" del CNR-ISTI per il trattamento congiunto del testo e della rappresentazione digitale del supporto materiale, secondo il modello proposto in (Lamé *et al.* 2012).

Il secondo aspetto riguarda il recupero del patrimonio testuale dell'ILC, già in formato digitale ma a rischio di estinzione a causa dell'evoluzione di codifiche e formati di rappresentazione. Fin dalle sue origini alla fine degli anni '70, l'ILC ha collaborato con importanti istituzioni pubbliche, culturali e accademiche nazionali per lo spoglio e l'elaborazione di archivi testuali che rappresentano importanti testimonianze del patrimonio culturale italiano. Ad oggi, una parte significativa dell'archivio testuale dell'ILC è costituita da testi caratterizzati da codifiche obsolete (EBCDIC, ASCII) e formati di rappresentazioni proprietari, che rendono problematici il recupero e la valorizzazione di tali archivi: oggi tale impresa appare impegnativa, ma con il passare del tempo il suo esito rischia di diventare sempre più incerto. Questa linea di attività è stata avviata all'interno di un accordo di collaborazione con l'Accademia della Crusca e al momento riguarda il recupero, la conservazione e la valorizzazione di importanti archivi testuali per lo studio dell'italiano post-unitario. La codifica e la rappresentazione originaria dei testi (che include anche annotazioni di varia natura, es. lemmatizzazione) è ricondotta al formato TEI (versione P5) che rappresenta ad oggi lo standard internazionalmente riconosciuto per la codifica digitale di testi umanistici.

Linea 2 - Progettazione e sviluppo di risorse e strumenti per il trattamento automatico di varietà storiche della lingua

Con la disponibilità sempre crescente di archivi testuali in formato digitale si è andata diffondendo la consapevolezza dell'ampio spettro di analisi che si aprono per questi testi e dunque per le discipline umanistiche di cui questi rappresentano l'espressione. Attraverso il processo di digitalizzazione i testi non solo diventano più facilmente accessibili, il che rappresenta già di per sé un risultato, ma possono essere interrogati a vari livelli di astrazione, che vanno anche al di là della loro rappresentazione superficiale. Interrogazioni astratte, ovvero basate non sulle stringhe di caratteri che costituiscono il testo ma su annotazioni di varia natura (es. morfo-sintattica, sintattica o semantica) associate ad esso o a sue porzioni, sono possibili a condizione che siano disponibili risorse linguistiche e strumenti per il trattamento automatico del testo.

Storicamente, i presupposti per ricerche più astratte venivano creati mediante l'arricchimento del testo con informazioni di varia natura (ad esempio, il lemma associato ad ogni forma del testo): tale processo era condotto tipicamente in modo manuale o – ove possibile – in modo semi-automatico. In tempi più recenti, grazie alla maturità raggiunta dalle tecnologie del linguaggio per l'annotazione linguistica del testo l'attenzione si è spostata sullo sviluppo di risorse e strumenti per il trattamento automatico di lingue antiche o di varietà storiche della lingua. All'interno di questa linea di attività, l'ILC è impegnato su diversi fronti, all'interno di progetti e di strategiche collaborazioni scientifiche a livello nazionale e internazionale, per un ampio spettro di lingue. I risultati consistono nella progettazione e lo sviluppo di risorse lessicali e terminologico-concettuali da un lato, e di strumenti per il trattamento automatico della lingua dall'altro.

Le risorse semantiche e lessicali sviluppate (in alcuni casi ancora in corso di sviluppo) fanno riferimento a modelli di rappresentazione ampiamente riconosciuti a livello internazionale. Esse includono:

- lessici computazionali secondo il modello WordNet per lingue classiche, in particolare Greco (Bizzoni *et al.* 2014), Latino e Arabo (sviluppo in corso). Tale attività riguarda sia l'acquisizione automatica dei "synset" candidati a partire dall'analisi di dizionari bilingui, sia la loro verifica e strutturazione all'interno di una rete di relazioni semantiche, sia il collegamento dei "synsets" estratti e validati alle risorse ItalWordNet per l'italiano o WordNet per l'inglese. Tali risorse, che vanno ad ampliare

la famiglia dei WordNets sviluppati a livello mondiale con importanti acquisizioni riguardanti lingue antiche, rappresentano un importante risultato di per sé ma anche utili strumenti a supporto dell'interrogazione e della navigazione di testi;

- risorse terminologico-concettuali relative a diversi domini e costruite a partire dall'edizione digitale dei manoscritti di Cristoforo Clavio, gesuita, matematico e astronomo tedesco vissuto nel '500, o del linguista Ferdinand De Saussure. In questo caso, il modello di rappresentazione lessicale adottato è costituito dal modello SIMPLE, alla base del "Lexical Markup Framework" che costituisce lo standard ISO per i lessici computazionali (Francopoulo 2013): tale modello è risultato il più adeguato in relazione alla tipologia di dati da trattare e alle finalità progettuali. Il lessico dedicato alla terminologia matematico-astronomica utilizzata da Clavius (il cui sviluppo è ancora in corso) così come quello dedicato alla terminologia linguistica di De Saussure (Piccini *et al.* 2013) sono parte delle edizioni digitali delle loro opere: oltre a costituire un importante risultato di per sé, tali risorse possono anche essere utilizzate a supporto della navigazione dei testi.

  Le risorse sviluppate includono anche corpora semanticamente annotati per lo studio dell'intertestualità di componimenti poetici plurilingui in greco, latino, italiano e arabo di natura epigrafica o letteraria (progetto PRIN 2010/11 "Memorata Poetis" che coinvolge numerose unità operative coordinate dall'Università di Venezia, Boschetti *et al.* 2014).

Sul versante degli strumenti per il trattamento automatico della lingua, gli sforzi sono concentrati a livello dell'analisi morfo-sintattica. Alle evoluzioni dell'analizzatore morfologico e lemmatizzatore del latino LEMLAT (Passarotti 2007), si affiancano:

- il raffinamento e l'estensione del componente software open-source "Aramorph" per l'analisi morfologica e la lemmatizzazione dell'arabo con un lessico esteso e filtri ortografici, sintattici e semantici, nell'ambito dell'Advanced Grant ERC (call Ideas 2009) "Greek into Arabic" (Nahli e Giovannetti 2013);
- analizzatori morfo-sintattici, il cui sviluppo è ancora in corso, per il greco antico così come per lingue semitiche come l'ebraico antico e l'aramaico (questi ultimi nell'ambito del progetto nazionale finanziato dal MIUR "Traduzione del Talmud Babilonese").

Va infine menzionato lo sviluppo di un analizzatore sintattico a dipendenze per il latino medievale: si tratta del parser stocastico "DeSR" adde-

strato sulla "Index Thomisticus Treebank", il cui algoritmo di analisi è stato specializzato in considerazione delle caratteristiche della lingua latina, in particolare l'ordine libero delle parole e la posizione finale del verbo all'interno della frase (Passarotti e Dell'Orletta 2010).

I risultati delle attività di ricerca e sviluppo raccolte all'interno di questa linea se da un lato costituiscono di per sé importanti e autonome acquisizioni dall'altro possono essere utilizzati per l'arricchimento dei testi creando i presupposti per interrogazioni avanzate basate sulla struttura linguistica e/o semantico-concettuale sottostante al testo.

Linea 3 - Progettazione e sviluppo di risorse e strumenti per il trattamento automatico di varietà storiche della lingua

Il testo, una volta digitalizzato ed eventualmente arricchito con annotazioni di varia natura provenienti da risorse e strumenti per il trattamento automatico del linguaggio, è pronto per essere predisposto per essere messo a disposizione della comunità scientifica, per il raffinamento con annotazioni manuali o traduzioni, per la consultazione da parte di più utenti, anche remoti, oppure per analisi qualitative e quantitative dei contenuti.

All'interno di questa linea di ricerca, le attività si suddividono all'interno di due filoni corrispondenti a due tipi di analisi del dato testuale, riconducibili alla dicotomia proposta da Franco Moretti (2005) per l'analisi letteraria "close reading" *vs* "distant reading". Con "close reading" si intende il tradizionale approccio accademico di "lettura ravvicinata", all'interno del quale l'attenzione è focalizzata sulle singole caratteristiche del testo, così come sulle sue variazioni e la sua storia; in altre parole, su di una sorta di micro-analisi dell'evidenza testuale, dei riferimenti, delle scelte lessicali e linguistiche, della semantica e delle scelte di stile. Invece, con "distant reading" o "lettura a distanza" si intende un approccio finalizzato ad estrarre generalizzazioni a partire dai testi: per dirla con le parole di Moretti, in questo approccio all'analisi del testo "la distanza fa vedere meno dettagli, vero: ma fa capire meglio i rapporti, i pattern, le forme".

Al primo filone di analisi testuale pertiene la progettazione e lo sviluppo di piattaforme per l'archiviazione, l'annotazione, l'interrogazione e la gestione di testi per: lo studio del testo, delle sue caratteristiche, delle sue varianti e della sua storia; la realizzazione di indici e concordanze; l'annotazione manuale; la creazione di una edizione critica. Tra le piattaforme recenti o ancora in corso di sviluppo vale la pena menzionare:

- la piattaforma Web per lo studio di traduzioni di testi antichi sviluppata all'interno del progetto ERC Advanced Grant "Greek into Arabic: Philosophical Concepts and Linguistic Bridges" (2010) finalizzato alla produzione dell'edizione critica della pseudo-Theologia di Aristotele (Bozzi e Marchi 2013);
- le applicazioni Web per lo studio e la produzione dell'edizione digitale dei manoscritti del matematico e astronomo Cristoforo Clavio (Abrate *et al.* 2014) o di quelli del linguista Ferdinand de Saussure (Del Grosso *et al.* 2013).

Gli strumenti e le tecniche che permettono esplorazioni ravvicinate del testo, in tutti i suoi minimi dettagli, rimangono e rimarranno sempre centrali nel lavoro quotidiano dell'umanista: l'arricchimento del dato testuale con annotazioni di varia natura, ottenute grazie all'inserimento all'interno della piattaforma di analisi testuale di risorse e strumenti per il trattamento automatico del linguaggio, contribuisce ad estendere le potenzialità di navigazione e ricerca all'interno del testo, facendo astrazione dal mero dato testuale.

L'approccio all'analisi del testo denominato "distant reading" rappresenta un nuovo paradigma di studio all'interno delle discipline umanistiche, che raccoglie al suo interno una varia tipologia di metodi e tecniche di analisi del testo. Al di là delle differenze legate al dominio e alla metodologia di analisi adottata, si tratta di analisi che fanno un passo indietro rispetto all'oggetto di analisi per poter osservare l'emergere di tendenze e generalizzazioni che vanno al di là del singolo testo/contesto. A questo filone sono riconducibili: gli studi dialettometrici sull'analisi aggregata dello spazio della variazione dialettale in Toscana, condotti all'interno di una collaborazione con l'Università di Groningen (Paesi Bassi) (Montemagni 2008, 2010); oppure l'analisi del significato lessicale attraverso la ricostruzione di una rete di rapporti topologici tra parole, determinati dalle loro modalità di co-occorrenza a livello sintagmatico a partire da un corpus di testi greci (Boschetti 2009); oppure, lo studio di sistemi di supporto alle decisioni che aiutano lo studioso nella formulazione di "ipotesi interpretative" relative ai testi e nella visualizzazione dei risultati delle proprie analisi (Bellandi *et al.* 2014; in stampa). A questo filone sono anche riconducibili i metodi e le tecniche di monitoraggio linguistico, finalizzato alla ricostruzione del profilo linguistico di collezioni di testi, rappresentative, ad esempio, di generi testuali o l'opera letteraria di un autore (Montemagni 2013; Dell'Orletta *et al.* 2013).

Linea 4 – Costruzione di un'infrastruttura italiana per le Digital Humanities

Le risorse e gli strumenti di analisi linguistica e testuale messi a punto all'interno delle precedenti linee di attività rappresentano contributi importanti per lo sviluppo del settore delle Digital Humanities. Tuttavia, perché questi singoli contributi possano trasformarsi in una rete di conoscenza distribuita a livello nazionale e internazionale è necessario definire nuove modalità di utilizzo, sviluppo e condivisione di strumenti e risorse a supporto della ricerca nel settore. Questo obiettivo è oggi perseguito creando e mettendo a disposizione infrastrutture di rete, condivise e distribuite, per la creazione, la fruizione, la distribuzione e la valutazione delle risorse e tecnologie linguistiche, che fungano da catalizzatore per lo sviluppo di una rete di eccellenza italiana e europea nei settori del trattamento automatico del linguaggio e delle Digital Humanities (Soria *et al.* in stampa). In questa ottica, appare cruciale la partecipazione dell'Italia alla rete europea CLARIN-ERIC (http://www.clarin.eu), in corso di avvio e all'interno della quale l'ILC costituirà uno dei pilastri dell'infrastruttura a livello nazionale. La creazione di CLARIN-IT consentirà alle comunità di ricerca nel settore delle scienze umane e sociali di trasformare la vasta collezione di risorse e infrastrutture locali esistenti o *in fieri* attualmente scollegate in un'unica infrastruttura di ricerca nazionale, integrandola al contempo con la rete esistente o in corso di sviluppo a livello Europeo.

CLARIN è un'infrastruttura nata all'interno della comunità internazionale della linguistica computazionale: come lo stesso acronimo dice, si tratta di una "Common Language Resources and Technology Infrastructure", ovvero di una infrastruttura che integra risorse e tecnologie linguistiche avanzate per l'accesso, l'elaborazione e l'analisi dei dati che stanno alla base delle ricerche nell'ambito delle scienze umane e sociali. In questo contesto, un interrogativo legittimo riguarda il rapporto tra CLARIN e l'altra infrastruttura di ricerca per il settore delle Digital Humanities, ovvero DARIAH, alla quale l'Italia ha già aderito. Si tratta di due iniziative che presentano interessanti aspetti di complementarità: appare esemplare in tal senso la recente iniziativa olandese di combinare i progetti CLARIN-NL e DIARIAH-NL all'interno di una nuova infrastruttura che integra dati (testi, immagini, dati strutturati e materiali audio-visivi) e strumenti per la loro analisi ed elaborazione. CLARIAH ("Common Lab Research Infrastructure for the Arts and Humanities") nasce come l'evoluzione naturale dei progetti nazionali CLARIN-NL (focalizzato sull'analisi linguistica di testi) e DIARIAH-NL (focalizzato sull'analisi di dati

socio-economici, storici e materiali strutturati) ai quali si è aggiunto un terzo aspetto, quello del trattamento di materiali audio-visivi. Un'analoga iniziativa potrebbe essere prospettata per l'Italia. In questo modo, sarebbe possibile federare le istituzioni italiane impegnate nella cura e salvaguardia del patrimonio culturale in un'autorevole rete integrata, dotare il Paese degli strumenti tecnologici per la fruizione e la condivisione del patrimonio culturale (costituito da un'ampia varietà di risorse, non limitato a risorse linguistiche e testuali), sopperire al problema dell'obsolescenza dei formati dei dati tramite un approccio basato sulla condivisione tecnologica e di know-how.

Per concludere, questa breve rassegna delle attività dell'ILC nel settore delle Digital Humanities mostra che esse costituiscono un settore strategico di applicazione dei risultati delle ricerche in corso, sul quale convergono i risultati di un'ampia gamma di attività che includono:

- lo sviluppo di metodi e tecniche per l'acquisizione e la conservazione di archivi testuali;
- lo sviluppo di risorse e strumenti per l'analisi automatica di testi antichi;
- lo sviluppo di piattaforme software a supporto dell'analisi testuale;
- lo sviluppo di metodi e strumenti per macro-analisi di corpora testuali;
- la costruzione di un'infrastruttura per l'integrazione e la condivisione delle risorse e degli strumenti sviluppati, che consoliderà l'ILC nella sua posizione di centro di riferimento nella rete delle Digital Humanities, nazionale e internazionale.

Dato il tema della tavola rotonda, in questo intervento mi sono focalizzata sui benefici che possono derivare per le Digital Humanities dal ricorso a tecnologie del linguaggio in senso lato, in particolare sulle nuove prospettive di analisi e ricerca che vengono ad aprirsi per le discipline umanistiche. Non si tratta tuttavia di un rapporto a senso unico: le Digital Humanities costituiscono infatti un fertile terreno di applicazione per lo sviluppo e il raffinamento di tecnologie innovative nel settore della linguistica computazionale, rendendo la sinergia ancora più stimolante.

2. Bibliografia

Abrate M., Del Grosso A. M., Giovannetti E., Lo Duca A., Luzzi D., Mancini L., Marchetti A., Pedretti I., Piccini S. (2014). *Sharing Cultural Heritage: the Clavius on the Web Project*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, 26-31 May 2014.

Bellandi A., Bellusci A., Carniani E., Giovannetti E. (2014). *Content Elicitation: Towards a New Paradigm for the Analysis and Interpretation of Text*. In Proceedings of the 13th IASTED International Conference on Software Engineering, Innsbruck, 2014.

Bellandi A., Bellusci A., Cappelli A., Giovannetti E. (in stampa). *Graphic Visualization in Literary Text Interpretation*. In Proceedings of the 18° International Conference of Information Visualization, University of Paris Descartes, 15-18 July, 2014.

Bizzoni Y., Boschetti F., Diakoff H., Del Gratta R., Monachini M., Crane G. (2014). *The Making of Ancient Greek WordNet*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, 26-31 May 2014.

Boschetti F. (questo volume). *Acquisizione e Creazione di Risorse Plurilingui per gli Studi di Filologia Classica in Ambienti Collaborativi*.

Boschetti F., Del Grosso A. M., Khan A. F., Lamé M., Nahli O. (2014). *A top-down approach to the design of components for the philological domain*. In Book of Abstracts of Digital Humanities Conference 2014 (DH2014).

Boschetti F. (2009). *Studio degli spazi semantici con strumenti informatici come metodo esplorativo per la valuatazione di congetture*. «Bollettino dei Classici», 30, pp. 41-53.

Bozzi A., Marchi S. (2013). *G2A: a Web application to study, annotate and scholarly edit ancient texts and their aligned translations.* «Studia Graeco-Arabica», 3, Pacini Editore.

Del Grosso A.M., Marchi S. (2013). *Una applicazione Web per la filologia computazionale. Un esperimento su alcuni scritti autografi di Ferdinand de Saussure*. In D. Gambarara e M.P. Marchese, a c. di, *Guida per un'edizione dei manoscritti di Ferdinand de Saussure*, Edizioni Dell'Orso, pp. 131-157.

Dell'Orletta F., Montemagni S., Venturi G. (2013). *Linguistic Profiling of Texts Across Textual Genre and Readability Level. An Exploratory Study on Italian Fictional Prose*. In Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013), 7-11 September, Hissar, Bulgaria, pp. 189-197.

Francopoulo G., a c. di (2013). *LMF Lexical Markup Framework*, Wiley-ISTE.

Lamé M., Valchera V., Boschetti F. (2012). *Epigrafia digitale: paradigmi di rappresentazione per il trattamento digitale delle epigrafi.* «Epigraphica», 74, pp. 386-392.

Montemagni S. (2008). *The space of Tuscan dialectal variation. A correlation study*. «International Journal of Humanities and Arts Computing», Edinburgh University Press, Oct 2008, vol. 2, no 1-2, pp. 135-152.

Montemagni S. (2010). *Esplorazioni computazionali nello spazio della variazione lessicale in Toscana*. In N. Prantera, A. Mendicino, C. Citraro, a c. di, Atti del Convegno *'Parole. Il lessico come strumento per organizzare e trasmettere gli etnosaperi'*, 2-4 luglio 2009, Rende, Centro Editoriale e Librario dell'Università della Calabria, 2010, pp. 619-644.

Montemagni S. (2013). *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*. «Studi Italiani di Linguistica Teorica e Applicata» (SILTA), Anno XLII, no 1, pp. 145-172.

Moretti F. (2005). *La letteratura vista da lontano*, Einaudi Editore.

Nahli O., Giovannetti E. (2013). *Computational contributions for Arabic language processing.* «Studia Graeco-Arabica», 3, Pacini Editore.

Passarotti M. (2007). *LEMLAT. Uno strumento per la lemmatizzazione morfologica automatica del latino*. «Papers on grammar», IX-3 (3), pp. 107-128 [http://hdl.handle.net/10807/1393].

Passarotti M., Dell'Orletta F. (2010). *Improvements in Parsing the Index Thomisticus Treebank. Revision, Combination and a Feature Model for Medieval Latin*. In Proceedings of LREC'10 – Seventh International Conference on Language Resources and Evaluation (Valletta, Malta, 17-23 May 2010), pp. 1964 – 1971.

Piccini S., Giovannetti E., Ruimy N. (2013). *Le lexique électronique de la terminologie de Ferdinand de Saussure: une première*. Actes du XXVII Congrès international de linguistique et de philologie romanes, Nancy, 15-20 July, 2013.

Soria C., Calzolari N., Monachini M., Quochi V., Bel N., Choukri K., Mariani J., Odijk J., Piperidis S. (in stampa). *The Language Resource Strategic Agenda: the FLaReNet synthesis of community recommendations*. «Language Resources and Evaluation», Springer.

Panels

The Digital Library to Support the Computer Humanist / La biblioteca digitale a supporto dell'umanista informatico

Digital libraries and digital humanities scholars: community context, workflow and collaboration*

Anna Maria Tammaro
Department of Information Engineering, University of Parma, Italy
annamaria.tammaro@unipr.it

Abstract. This paper introduces the Round Table about Digital Libraries and Digital Humanities Scholars. Despite more than fifteen years of digitization by digital libraries, the digital humanities scholars do not seem to have completely innovated the processes of scholarly communication, with the exception of a few pioneers and some specialized research centers such as those of the speakers invited to the Round Table. More research is needed to understand the institutional and community context of the digital scholarship for bridging the gap between digital libraries and digital humanities scholars.

Keywords: DELOS Reference Model, Virtual Research Environment, Digital Libraries.

1. Introduction

The concept of a digital library discussed in the Round Table is that of a virtual research environment (VRE) for digital humanities scholars and the value of the digital library is measured by its inclusion in the workflow of scholars, to become an indispensable tool, at least in some humanities disciplines, for facilitating research methodologies and innovative results. The concept of the digital library as infrastructure for e-science defines an organization that is able to integrate the complex technological infrastructure of the digital library with the needs of digital humanities communities, their collaborative relationships, the flow of organizational practices, in the broader context of the scholarly communication. The central idea of this concept is that the digital library should respond to the needs and research objectives pursued by the scholarly community and must be able to combine the infrastructure and digital content as part of a community based partici-

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

patory approach to service. In a standard cognitivist approach, the context is understood as a given cognitive models of the individual and cognitive structures as the context of problem solving, thinking and learning. In an other view, the context of the scholars includes the societal and cultural aspects of the scholarly community, attempting to define contexts as social situations, virtual space of interactive experience, fields of discourse. Contexts are here seen as interpersonal constructions which impact on the information behaviour of the individual, the community rules and institutional regulations and values.

1.1 *The DELOS Reference Model*

The DELOS European Network of Excellence on Digital Libraries (DELOS 2007) has described the digital library as "as a tool at the centre of intellectual activity having no logical, conceptual, physical, temporal or personal borders or barriers on information" (p. 17), and argued that digital libraries, in pursuit of the interaction with the user, "have evolved from a content-centric system that simply organises and provides access to particular collections of data and information to a person-centric system that aims to provide interesting, novel, personalised experiences to users" (p. 17). Moreover, Nicola Ferro has evidenced during the Round Table that the DELOS Reference Model distinguishes among three different "systems" which constitute the digital library universe and rely on the six domains introduced above for their definition: *Digital Library (DL)* is an organisation, which might be virtual, that comprehensively collects, manages and preserves for the long term rich *digital content*, and offers to its *user* communities specialised *functionality* on that content, of measurable *quality* and according to codified *policies*; *Digital Library System (DLS)* is a software system that is based on a defined (possibly distributed) *architecture* and provides all functionality required by a particular Digital Library. Users interact with a Digital Library through the corresponding Digital Library System; *Digital Library Management System (DLMS)* is a generic software system that provides the appropriate software infrastructure both (i) to produce and administer a Digital Library System incorporating the suite of functionality considered fundamental for Digital Libraries and (ii) to integrate additional software offering more refined, specialised or advanced functionality.

The DELOS model shows effectively that users are no longer passive users of the collections, but they are "creators" themselves of digital content

and active agents in the research cycle. What is lacking – or rather implicit in the *policy* concept – in this model DELOS, and instead should be better highlighted, is the importance of adapting to the context of the scholarly community, understanding the broader institutional context and its impact on individuals.

1.2 *The scholarly community context*

Digital libraries are tools for the creation of knowledge, ie they are tightly integrated into the cycle of academic research, bringing together digital resources and services to facilitate sharing and collaboration among the community of scholars in innovative ways. The concept of a digital library as a virtual space for collaboration of specific communities of scholars (Bishop 1996) overlaps with the "Virtual Research Environment VRE" or virtual research space. The digital library as VRE are designed to help scholars of all disciplines to manage the complex activities that are necessary to innovate the knowledge creation in the digital domain. Stefano Casati and Maurizio Lana have described the context of two different scholarly community, historians and philologists, and identified the user's activities and the innovation cycle of scientific communication in the digital environment.

Once text is digital, scholars realize that they can do "more" than simply offer content to be searched and to browse. Digital libraries as VRE represent an ecosystem of scholars, digital objects, infrastructure along with the values of the particular scholarly community context, to do this "more". How can the digital library put the activities of digital humanities scholars at the centre of its organization?

The digital library as a virtual research space, involves new partnerships and renewal of relationships and established practices. The Galileo//thek@ model called for new ways to collaborate amongst different professionals. Computer scientists, system analysts and librarians worked side-by-side with scholars and researchers to create a digital library of this kind of complexity.

It is not enough to collect digital content in a VRE, but the digital texts should be shared, personalized, reused, again distributed to a larger user base. Digital libraries assume the role of facilitator of innovation when they become "boundary objects" (Bishop *et al.* 2003) between communities with different specializations and levels of ability and are able to stimulate innovative use of content. Unsworth (2000) has defined what are the essential

activities (scholarly primitives) that digital humanities scholars do: find text, annotate, compare it with other texts, give examples, describe, interpret and make citations. Having access to a digital library of Latin texts, Lana explains the availability in digilibLT of the scholarly primitives on Latin texts, for a research which was previously not feasible.

What changes in the complex ecosystem of digital information distributed on the net? What activites are considered to be critical? The digital humanities scholars can now analyze large amounts of texts (corpora) and databases accessible on-line on which to perform the essential activities defined by Unsworth and the computer offers a different perspective from before, when it was possible to analyze a text at a time (Nichols 2013). The digital ecosystem has also made it possible collaborative research: this is a real innovation, which increases the value of research, for the possibility of integrating the analysis of texts with multidisciplinary skills and expertise. To achieve this collaboration, the digital library must understand that digital text is not only a feature of the digital library but a form of organization of digital humanities. The Project Geolat, described by Maurizio Lana is an example of this functionality.

2. Cultural institutions and e-infrastructures context

Rossella Caffo has described the European goal for the near future: to develop policies and strategies at the European level based on collaboration between the cultural heritage sector, the research sector and the e-Infrastructure technologies. This is a challenging job, because these organizations are organized differently from country to country, and the data and also the tools and services developed by each of these communities are not easily integrated with each other. The objective is that of creating a federated infrastructure for interoperability, storage, preservation of digital cultural heritage and the development of virtual research environment for scholarly communities linked to it. But how to harmonize the different policies? the main focus of the ongoing European projects is on establishing a strategic and operational coordination between cultural heritage institutions and providers of e-infrastructure for storage and preservation of digital content but there is still research to be done to better understand the role of users and the scholarly community.

3. Conclusions

Despite more than fifteen years of digitization by digital libraries, the digital humanities do not seem to have completely innovated the processes of scholarly communication, with the exception of a few pioneers and some specialized research centers. This delay may be due to the lack of the e-infrastructure, such as a digital library as VRE?

The functionality of VRE were actually carried out by research centres, not from digital libraries. The research centres have been opened in many universities, to put together resources and technologies and stimulate the development of new technological tools and to facilitate the collaboration of scholars, computer scientists and librarians, but outside of the digital libraries (Zorich 2008). However, research centres cannot provide the same support which could be given by the digital libraries. For example, the centres are not concerned with the curation and preservation of data, do not offer a collection service for the general users, not disseminate the results of research done by scholars through their repositories and research platforms.

4. References

Bishop A. P., Star S. L. (1996). *Social informatics of digital library use and infrastructure*. In M. Williams, ed., *Annual review of information science and technology*, vol. 31, pp. 301-401. Information Today.

Bishop A. P., Mehra B., Bazzell I., Smith C. (2003). *Participatory Action Research and Digital Libraries: Reframing Evaluation*. In A. P. Bishop, N. A. Van House, B. P. Buttenfield, eds., *Digital Library Use: Social Practice in Design and Evaluation*, edited by, pp. 161-189. MIT Press.

Candela L., Castelli D., Ferro N., Ioannidis Y., Koutrika G., Meghini C., Pagano P., Ross S., Soergel D., Agosti M., Dobreva M., Katifori V., Schuldt H. (2007). *The DELOS Digital Library Reference Model*. Foundations for Digital Libraries. ISTI-CNR at Gruppo ALI, Pisa, Italy.

Nichols S. J. (2013). *Anxiety of irrelevance Digital humanities and contemporary literary theory.* URL=https://www.academia.edu/4441824/The_Anxiety_of_Irrelevance_Digital_Humanities_and_Contemporary_Literary_Theory.

Zorich D. M. (2008). *A Survey of Digital Humanities Centers in the United States*. Tech. rep., OCLC Programs and Research, Dublin, Ohio, USA, 78 pp.

e-Infrastructures per le esigenze della ricerca*

Rossella Caffo
Direttore dell'Istituto Centrale per il Catalogo Unico delle biblioteche italiane, Roma, Italia
rosa.caffo@beniculturali.it

Abstract. L'obiettivo per il prossimo futuro è di sviluppare politiche e strategie a livello europeo che siano basate sul dialogo tra il settore culturale, il settore della ricerca e le e-Infrastructure: ICCU è impegnato nella realizzazione di questa infrastruttura ma è un lavoro impegnativo, perché questi soggetti sono organizzati in modo diverso da paese a paese ed i dati ed inoltre gli strumenti e i servizi messi a punto da ciascuna di queste comunità non sono facilmente integrabili tra loro.

Keywords: Dariah, e-infrastructure, digital humanities, cultural heritage.

1. Introduzione

L'accelerazione della rivoluzione digitale contribuisce in modo decisivo alla concezione di un nuovo modello, aperto e diffuso del patrimonio culturale e le potenzialità che offrono le tecnologie digitali pongono questioni cruciali, quali la proprietà intellettuale, la conservazione del patrimonio culturale digitale, lo sviluppo di nuovi servizi legati alla valorizzazione e fruizione del patrimonio culturale per il turismo e per la didattica in funzione di una rinnovata memoria e identità culturale, condivisa a livello europeo. L'innovazione, la creatività e l'interesse del pubblico per il patrimonio culturale sono elementi decisivi per lo sviluppo di una *e-Infrastructure* europea per il patrimonio culturale digitale, in grado di offrire nuovi servizi ad un bacino sempre più ampio di utenti.

L'obiettivo per il prossimo futuro è di sviluppare politiche e strategie a livello europeo che siano basate sul dialogo tra il settore culturale, il settore della ricerca e le *e-Infrastructure*: si tratta di un lavoro impegnativo, perché questi soggetti sono organizzati in modo diverso da paese a paese ed i dati ed inoltre gli strumenti e i servizi messi a punto da ciascuna di queste co-

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

munità non sono facilmente integrabili tra loro. Infatti le *e-Infrastructure*, oggi utilizzate nell'ambito delle e-Science sono in grado di offrire soluzioni efficienti anche per la creazione e gestione dei dati prodotti nel settore delle scienze umanistiche e del patrimoni culturale, dall'acquisizione all'accesso e alla gestione, allo storage e conservazione a lungo termine, collegando comunità di ricercatori che operano in diversi settori grazie al sistema di Identità Federato per l'accesso e l'uso dei dati e dei servizi messi a punto dalle Infrastrutture di Ricerca.

1.1 *Il ruolo di* Iccu

La prospettiva è quella di realizzare un'infrastruttura federata dedicata agli istituti culturali per la connettività, lo storage, la preservazione del patrimonio culturale digitale e per lo sviluppo di *virtual research communities* ad esso collegate.

Ciò implica:

- stabilire un coordinamento strategico e operativo tra gli istituti culturali e i provider di e-infrastrutture;
- definire una roadmap e una serie di strumenti pratici a supporto degli istituti culturali;
- sfruttare i common e-Infrastructure layers (federazioni di identità, federated cloud, servizi di infrastruttura di dati, ecc);
- definire il ruolo degli utenti e delle comunità di ricerca.

L'Istituto Centrale per il Catalogo Unico delle biblioteche italiane, per conto del Ministero italiano dei Beni e delle Attività Culturali e del Turismo, da anni impegnato sul fronte del dibattito europeo per la digitalizzazione e l'accesso in rete al patrimonio culturale, ha avviato due linee di programmazione complementari: 1) da un lato l'impulso alla digitalizzazione per favorire la diffusione della conoscenza e lo sviluppo delle industrie creative e del riuso delle risorse digitali, intesi come volano di nuova produttività, 2) dall'altro l'armonizzazione dei programmi nazionali di ricerca sul patrimonio culturale digitale e lo sviluppo di un'infrastruttura europea per l'interoperabilità dei sistemi nazionali di gestione e accesso al patrimonio.

1.2 *Progetti europei* Dc-Net, Indicate, Dch-Rp

L'Iccu ha coordinato tre progetti che hanno indagato l'impatto delle e-Infrastructure nell'ambito del patrimonio culturale, hanno avviato il dialogo

e la collaborazione tra i settore tecnologico delle e-Infrastructure e della ricerca con quello delle istituzioni del patrimonio culturale e individuato le criticità e le priorità necessarie per lo sviluppo di un piano d'azione comune.

L'avvio del dialogo è stato prima realizzato con il progetto DC-NET (2009-2012) e poi rinnovato con i progetti INDICATE (2010-2012), e DCH-RP che hanno indagato gli aspetti politici e tecnici che riguardano la relazione tra il settore del patrimonio culturale e le e-infrastructure.

DC-NET (Digital Cultural Heritage Network)[1] è stato un progetto ERA-NET (European Research Area Network) finanziato dalla Commissione Europea nell'ambito del Settimo Programma Quadro (FP7), programma specifico e-infrastructure, avviato nel 2009 e conclusosi nel 2012.

DC-NET ha sviluppato e consolidato il coordinamento dei programmi europei di ricerca pubblica nel settore del patrimonio culturale digitale. Con il coordinamento dell'ICCU, otto partner iniziali più altri cinque che si sono aggiunti in corso d'opera hanno avviato un piano di attività congiunte con il fine di realizzare una nuova infrastruttura digitale per la ricerca nel campo del patrimonio culturale digitale, che contenga una massa critica di contenuti e offra un'ampia gamma di servizi e strumenti che ne facilitino l'integrazione e l'analisi. Le attività sviluppate nell'ambito di gruppi di lavoro nazionali e internazionali, hanno visto la partecipazione di rappresentati del mondo della ricerca ICT, del patrimonio culturale e delle *e-infrastructure*. Tra i risultati principali del progetto è un documento strategico, '*Service Priorities and best Practice for Digital Cultural Heritage*'[2] condiviso dai tredici paesi partecipanti, che individua i temi prioritari per la ricerca sul patrimonio culturale digitale e un'analisi dettagliata dei servizi esistenti in Europa per la *digital preservation*.

Il progetto INDICATE (International Network for a Digital Cultural Heritage e-Infrastructure)[3], avviato nel 2010 e conclusosi nel 2012, è stata un'azione di coordinamento, sostenuta dalla Commissione europea nell'ambito del Capacities Programme of FP7. Il progetto, coordinato dall'ICCU, ha visto la partecipazione di otto paesi partner: Francia, Grecia, Italia, Slovenia, Spagna in Europa ed Egitto, Giordania e Turchia nell'area del Mediterraneo. INDICATE era integrato con il progetto DC-NET di cui rappresentava la prima messa in pratica dei suoi risultati. Grazie a INDICATE si è avviato il coordi-

[1] http://www.dc-net.org/
[2] http://www.dc-net.org/getFile.php?id=450
[3] http://www.indicate-project.eu/

namento dei piani di ricerca sul patrimonio culturale sviluppati intorno ai servizi offerti dalle *e-infrastructures* nei paesi attorno all'area del Mediterraneo ed ha realizzato esperimenti pilota e definito casi di studio che hanno funzionato da modelli di riferimento per le istituzioni culturali interessate all'utilizzo di piattaforme basate sulle *e-infrastructures*. Tra i maggiori risultati del progetto, elenchiamo:

- La Paris Declaration[4], una visione condivisa del Consorzio INDICATE per lo sviluppo del Patrimonio culturale digitale negli anni futuri;
- La pubblicazione di '*Handbook on virtual exhibitions and virtual performances*'[5];
- La pubblicazione di '*Best practice for applying research pilots and use case studies to digital cultural heritage*'[6].

Il progetto DCH-RP (Digital Cultural Heritage Roadmap for Preservation)[7] è un'azione di coordinamento sostenuta dalla Commissione europea nell'ambito di EC FP7 e-Infrastructures Programme. Il progetto partito nel 2012 ha sviluppato l'indagine già avviata dai progetti INDICATE e DC NET sugli strumenti e i servizi esistenti per la conservazione del digitale, da cui è emerso un gap tra le soluzioni disponibili e le esigenze degli istituti culturali. Se per la digitalizzazione dei contenuti culturali gli approcci comuni e le buone pratiche sono in generale ben sviluppate, la conservazione del digitale è ancora un settore molto frammentato, in cui mancano flussi di lavoro condivisi e strumenti efficienti a disposizione degli istituti culturali. È necessario perciò migliorare le pratiche di conservazione digitale nelle istituzioni culturali e predisporre le *e-Infrastructures* della ricerca alle esigenze di musei, biblioteche e archivi, con un focus particolare ai servizi offerti da NREN, NGI e altre infrastrutture di dati che integrano servizi GRID e CLOUD, in quanto canali efficaci per la fornitura di tecnologie avanzate.

Le soluzioni attualmente disponibili richiedono infatti sempre l'adattamento al compito specifico dell'istituzione, alla infrastruttura tecnologica a disposizione e alle competenze del personale. Inoltre nell'uso dei sistemi oggi a disposizione si presentano inevitabilmente problemi legati all'interoperabilità tecnica e semantica e alle barriere legali: in generale gli strumenti e/o i servizi esistenti nell'ambito della conservazione del digitale non sono

[4] http://www.indicate-project.org/index.php?en/187/paris-declaration
[5] http://www.indicate-project.org/getFile.php?id=412
[6] http://www.indicate-project.org/getFile.php?id=418
[7] http://www.dch-rp.eu/

sufficientemente sviluppati per soddisfare in modo efficace le esigenze degli istituti culturali.

Per colmare questo gap è necessario partire innanzitutto dalla elaborazione di un modello di conservazione distribuito basato sulle e-Infrastructures in grado di soddisfare le esigenze individuate dalla comunità degli istituti della memoria impegnati nello sviluppo di best practise di conservazione del patrimonio culturale.

DCH-RP cerca di offrire una roadmap coerente e realistica che può supportare i responsabili politici e i manager degli istituti nella pianificazione di programmi mirati alla conservazione del digitale.

In sintesi gli obiettivi del progetto DCH-RP sono:

- armonizzare le politiche di conservazione dei dati nel settore dei beni culturali digitali a livello europeo e internazionale;
- identificare i modelli più adatti per la gestione e la sostenibilità di un infrastruttura dedicata alla conservazione dei contenuti digitali;
- elaborare una Roadmap per lo storage e la conservazione;
- realizzare un Registro di strumenti e servizi;
- condurre dei test sperimentali (Proof of concept).

La proposta è di riunire e integrare le infrastrutture esistenti della ricerca con servizi e dati sul patrimonio culturale, in modo da consentire ai ricercatori e agli studiosi dei beni culturali di utilizzare i dati distribuiti e le nuove tecnologie per migliorare la metodologia di ricerca nel campo della valorizzazione e conservazione del patrimonio culturale. Si tratta di una infrastruttura per sostenere la ricerca digitale realizzata dalle scienze umane e le arti, attraverso l'applicazione e l'uso della tecnologia, utilizzando metodi e contenuti digitali, inserito in tutto il ciclo di vita di ricerca e incorporando il ciclo di vita dei contenuti / dati per garantire la cura (digital curation) dei dati come parte del processo di ricerca.

2. Dariah e Ariadne

DARIAH (Digital Research Infrastructure for the Arts and Humanities)[8] è un'infrastruttura di ricerca a supporto delle scienze umane e delle arti ed è ha l'obiettivo di creare un "ambiente" in cui condividere tecnologie digitali, dati, strumenti e metodologie innovative. DARIAH è stata realizzata prima

[8] http://www.dariah.eu/

come un progetto preparatorio, e recentemente come entità costituita nella veste di persona giuridica di diritto europeo denominata ERIC (European Research Infrastructure Consortium). L'Italia ha aderito a DARIAH attraverso i Ministeri dei beni culturali e della ricerca e università, e creato un gruppo di lavoro nazionale, DARIAH-IT, guidato dal CNR e con la partecipazione dell'ICCU per conto del MIBACT.

L'infrastruttura DARIAH favorisce la collaborazione multidisciplinare di diversi settori della ricerca che operano nel campo delle scienze umane, promuove l'interoperabilità e la condivisione di servizi digitali, facilita l'accesso a lungo termine e l'uso dei dati digitali prodotti dalla ricerca e rafforza le competenze e lo scambio di standard e buone pratiche nei paesi partecipanti. DARIAH è infatti una rete di persone, risorse informative, tecnologie, strumenti e metodologie per individuare, esplorare e sostenere il lavoro e la ricerca nel campo delle Digital Humanities.

L'Italia coordina inoltre un importante progetto per la realizzazione di una infrastruttura di ricerca nel campo dell'archeologica: il progetto ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe)[9], coordinato dal Laboratorio Servizi Didattici e Scientifici (Pin) dell'Università di Firenze finanziato dalla Commissione europea nell'ambito di EC FP7 e-Infrastructures Programme. ARIADNE è una Azione di coordinamento per lo sviluppo di una infrastruttura di ricerca per la gestione e l'integrazione dei dati archeologici a livello europeo. L'ICCU è partner del progetto: coordina il gruppo italiano formato dalla DG Antichità e dalle soprintendenze archeologiche di Roma, del Lazio e dell'Etruria Meridionale.

3. Conclusioni

L'ICCU ha inoltre coordinato una serie di progetti correlati ad Europeana, come ATHENA (2008-2011), LINKED HERITAGE (2011-2013) e ATHENA PLUS (2013-2015), coinvolgendo centinaia di istituzioni culturali per la digitalizzazione e l'integrazione in rete di contenuti culturali digitali e il loro ri-uso, dando impulso allo sviluppo delle industrie creative. In vista del semestre di presidenza italiana dell'Unione europea l'ICCU sta organizzando due importanti eventi internazionali:

[9] http://www.ariadne-infrastructure.eu/

Roma, 2 ottobre 2014, Biblioteca Nazionale, Conferenza internazionale sul tema del riuso del patrimonio culturale digitale per la didattica, il turismo e l'*edutainment*. La Conferenza ospiterà tre sessioni: la prima dedicata al riuso e alla scoperta del patrimonio culturale digitale, la seconda illustrerà buone pratiche nei settori dell'istruzione, dell'edutainment e del turismo, la terza si focalizzerà sui risultati di alcuni progetti europei rispetto alla creatività e ai bisogni degli utenti.

Roma 13 e 14 novembre 2014, Biblioteca Nazionale, Conferenza internazionale "*Infrastrutture digitali e infrastrutture di ricerca per il patrimonio culturale*". Il programma della conferenza presenterà una sessione di apertura con le considerazioni politiche, una sessione che introduce i servizi offerti dalle Infrastrutture digitali e alcuni aspetti organizzativi per la cooperazione internazionale della ricerca e una sessione in cui saranno presentate le più importanti infrastrutture di ricerca operanti nel settore dei beni culturali, delle arti e delle scienze umane.

(Formal) Models for systems, infrastructures, communities, and cultures*

Nicola Ferro
Department of Information Engineering, University of Padua, Padova, Italy
ferro@dei.unipd.it

Abstract. This paper highlight the role (formal) model for digital libraries can play to gap the bridge between different communities and cultures, such as libraries and archives, in order to enable interoperability among systems and infrastructures.

Keywords: DELOS Reference Model, 5S Model, Libraries, Archives.

1. Digital Libraries Models

Since the field of digital libraries has come to light in the early nineties of the past century, a lot of improvements and a dramatic change in the viewpoint has happened. In the beginning, digital libraries were almost monolithic systems, each one built for a specific kind of information resources – e.g. text, images, or videos – and with very specialised functionalities developed ad-hoc for those contents. This approach caused a flourishing of systems where the very same functionalities, e.g. user management or repositories, were developed and re-developed from scratch many times, causing them to be different and often incompatible one with the other.

With the passing of time and by exploiting the previous research results and achievements, a more mature way of facing the design and development of digital libraries has taken place. Digital libraries moved from being monolithic systems to being component and service-base systems, where easily configurable and deployable services can be plugged together and re-used in order to create a digital library. Moreover, digital libraries started to be seen as more and more user-centered systems, where the original content management task is partnered with new communication and cooperation tasks, so that digital libraries become "a common vehicle by which everyone will

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

access, discuss, evaluate, and enhance information of all forms" (Ioannidis *et al.* 2005).

In this evolving scenario, the design and development of effective services which foster the cooperation among users and the integration of heterogeneous information resources become a key factor which needs to be pursued by researchers and developers. A relevant example of this kind of new services are annotations, i.e. providing users or groups of users with the possibility of adding personal annotations on the managed information resources, even crossing the boundaries of the single digital library (Agosti *et al.* 2013; Agosti and Ferro 2008).

In this context, building foundations and a formal theory for digital libraries is a longstanding issue in the field, dating back to the mid-1960s (Licklider 1965), and this challenge has been accepted only very recently, for example, by the 5S model (Gonçalves *et al.* 2004) and the DELOS Reference model (Candela *et al.* 2007).

1.1 *The DELOS Reference Model*

The DELOS Reference Model lays the foundations of digital libraries and defines what are the constituent entities and stakeholders of the digital library universe as well as the relationships among them; in particular, the reference model provides a clear picture of what a digital library is and on what concepts and functionalities we can leverage in order to promote co-operation and interoperability (Candela *et al.* 2007). The DELOS Reference Model approaches the problem of modelling the digital library universe by highlighting six domains or main concepts: *content* is the data and information that digital libraries handle and make available to their users; *user* is the actors (whether human or not) entitled to interact with digital libraries; *functionality* is the services that digital libraries offer to their users; *quality* is the parameters that can be used to characterize and evaluate the content and behaviour of digital libraries; *policy* is a set of rules that govern the interaction between users and digital libraries; *architecture* is a mapping of the functionality and content offered by a digital library onto hardware and software components.

Moreover, the DELOS Reference Model distinguishes among three different "systems" which constitute the digital library universe and rely on the six domains introduced above for their definition: *Digital Library (DL)* is an

organisation, which might be virtual, that comprehensively collects, manages and preserves for the long term rich *digital content*, and offers to its *user* communities specialised *functionality* on that content, of measurable *quality* and according to codified *policies*; *Digital Library System (DLs)* is a software system that is based on a defined (possibly distributed) *architecture* and provides all functionality required by a particular Digital Library. Users interact with a Digital Library through the corresponding Digital Library System; *Digital Library Management System (DLMS)* is a generic software system that provides the appropriate software infrastructure both (i) to produce and administer a Digital Library System incorporating the suite of functionality considered fundamental for Digital Libraries and (ii) to integrate additional software offering more refined, specialised or advanced functionality.

1.2 *The 5S Model*

The Streams, Structures, Spaces, Scenarios, Societies (5S) (Gonçalves *et al.* 2004) is a formal model for digital libraries based on the following abstractions: *Streams* are sequences of elements of an arbitrary type (e.g. bits, characters, images) and thus they can model both static and dynamic content; *Structures* are the way through which parts of a whole are organised. In particular, they can be used to represent hypertexts and structured information objects, taxonomies, system connections and user relationships; *Spaces* are sets of objects together with operations on those objects conforming to certain constraints; *Scenarios* are sequences of events that may have parameters, and events represent state transitions; *Societies* are sets of entities and relationships. The entities may be humans or software and hardware components, which either use or support digital library services.

As shown in Figure 1, from the five abstractions of streams, structures, spaces, scenarios, and societies, a series of concepts are derived, which are then used to define what a digital library is. Indeed, in accordance with this framework, a minimal digital library is defined a constituted by: a *repository*, that is a service encapsulating a family of collections and specific services to manipulate the collections; a set of *metadata catalogues* for all the collections in the repository; a set of *services* containing, at least, services for indexing, searching and browsing; and, a society whose information needs have to be satisfied. As you can note from Figure 1, only three out of the six domains of the DELOS Reference Model are taken into consideration in the

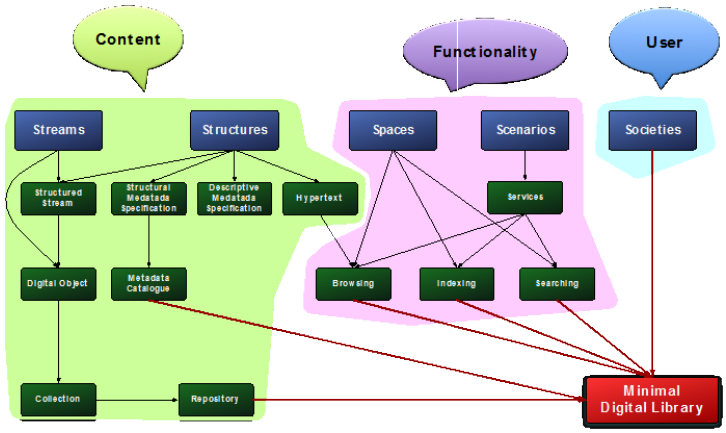


Fig. 1. Main definitions of the 5S model and their relationships with the domains of the DELOS Reference Model.

5S model, namely the Content, Functionality, and User domains; the other three – Quality, Policy, and Architecture – are not dealt with but are left to additional models that can be built starting from the 5S model.

2. Bridging between Libraries and Archives

In the context of Libraries, Archives, and Museums (LAM) unifying a variety of organizational settings and providing more integrated access to their contents is an aspect of utmost importance. Indeed, LAM collect, manage and share digital contents; although the type of materials may differ and professional practices vary, LAM share an overlapping set of functions. Fulfilling these functions in "collaboration rather than isolation creates a win–win for users and institutions" (Zorich *et al.* 2008). Although the convergence between libraries, archives and museums has been a topic of much discussion in the digital library community, the emerging similarities between these three types of cultural heritage institutions are not yet evident in the proposed formal models, developed systems, and education of professionals (Trant 2009; Timms and Fall 2009).

Archives are a fundamental constituent of our cultural heritage and digital libraries are the natural choice for managing and providing access to their

assets. Unfortunately, there have been almost no formal models for archives and this has prevented them from being fully integrated in digital library communities, methodologies and technologies. We think that the archival domain deserves a formal theory as well and that this theory has to be reconciled with the more general theories for digital libraries in order to provide archives with the full breadth of methodologies and technologies which have been developed over the last two decades in the digital library field.

We proposed the NESTOR formal model (Ferro and Silvello 2013) to settle a common ground for dealing with hierarchies open to existing models, solutions and technologies. The set data models composing NESTOR are well-suited for archival practice; indeed, the idea of "set" shapes the concept of archival division which is a "container" comprising distinct elements that have some properties in common. If we consider the Chinese boxes metaphor, a hierarchy is composed of a sequence of boxes contained one inside the other; if we look at an archive from the physical point-of-view, we can see that it resembles the Chinese boxes structure as there are boxes, folders, sheets, etc. contained one inside the other. Nested sets are closer to this view of reality than trees are. Indeed, although archival practice commonly considers ar- chives as trees, a tree is actually a higher level abstraction than the nested sets as it only focuses on structural relationships. Indeed, NESTOR comprises both the structure and the content of the archive, where the inclusion relationships represent the structure and the elements belonging to the sets represent the content.

Then, we extended the 5S model to introduce the notion of digital archive as a specific case of digital library complying with the NESTOR archival constraints. This, in turn, will open up the possibility to further extend the 5S model. Indeed, according to this model, a minimal digital library has to offer indexing, searching and browsing services (Gonçalves *et al.* 2004). The formal definition of the query and update operations in NESTOR will thus allow us to precisely describe what these services are in the case of digital archives.

3. References

Agosti M., Conlan O., Ferro N., Hampson C., Munnelly G. (2013). *Interacting with Digital Cultural Heritage Collections via Annotations: The CULTURA Approach*. In Proc. 13th ACM Symposium on Document Engineering (DocEng 2013), ACM Press, pp. 13-22.

Agosti M., Ferro N. (2008). *A Formal Model of Annotations of Digital Content*. «ACM Transactions on Information Systems» (TOIS), 26, 1, 3, pp. 3-57.

Candela L., Castelli D., Ferro N., Ioannidis Y., Koutrika G., Meghini C., Pagano P., Ross S., Soergel D., Agosti M., Dobreva M., Katifori V., Schuldt H. (2007). *The DELOS Digital Library Reference Model*. Foundations for Digital Libraries. ISTI-CNR at Gruppo ALI, Pisa, Italy.

Ferro N., Silvello G. (2013). *NESTOR: A Formal Model for Digital Archives*. «Information Processing & Management», 49, 6, pp. 1206-1240.

Gonçalves M.A., Fox E.A., Watson L.T., Kipp N.A. (2004). *Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries*. «ACM Transactions on Information Systems» (TOIS), 22, 2, pp. 270-312.

Ioannidis Y., Maier D., Abiteboul S., Buneman P., Davidson S., Fox E.A., Halevy A., Knoblock C., Rabitti F., Schek H.-J., Weikum G. (2005). *Digital library information-technology infrastructures*. «International Journal on Digital Libraries», 5, 4, pp. 266-274.

Licklider J.C.R. (1965). *Libraries of the future*. The MIT Press.

Timms K. (2009). *New partnerships for old sibling rivals: The development of integrated access systems for the holdings of archives, libraries, and museums*. «Archiviaria», 68, pp. 67-96.

Trant J. (2009). *Emerging convergence? Thoughts on museums, archives, libraries, and professional training*. «Museum Management and Curatorship», 24, 4, pp. 369-387.

Zorich D.M., Waibel G., Erway R. (2008). *Beyond the silos of the LAMs: Collaboration among libraries, archives and museums*. Tech. rep., OCLC Programs and Research, Dublin, Ohio, USA.

Biblioteche digitali e studi umanistici*

Maurizio Lana
Dipartimento di Studi Umanistici/Università del Piemonte Orientale, Vercelli, Italia
m.lana@lett.unipmn.it

Abstract. Il complesso e costoso lavoro di creazione di biblioteche digitali non si può giustificare solo sulla base della semplice creazione di una risorsa per la lettura dei testi nell'ambito digitale. Occorre che le biblioteche digitali siano utilizzate per attività complesse non praticabili con le edizioni a stampa. Una di esse è l'annotazione formale dei testi e l'impiego dell'annotazione per dar luogo a nuove modalità di lavoro sul testo stesso.

Parole chiave: geografia, ontologia, TEI, Linked Open Data.

1. Introduzione

Accade in genere che davanti ad un fenomeno nuovo si provi in primo luogo a leggerlo in analogia con qualcosa di preesistente che appare simile. Fotografia digitale e fotografia tradizionale (analogica) sono diversissime, le loro somiglianze in fin dei conti sono marginali. Però si parla sempre di fotografia, stampe, esposizione, fotocamera, e così via. La diffusione è facilitata dal fatto che ciò che nella forma tradizionale appariva difficile – fare buone foto – sembra ora facile con tutte le preimpostazioni disponibili per cui le foto sbiancate per la sovraesposizione o nere di sottoesposizione non esistono più. Così facendo però si perde di vista che la fotografia digitale permette un controllo diretto sulla produzione dell'immagine più ampio e diretto di quello che era possibile con la pellicola. E che permette anche operazioni non possibili con la pellicola.

Con le biblioteche digitali si rischia che accada qualcosa di simile. Costruire una biblioteca digitale, soprattutto se si tratta di una biblioteca costruita a partire da fonti a stampa digitalizzate e poi accuratamente corrette e annotate, richiede investimenti cospicui in tempo e denaro. E nel corso

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

degli anni occorre aggiornare il formato dei dati, il software di gestione della biblioteca, e le opere digitalizzate sono meno leggibili che se si trovassero su un supporto fisico: la conservazione è delicata e difficile – non basta 'tenere i libri lì', occorre curarli e accudirli costantemente come bottiglie di champagne nella *cave*.[1]

Di qui la domanda: perché costruire biblioteche digitali? La domanda ovviamente si pone essenzialmente per le biblioteche le cui collezioni sono state digitalizzate, perché quando esse invece custodiscono collezioni *born digital* la risposta in certo modo è apparentemente data dal fatto stesso che esistono contenuti che nascono digitali – libri, giornali, film, musiche, giochi e dunque richiedono una biblioteca digitale che li raccolga e li renda disponibili. La risposta è che la digitalizzazione deve essere il ponte, la rampa di lancio, verso utilizzi, attività, impossibili o comunque non facili da realizzare se i contenuti si trovano ancorati ad un supporto fisico.

Dunque – provocatoriamente e paradossalmente – una biblioteca che contenga collezioni *digitalizzate*[2] non ha come come scopo primario di offrire l'accesso in lettura ai libri che possiede: le edizioni a stampa già esistenti sono più facili da gestire, distribuire, leggere. In misura più o meno grande, tutti i modi di accesso ai contenuti digitali che ripropongano *primariamente* le forme consuete per l'accesso ai corrispondenti contenuti su supporto fisico costituiscono uno spreco di risorse. Il nucleo della questione sta nell'annotazione: cioè nello scrivere in modo formalizzato informazioni associate ad uno specifico passo di testo. Nulla di nuovo: è 'solo' la versione tecnologicamente evoluta dell'annotazione a margine che qualsiasi lettore esperto è abituato a fare, e che ha la sua forma più illustre negli *scholia* che ci testimoniano l'attività di commentatori medievali delle opere della letteratura greco-latina. Non solo: il testo, soprattutto letterario (ma in realtà tutti, in vario modo) vive solo nella lettura di un soggetto che con quell'atto lo attiva. Perché il lettore interpreta ciò che legge, individua significati, costruisce visioni delle cose, a partire dal testo letto. Legge con un'intenzionalità, con uno scopo grazie ai quali acquisisce conoscenza, individua informazioni, formula ragionamenti, relativi a specifici punti o passi del testo. Tutto ciò si

[1] Più obbligato il percorso per le biblioteche digitali che danno accesso a collezioni nativamente digitali: riviste scientifiche, contenuti audio o video o immagini fotografiche.

[2] Proprio solo per semplicità espositiva parleremo nella pagine seguenti di *libri* come contenuto delle biblioteche digitalizzate, ma essi sono linguisticamente solo un modo semplice per indicare "i contenuti digitalizzati, di qualsiasi tipo".

può tradurre in annotazioni formalmente rigorose che possono essere riusate e rielaborate: e ciò è precisamente uno dei modi in cui si manifesta lo scopo proprio di una biblioteca digitale, non semplicemente leggere i testi, ma lavorare su di, e con, essi in modi altrimenti (praticamente) impossibili.

Il percorso quindi è segnato da queste tappe:

> testo a stampa – digitalizzazione – testo digitale – lettura/correzione – annotazione – rielaborazione del testo annotato.

Nelle pagine che seguono si descriverà una biblioteca digitale in corso di realizzazione e si mostrerà come la sua esistenza permetta di costruire un progetto di ricerca altrimenti irrealizzabile.

2. La biblioteca digitale digilibLT

Gli studiosi del mondo classico hanno a disposizione due raccolte di testi su CDROM realizzate negli anni 90 del secolo scorso, note come TLG (Thesaurus Linguae Graecae) e PHI (Packard Humanities Institute) CDROM. Mentre il primo raccoglie tutti i testi in lingua greca antica dalle origini all'epoca bizantina, esaurendo così l'orizzonte temporale di quella che viene chiamata "letteratura greca" sia nella componente classica sia in quella bizantina, il PHI CDROM raccoglie testi in lingua latina dalle origini all'epoca classica con ciò intendendo testi che si collocano intorno al I secolo dopo Cristo. Rimane scoperto quindi il periodo che va da dal I/II secolo dopo Cristo fino alla caduta dell'impero romano d'Occidente nel 456 d.C. cioè il periodo del latino tardo (o tardoantico). Di qui l'idea di colmare la lacuna sia per amore di completezza sia perché i testi latini tardi sono quelli che hanno trasmesso il pensiero e la civiltà latina al Medioevo[3].

Poiché le opere latine tarde sono tutte disponibili in edizioni a stampa, oppure in edizioni online costose, oppure in edizioni online di non eccelsa qualità e/o sparse e/o dallo statuto formale non sempre chiaro, creare una biblioteca digitale specialistica che le accogliesse significava digitalizzare il testo stabilito delle edizioni a stampa, con l'intento di costruire una risorsa di alto livello non solo per i latinisti ma anche più in generale per tutti coloro

[3] Idea di Raffaella Tabacco, latinista del Dipartimento di Studi Umanistici dell'Università del Piemonte Orientale.

che studiano il mondo antico. Prima questione da affrontare, la necessità di fondi mirati alla realizzazione della biblioteca. Nel 2008 in risposta ad un Bando della Regione Piemonte che finanziava iniziative nell'ambito delle scienze umane e sociali[4] venne dunque concepito e redatto da chi scrive un progetto che fu presentato, e dopo aver attraversato un processo di *blind peer evaluation* nell'agosto 2009 fu approvato e cofinanziato. La biblioteca attualmente è in corso di realizzazione presso il Dipartimento di Studi Umanistici dell'Università del Piemonte Orientale ed è disponibile all'indirizzo http://www.digiliblt.unipmn.it. Il progetto non nasceva da bibliotecari ma da studiosi interessati ad uno specifico ambito disciplinare i quali poi avevano coinvolto un bibliotecario nella progettazione[5].

Molti aspetti si potrebbero qui ricordare ma vale la pena soffermarsi su alcuni in particolare in relazione allo spazio disponibile. In primo luogo il tema della *openness*: il sito della biblioteca è costruito interamente con software *open source*[6] e tutto il suo contenuto è disponibile in *open access*. Il tema immediatamente sotteso a queste scelte è che ciò che viene realizzato con fondi pubblici deve essere liberamente accessibile al pubblico che tramite il pagamento delle tasse ne ha resa possibile la realizzazione; ma nel contempo deve anche costare il meno possibile – senza detrimento per la qualità sia specifica delle varie componenti, sia complessiva – così da massimizzarne l'efficacia; e deve per quanto possibile durare nel tempo con costi vivi[7] ridotti al minimo. Ma è almeno altrettanto rilevante, se non di più, il fatto che una biblioteca basata su un sistema aperto è più facilmente interoperabile (cioè predisposta per poter scambiare dati con altre biblioteche): i formati dei dati sono aperti e pubblici anch'essi[8], le librerie di software necessarie per l'interazione sono pubbliche, i programmi stessi sono configurabili e modificabili secondo le necessità dell'utilizzatore. Naturalmente ci si potrebbe chiedere

[4] http://www.regione.piemonte.it/innovazione/ricerca/bandi-e-finanziamenti /bandi-aperti/bando-scienze-umane-e-sociali.html

[5] La direttrice della biblioteca del Dipartimento, Silvia Botto.

[6] Il sistema operativo è Ubuntu Server, il server web Apache, il linguaggio di programmazione PHP, il library manager è XTF, il database MySQL.

[7] Quali lo spazio sul server, la corrente elettrica di alimentazione, un servizio di backup, e simili, che consentono l'esistenza del servizio offerto. Non costi di licenze di software, ad esempio.

[8] È noto per concetto e per esperienza che formati proprietari per i dati, derivanti dall'utilizzo di un software commerciale, possono rendere impossibile l'interazione profonda con altre biblioteche.

perché mai si debba ritenere importante l'interoperabilità. La ragione è che in generale i dati – e il contenuto di una biblioteca digitale è costituito da dati – sono tanto più interessanti quanto più sono numerosi e vari; e per aumentarne la numerosità e varietà il modo più semplice e valido è quello di riunire dati provenienti da differenti origini. Nello specifico: 'riunire' libri digitali provenienti da differenti biblioteche (che cosa esattamente significhi 'riunire' dipende di volta in volta dalle intenzioni e dalle modalità operative scelte[9]) le quali devono essere interoperabili perché questa unificazione dei dati possa avvenire.

Poi c'è il tema della qualità dei testi e del loro trattamento digitale. È interessante un'osservazione formulata in ambito storico ma ugualmente valida anche in ambito letterario da Tim Hitchcock, condirettore di Old Bailey Online, che ha scritto:

> This discussion piece argues that the design and structure of online historical resources and the process of search and discover embodied within them create a series of substantial problems for historians.
> Algorithm-driven discovery and misleading forms of search, poor OCR, and all the selection biases of a new edition of the Western print archive have changed how we research the past, and the underlying character of the object of study. (Hitchcock 2013).

L'aspetto importante qui è la notazione "poor OCR", Hitchcock afferma che quando si digitalizzano fonti testuali a stampa, la scarsa qualità del riconoscimento ottico del testo crea una serie di problemi sostanziali agli studiosi. Consapevoli di questo in digilibLT ogni testo digitalizzato viene (ri)letto almeno tre volte: le prime due sono vere e proprie correzioni di bozze[10], la terza è la fase di annotazione formale del testo.

Un terzo aspetto rilevante è quello della licenza d'uso dei contenuti. Distribuire i contenuti in una prospettiva di *open access* non significa abbandonarli nel web come un tronco portato dal mare su una spiaggia; signi-

[9] Due principalmente sono le opzioni: costruire un'interfaccia di consultazione e ricerca comune a differenti biblioteche e quando l'utente formula una ricerca ridistribuirla alle funzioni di ricerca delle varie biblioteche aderenti; oppure raccogliere centralmente i libri digitali delle varie biblioteche, e lì effettuare le ricerche richieste dall'utente; in entrambi i casi quando l'utente clicca su un esito viene indirizzato al sito della biblioteca pertinente.

[10] Accade talora che in questa rilettura accurata vengano individuati veri e propri errori tipografici presenti nell'edizione a stampa.

fica invece assegnare ad essi in modo esplicito e formale una licenza d'uso che indichi in quali modi se ne consente la ridistribuzione e circolazione in forme diverse da quelle del diritto d'autore tradizionale. C'è oggi una spinta crescente da parte per esempio del programma *Horizon 2020* e della *Budapest Open Access Initiative 2012* (giusto per citare due soggetti importanti in questo ambito) verso l'adozione di licenze Creative Commons molto aperte, basate sulle clausole BY (attribuzione) e SA (condividi allo stesso modo) orientate a permettere il riuso facile e privo di vincoli economici dei contenuti così distribuiti, su cui la discussione è aperta[11].

Ultimo, tra i tanti spunti possibili da approfondire, la biblioteca si sta espandendo oltre i confini del latino letterario tardoantico: da un lato con l'acquisizione degli scritti noti come "grammatici latini": un insieme di opere di argomento linguistico/grammaticale di autori che si collocano tra il II e il V/VI sec. d.C.; dall'altro con l'acquisizione delle opere già raccolte nel PHI CDROM perché su tale collezione di testi i diritti di proprietà intellettuale sono scaduti e dunque essa può essere riutilizzata. L'idea guida è che l'esistenza di molteplici iniziative che mirano a costruire una biblioteca digitale globale del latino[12] è positiva in quanto ciascuna di essa potrà offrire ai lettori edizioni differenti delle medesime opere. E l'interoperabilità, in questo caso, significherebbe che lo studioso può cercare l'opera di un autore e vedere in quali edizioni essa è disponibile nelle varie biblioteche digitali.

3. Il progetto geolat

Avendo a disposizione una biblioteca digitale di testi latini si può pensare a modalità di studio e lavoro sui testi prima non praticabili. Il progetto

[11] Non è questa la sede, e non c'è lo spazio, per discutere approfonditamente di queste licenze che sono descritte nel sito Creative Commons all'indirizzo http://www.creativecommons.it/. Sulla complessa questione dei maggiori o minori vincoli all'utilizzo a fini di lucro dei contenuti distribuiti con licenze Creative Commons, si veda (Lana 2014).

[12] Si possono citare almeno l'iniziativa dell'American Philological Association – che non ha ancora dato luogo ad una biblioteca digitale esistente; e quella dell'Open Philology Project di Gregory Crane Humboldt Professor a Lipsia, http://www.dh.uni-leipzig.de/wo/projects/. Senza dimenticare, per la sua ampiezza, la biblioteca disponibile in Perseus, http://www.perseus.tufts.edu/.

geolat (http://www.geolat.it)[13] lavora per costruire nuovi modi di accesso ai testi latini di digilibLT valorizzandone i contenuti geografici: per mezzo di un'ontologia geografica del mondo antico appositamente costruita, sarà possibile annotare formalmente i nomi geografici presenti nei testi utilizzando lo standard TEI e pubblicare le annotazioni in forma di Linked Open Data. Sulla base di tali annotazioni verranno prodotte carte geografiche arricchite dall'accesso ai testi della biblioteca: tracciando un'area sulla carta si otterrà l'elenco degli autori, opere, e passi, in cui sono presenti nomi geografici appartenenti a quell'area; scelto un luogo sarà possibile individuare sia in forma testuale sia in forma cartografica i luoghi che cooccorrono con il primo; oppure, scelto un luogo leggere i passi dei testi che lo menzionano, e così via.

Anche geolat si basa sui medesimi principi esposti per la biblioteca digilibLT, cioè *open access*, licenze Creative Commons, software open source; e l'utilizzo del meccanismo dei Linked Open Data permetterà l'interoperabilità con altre risorse riguardanti il mondo antico, prima fra tutte il gazzettiere Pleiades[14]: da Pleiades geolat prenderà per ogni nome di luogo le coordinate geografiche e la traduzione del nome latino nelle lingue contemporanee in cui esso è disponibile; mentre Pleiades potrà prendere da geolat per ogni nome di luogo l'elenco dei passi in cui esso ricorre.

In questo modo due differenti risorse si arricchiscono reciprocamente e ciò è reso possibile dalla disponibilità di una biblioteca digitale.

4. Bibliografia

Talbert R., Bagnall R., a c. di (2000). *Barrington Atlas of Greek and Roman World*, Princeton University Press.

Lana M. (2014). *Licenze d'uso e valorizzazione della ricerca umanistica*. «DigItalia», no 1. URL=http://digitalia.sbn.it/ [in stampa].

Hitchcock T. (2013). *Confronting the Digital: or How Academic History Writing Lost the Plot*. «Cultural and Social History», vol. 10, no 1, pp. 9-23.

[13] Finanziato dal 2012 al 2015 dalla Compagnia di San Paolo a seguito di una blind peer evaluation effettuata da European Science Foundation.

[14] http://pleiades.stoa.org/. La fonte di autorità di Pleiades è (Talbert 2000).

Some remarks about Museo Galileo's digital collections*

Stefano Casati, Fabrizio Butini
Museo Galileo, Florence, Italy
{s.casati, f.butini}@museogalileo.it

Abstract. The Museo Galileo's Digital Library (MGDL) began in 2004, for offering an online consultation service of rare and important historical scientific works to scholars, by turning specialized bibliographies, such as the International Galilaean Bibliography, into dedicated digital libraries. The analysis of specific needs of Museo Galileo soon brought about a change of the MGDL's primary goal and highlighted the demand for a much more ambitious project. This new phase was aimed at creating an information system suitable for collecting and improving the different typologies of digital resources – images, texts, videos, sounds, animations, etc. This created the need to for a better managing system of digital resources, their work flow and their conservation. Galileo//thek@ – the ambitious project aimed at collecting all Galilaean resources – is placed inside of this development of the MGDL.

Keyword: Digital Library, History of science, Galilean studies.

1. Museo Galileo's Digital Collections

Digital libraries are usually viewed as effective tools to consult books and manuscripts online, however, their true nature in reality is much more complex, structured and articulated. The restrictive view point partially comes from the coupling of the contemporary term "digital" with the well-established term of "library" which carries with it the centuries-old traditional idea. This kind of cultural uneasiness almost naturally produces the idea that the digital world is essentially a technological branch of the universe of libraries. The experience of Museo Galileo's Digital Library (MGDL)[1] seems to refute this point of view.

The MGDL began in 2004 within a cultural context which marked its characteristic features and future development. Museo Galileo is comprised

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.
[1] Info about Museo Galileo: http://www.museogalileo.it/

of various departments which produce digital data and resources that are then utilized by the Digital Library. These include the Library that produces different kinds of bibliographies, the Historical Archive which keeps important 19th-century collections, the Institute which is devoted to research and the dissemination of scientific culture, the Photographic Laboratory, and the Multimedia Department which develops new information technologies and manages the museum's website.

Initially the project aimed at offering an online consultation service of rare and important historical scientific works to scholars, by turning specialized bibliographies, such as the *International Galilaean Bibliography*,[2] into dedicated digital libraries. It was a demanding job – especially with regard to the difficulty in finding many of these referenced works – which resulted in the creation of a digital collection of about 5,000 works, tallying about two million pages. A great part of them were collected in the framework of various projects sponsored by the Italian Digital Library.[3]

The analysis of specific needs of Museo Galileo's various departments soon brought about a change of the MGDL's primary goal and highlighted the demand for a much more ambitious project. This new phase was aimed at creating an information system suitable for collecting and improving the different typologies of digital resources – images, texts, videos, sounds, animations, etc. This created the need to for a better managing system of digital resources, their work flow and their conservation.

As a result of this change, the DL is now managing all digital resources produced by Museo Galileo. The birth of MGDL can be dated back to the publication of *Bibliotheca Perspectivae: arte e scienza della rappresentazione. I trattati di prospettiva del Rinascimento* in 2005,[4] when the first DL (mostly

[2] Info about MGDL: http://www.museogalileo.it/esplora/biblioteche/biblioteca.html

[3] Many of the MGDL contents come from other libraries and have been included thanks to the partnership with important institutions, such as the Accademia delle Scienze di Torino, the Accademia Nazionale dei Lincei, the Biblioteca di Storia delle Scienze Carlo Viganò in Brescia and the Biblioteca d'Informazione e Cultura in Milan. Internet access is totally free. Downloading is also available, except for copyrighted works, which can be consulted only inside the (Museo Galileo) Library. See info about the Italian Digital Library at http://www.librari.beniculturali.it

[4] *Bibliotheca Perspectivae: art and science of representation. Renaissance perspective treatises* well illustrates this new concept. Firstly organised as an "on-line library" of rare and important treatises on the history of perspective, *Bibliotheca Perspectivae* was then turned into an informational tool supplemented with iconographic and multimedia resources, which made it possible to map out knowledge routes to be explored on different levels which are suit-

devoted to reproduce analogue documents) was turned into an information system included records coming from different archives and heterogeneous digital collections.

The MGDL cumulative archive includes both records originating from the iconographic archives and from bibliographical archives, such as the catalogue and bibliographies produced by the Library. It also includes data related to the cataloguing of scientific instruments. The integration between data populating the cumulative archive and digital resources metadata allowed us to create knowledge routes which are suitable to the web modalities and potentials, thus highlighting one of most innovative aspects of digital libraries.[5] This made inductive navigation systems possible. For example this is the case of the online publication of *Saggi di naturali esperienze*, the book describing experiments conducted by members of the *Accademia del Cimento*[6] which starts with the original work as a basis. Starting from the digital edition in both image and full-text formats of the well-known book, we created a reading route consisting of a "spider web" made up of cross-references and teaching aids which enable users to go further in-depth in related topics as well as gather information about the various editions, translations and editor's notes. Moreover, the digital index makes it is possible to analytically explore the different entries through multimedia and other information resources.

2. Galileo//thek@

Galileo//thek@ – the ambitious project aimed at collecting all Galilaean resources– is placed inside of this digital context. The *Galileo//thek*@

able for both scholars and curious users. Visit http://brunelleschi.imss.fi.it/bdtema/ibpr.asp?c=684&xsl=5.

[5] MGDL now includes 13 digital collections which are thematic historical scientific groups, mostly dedicated to Galilaean corpus (visit http://www.museogalileo.it/esplora/biblioteche/bibliotecadigitale.html). MGDL also includes virtual exhibitions (visit http://www.museogalileo.it/esplora/mostre/mostrevirtuali.html) offering well organized and complex browsing routes as well as digital collections. The distinction between digital collections and virtual exhibitions is strictly related to the architecture of Museo Galileo's website: the so-called virtual exhibitions are to be considered as digital libraries for all intents and purposes.

[6] *Saggi di naturali esperienze fatte nell'Accademia del Cimento sotto la protezione del serenissimo principe Leopoldo di Toscana e descritte dal segretario di essa Accademia*, 1666 [i.e. 1667].

project began in the early 2000s. Its second edition, which has completely new graphics and contents, is projected to be released by the end of 2014. *Galileo//thek*@ is an innovative, integrated federation of Galilaean digital resources, which can be easily explored through refined research tools. The new edition supplements and updates previously released resources, improves research capacities and the interface, making available tools interactive with online contents for users. It will be then be possible to make one, simple search into browse data related to manuscripts, bibliographical records, thematic and biographical indexes, and lexicon of every single record referenced by the National Edition of Galileo's Works.

The *Galileo//thek*@ model called for new ways to collaborate amongst different professionals. Computer scientists, system analysts and librarians worked side-by-side with scholars and researchers to create a digital library of this kind of complexity. Since this project called for work phases including the sorting and digital acquisition of documents, data and metadata processing, the creation of multimedia systems as well as the adoption of a sound infrastructure for web storage and releasing its whole operations demanded input from a number of professional competences.

In order for users to benefit from all of the resources in an effective way, MGDL designed a technological infrastructure able to support all of the phases necessary for its creation. This infra-structure has been conceived by and developed according to the following guidelines:

- Reliability: The system must always be in operation and the inactivity period in case of default must be as short as possible.
- Sustainability: The costs of the infra-structure, as far as both human and material resources are concerned, are to be sustainable and consistent within budget needs while providing a high quality of service.
- Usability: All of the elements (not only those made available to final users) must be accessible and used in an effective way by people involved in the production/use of digital objects.
- Open Source: All software components must be developed through Open Source technologies.

The following diagram explains the architecture of DL.

– Digital TECA: web-application designed to use digital works. This application allows users to:

- Explore a work through the structural map;
- View texts and images in a single- or double-page view, in normal or zoom view, rotated by 90, 180, 270 degrees;
- Free search on texts (where OCR is available);
- Download the work in PDF format (if available);
- View and click external links, if any;
- Link one to the context in a sensible way through other web-applications (OPAC, *Galileo//thek*@, etc).

– DL Managing System: web-application devoted to Intranet use and designed to manage all the phases of processing, from image acquisition to their publication in Digital TECA;

- TECA Database: database used by Digital TECA;
- Database Managing System: database used by DL managing system.

Storage

Storage is the most critical component with regards to reliability and sustainability. In substance, there are two kinds of digital storage:

1. Long Term Storage (LTS): Oriented to preserve the originals from whom digital works are generated through time;
2. Work Storage: Oriented to processing and using the digital works.

LTS is the most complex part of the system, since it guarantees the conservation of the originals through time by using the best-quality-format scans of the MG's originals.

While other commercial solutions of strong reliability were available their installation costs and hardware/software updating were not within MG's budget. We therefore, chose the following solutions:

- We acquired two independent but identical machines, both equipped with a battery of SATA discs of moderate cost.
- These machines are synchronized through a scheduled rsync process in order to reduce displacements and make them last no longer than a few hours. These dis-alignments can easily be corrected through manual interventions.
- We decided not to use systems such as RAID and/or iSCSI since they depend too much on hardware (disk controller, ethernet cards) and are complicated when managing some kinds of faults.

- The whole architecture was chosen due to its contained costs which also includes implementing all hardware updating.

Beyond LTS there are two more storage systems, based on RAID 1 systems and containing the images in the formats available on the Internet.

1. TECA storage which contains images and texts used by Digital TECA. This system is subject to damages caused by faults and hacker attacks, etc. in spite of Firewall protection.
2. Work storage which contains a superset of the images/texts in the Digital TECA. This system is used both to backup TECA and for publishing purposes.

Development and Operative Environment

- All software products used are Open Source or developed by MG.
- Main language is JAVA, JSF 2/RichFaces and ORM Hibernate frameworks.
- PDFs are created through ImageMagick libraries.
- Text search is based on Apache SOLR.
- All procedures use Apache Tomcat as application server and Apache HTTPD as front-end toward Internet users.

3. References

Casati S. (2012). *The Digital library. in Displaying scientific instruments: from the Medici wardrobe to the Museo Galileo*, edited by Filippo Camerota (numero monografico, Annali del Laboratorio museotecnico), Goppion, c1997, pp. 341-347.

Casati S. (2005). *La Biblioteca digitale dell'Istituto e Museo di storia della scienza di Firenze: il modello Bibliotheca perspectivae, arte e scienza della rappresentazione*. Seminario tenuto per la XV Settimana della cultura scientifica, Parma, 18 marzo 2005. [S.l. : s.n., 2005], Casalini Libri.

Saggi di naturali esperienze fatte nell'Accademia del Cimento sotto la protezione del serenissimo principe Leopoldo di Toscana e descritte dal segretario di essa Accademia. In Firenze, per Giuseppe Cocchini ..., 1666 [i.e. 1667].

Papers

Digital Philology / Filologia digitale

L'*Open Philology Project* dell'Università di Lipsia. Per una filologia 'sostenibile' in un mondo globale*

Monica Berti[1], Greta Franzini[1], Emily Franzini[1],
Giuseppe G.A. Celano[1], Gregory R. Crane[1,2]
[1] Humboldt Chair of Digital Humanities / Universität Leipzig, Leipzig, Germany
{berti, franzini, efranzini, celano, crane}@informatik.uni-leipzig.de
[2] Perseus Project / Tufts University, Medford, MA, USA
gregory.crane@tufts.edu

Abstract: Argomento di questo articolo è la presentazione dell'*Open Philology Project* della Humboldt Chair in Digital Humanities dell'Università di Lipsia. Il progetto nasce nell'ambito delle attività del Perseus Project della Tufts University e ha come scopo primario lo sviluppo di una collezione di risorse linguistiche greche e latine leggibili dalla macchina, la creazione di manuali dinamici basati su *corpora* annotati e l'avvio di nuove forme di pubblicazione riguardanti le lingue classiche, che possono includere sia annotazioni individuali che edizioni tradizionali integrate con dati elaborabili dalla macchina. L'*Open Philology Project* include tre componenti principali costituite dall'*Open Greek and Latin*, dall'*Historical Languages e-Learning Project*, e dall'*Open Access Publishing*.

Parole chiave: big data, OCR, e-Learning, greco, latino, didattica, publishing, business, treebanking, annotazione linguistica, riusi testuali.

1. Introduzione

L'*Open Philology Project* (OPP) della Humboldt Chair in Digital Humanities dell'Università di Lipsia aspira a riaffermare il ruolo e il valore della filologia nel senso più ampio del termine[1]. Due secoli fa, nella sua fondamentale opera di ripensamento degli studi classici, il filologo tedesco August Böckh definiva la filologia come *universae antiquitatis cognitio historica et philosophica* (Böckh 1858, 105; Id. 1877, 12). Prendendo spunto da questa affermazione, s'intende recuperare il significato originario della parola greca *philologia* (φιλολογία), la quale denota lo studio più vasto ed esaustivo

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.
[1] L'indirizzo del progetto è http://www.dh.uni-leipzig.de/wo/projects/

possibile delle testimonianze linguistiche al fine di promuovere una conoscenza approfondita dell'attività intellettuale prodotta dall'uomo. Nel caso specifico, l'OPP mira a concentrare l'attenzione sul greco e sul latino per quattro diversi motivi: 1) sono già disponibili in rete collezioni e strumenti dedicati a queste lingue; 2) esistono comunità di utenti particolarmente numerose (circa 35.000 utenti al mese accedono alle collezioni di fonti greche e latine della Perseus Digital Library[2]); 3) il progetto ha sede in Europa, il cui patrimonio culturale costituisce un bacino naturale per la creazione, lo sviluppo e la distribuzione di materiali pertinenti all'antichità greco-latina; 4) la città di Lipsia vanta una tradizione editoriale e libraria di prim'ordine nel campo della filologia classica – basti pensare alle edizioni critiche di testi greci e latini pubblicate dalla casa editrice Teubner[3] – e si pone dunque come spazio privilegiato per la ridefinizione della filologia nell'ambito degli studi di informatica umanistica.

L'OPP è stato concepito con la speranza di creare un modello applicabile anche allo studio di altre lingue storiche. Più in particolare, esso persegue tre obiettivi diversi ma strettamente connessi fra loro: 1) la creazione di una collezione di risorse linguistiche leggibili dalla macchina, le quali siano aperte, estensibili e riutilizzabili; 2) lo sviluppo di manuali dinamici basati su *corpora* annotati, che permettano di personalizzare il vocabolario e la grammatica dei testi esistenti e coinvolgere gli studiosi e gli studenti a produrre nuove annotazioni in maniera collaborativa; 3) la promozione di nuove forme di pubblicazione, che possono consistere sia in annotazioni individuali argomentate che in edizioni tradizionali integrate con dati elaborabili dalla macchina. Questi obiettivi sono definiti attraverso le tre componenti dell'OPP presentate qui di seguito: 1) *Open Greek and Latin Project*; 2) *Historical Languages e-Learning Project*; 3) *Open Access Publishing*.

2. Open Greek and Latin Project

L'*Open Greek and Latin Project* (OGL) si sta attualmente dedicando alla raccolta e alla scannerizzazione di edizioni di testi classici al fine di realizza-

[2] Il progetto OPP nasce nell'ambito delle attività del Perseus Project presso la Tufts University http://www.perseus.tufts.edu/

[3] Sulla *Bibliotheca scriptorum Graecorum et Romanorum Teubneriana* si veda la pagina della casa editrice De Gruyter http://www.degruyter.com

re la più grande biblioteca digitale in materia, contribuendo nel contempo all'arricchimento della collezione greca e latina di Google Books. In questo ambito l'OGL riveste anche un ruolo di consulenza sulla legge europea sul diritto d'autore, dato che redige una lista di edizioni europee che Google Books può digitalizzare, offrendo dunque una tutela contro eventuali cause legali[4].

Tale raccolta, che è *open source* e *open access*, fornisce anzitutto immagini ricercabili di edizioni di testi classici libere dai vincoli del *copyright*, corredandole di traduzioni multilingue e codificandole secondo lo standard TEI XML (*subset* EpiDoc[5]). L'architettura dell'OGL è concepita per gestire e mettere a disposizione degli utenti edizioni e traduzioni diverse per ogni opera classica prodotta dall'antichità greco-latina, coprendo un arco di tempo che va dall'epoca arcaica al 600 d.C. Questa caratteristica distingue l'OGL dalla maggior parte dei *corpora* esistenti (i quali prevedono generalmente un'unica edizione per opera) e costituisce un presupposto imprescindibile sul quale fondare edizioni digitali che siano realmente critiche e multitestuali (sul concetto di 'multitesto' si veda Blackwell-Crane 2009). Per poter realizzare questo obiettivo, l'OGL ha avviato collaborazioni con istituzioni accademiche di altri paesi al fine di promuovere lo scambio di dati con progetti di respiro internazionale. Tra i paesi coinvolti si annoverano la Bulgaria (progetti *Romulus Bulgaricus* e *Theseus*[6]), la Croazia (Università di Zagabria, Dipartimento di Filologia Classica, progetto *Croala*[7]), la Georgia (Ivane Javakhishvili Tbilisi State University[8]), il Nebraska (progetto *Digital Athenaeus*[9]) e l'Italia (Università del Piemonte Orientale, progetto *digilibLT*[10]). Tale iniziativa vorrebbe naturalmente estendersi ad altri paesi europei, sperando di spostarsi anche su zone meno esplorate come l'Est Europeo e il Medio Oriente.

Lo sforzo intrapreso dall'OGL comporta un lavoro di inserimento di dati e l'uso di tecnologie OCR per arricchire un *corpus* potenzialmente già esistente, che sia aperto e sufficientemente ampio da includere i circa 100.000.000

[4] Per quanto riguarda le leggi sul *copyright* vigenti in diversi paesi si può consultare la voce Wikipedia http://en.wikipedia.org/wiki/List_of_countries%27_copyright_lengths
[5] http://sourceforge.net/p/epidoc/wiki/Home/
[6] Si vedano rispettivamente http://romulus-bg.net e http://theseus.proclassics.org
[7] http://www.tei-c.org/Activities/Projects/cr02.xml
[8] http://www.tsu.edu.ge/en/
[9] http://www.dh.uni-leipzig.de/wo/open-philology-project/digital-athenaeus/
[10] http://digiliblt.lett.unipmn.it

di parole prodotte dai primordi della classicità sino al VII secolo d.C. A questo riguardo OGL ha firmato contratti con due aziende in grado di produrre tale mole di lavoro. Il primo contratto è stato firmato con la compagnia francese Jouve, la quale si de-dica alla digitalizzazione e, dove necessario, all'inserimento manuale dei dati del *Corpus Scriptorum Ecclesiasticorum Latinorum* (CSEL) e dei primi cinquanta volumi della *Patrologia Latina*[11]. Inoltre, alla luce della collaborazione italo-tedesca, Jouve si occuperà della digitalizzazione di volumi destinati all'arricchimento di *digilibLT*, la biblioteca digitale dei testi latini tardoantichi dell'Università del Piemonte Orientale di Vercelli[12]. Il secondo contratto è stato firmato con Digital Divide Data (DDD), un'azienda americana con filiali in Laos e Cambogia[13]. DDD si occupa di digitalizzare i volumi 51-122 della *Patrologia Latina* e di altre opere greche, tra le quali quelle di Ateneo, Filone Alessandrino, Libanio, i commenti greci ad Aristotele (*Commentaria in Aristotelem Graeca*) e, in un prossimo futuro, la *Patrologia Graeca*[14]. L'intento, infatti, è quello di testare entrambi i *workflows* e, qualora portassero a buoni risultati, rinnovare i contratti per produrre edizioni elettroniche di Eschilo, della raccolta dei frammenti degli storici romani (*Historicorum Romanorum Reliquiae*) e di qualsiasi altro autore fosse richiesto dai collaboratori dell'OGL.

Il *workflow* di queste attività prevede che il gruppo di ricerca dell'Università di Lipsia gestisca l'*input* e verifichi la validità del prodotto finale, mentre le compagnie con le quali sono stati stipulati i contratti si occupano della parte tecnica e meccanica del progetto. L'organizzazione del lavoro può essere riassunta nel modo seguente:

Università di Lipsia – Ogni autore o volume o serie di volumi deve essere codificata secondo la struttura dell'edizione di riferimento. Questa necessità comporta la creazione di *templates* molteplici che riflettono la diversità delle edizioni, pur rimanendo sempre compatibili con le specifiche di EpiDoc e

[11] Per digitalizzazione si intende il riconoscimento ottico dei caratteri (OCR), la correzione dell'*output* dell'OCR, nonché la codifica in EpiDoc XML. Per informazioni sul gruppo Jouve si veda http://www.jouve.com/

[12] Vd. n. 10.

[13] http://www.digitaldividedata.org

[14] La decisione di dividere i volumi della *Patrologia Latina* fra Jouve e DDD è stata dettata dal desiderio di paragonare due differenti *workflows* e *outputs*. I risultati prodotti permetteranno di scegliere il procedimento migliore in termini di metodo/qualità/prezzo.

in particolare con la classe di marcatori (*tags*) CITE-friendly[15]. Il gruppo di lavoro dell'Università di Lipsia si occupa di analizzare la struttura di ogni edizione e di ricavarne un documento descrittivo con *template* allegato da inoltrare alle aziende ingaggiate per il lavoro. Esso, inoltre, si occupa anche di scaricare e fornire le scansioni esistenti delle suddette edizioni in formato TIFF, PNG o JP2[16]. Queste immagini vengono correttamente catalogate e caricate su un *server* che contribuisce alla creazione del *corpus Open Greek and Latin*. Per quanto concerne le collaborazioni, è responsabilità dell'ente collaboratore fornire al gruppo di lavoro dell'Università di Lipsia le immagini necessarie. Sebbene la correzione degli errori prodotti dall'OCR venga effettuata dalle due aziende Jouve e DDD, lo strumento che queste ultime utilizzano per svolgere tale compito è stato sviluppato dall'Università di Lipsia. Nello specifico, il *Proofreader* (cfr. fig. 1) ottimizza uno strumento sviluppato da Bruce Robertson e Federico Boschetti e permette di allineare l'*output* dell'OCR a edizioni conosciute e consentire correzioni semi-automatiche tramite un'interfaccia semplice e intuitiva (Boschetti *et al*. 2009; cfr. inoltre Manmatha-Feng 2006 e Bryant *et al*. 2010)[17]. Il gruppo di lavoro di Lipsia si occupa infine di supervisionare il *workflow* e assicurarsi che i termini e le scadenze previste siano rispettati.

Jouve e DDD – Come si è detto, le due aziende si occupano di 'OCRizzare' le immagini fornite dall'Università di Lipsia, di correggere eventuali errori utilizzando lo strumento loro fornito e di codificare il testo secondo le specifiche EpiDoc che sono state predisposte. Eventuali commenti, problemi e richieste vengono gestite tramite posta elettronica e videoconferenze a scadenza regolare.

[15] Sulla CTS/CITE Architecture sviluppata dall'*Homer Multitext Project* per la codifica dei manoscritti omerici si veda http://www.homermultitext.org/hmt-doc/cite/

[16] Le biblioteche digitali di riferimento sono Archive.org (https://archive.org/details/texts), HathiTrust Digital Library (http://www.hathitrust.org/) e Deutsche Digitale Bibliothek (https://www.deutsche-digitale-bibliothek.de/). Le scansioni vengono scaricate e convertite nel formato richiesto da Jouve e DDD in maniera semi-automatica con strumenti *ad hoc* sviluppati dal gruppo di lavoro dell'Università di Lipsia. Ogni scansione necessita anche una corretta catalogazione in quanto le biblioteche digitali di riferimento spesso presentano *metadata* errati. Un esempio è la *Patrologia Graeca* sotto la quale risultano essere stati erroneamente catalogati molti volumi, che richiedono pertanto un ulteriore controllo manuale.

[17] Il Proofreader è stato sviluppato e ottimizzato da Frederik Baumgardt (Università di Lipsia), Bruce Robertson (Mount Allison University) e Federico Boschetti (CNR Pisa). Su questi strumenti si vedano https://github.com/CoPhi e http://heml.mta.ca/rigaudon

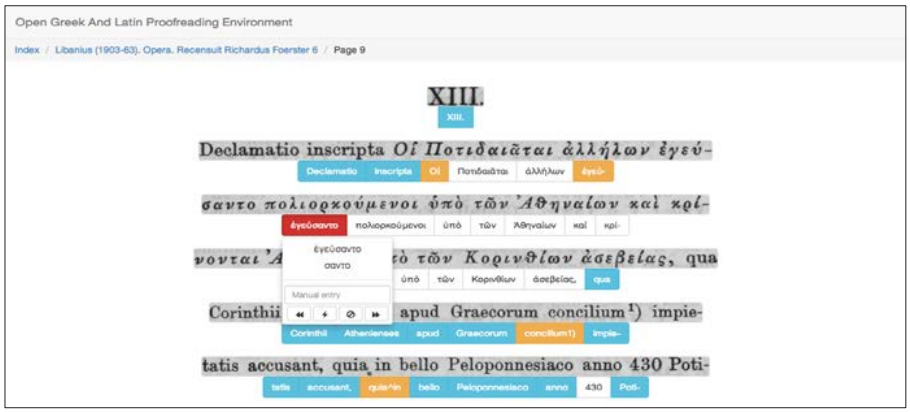


Fig. 1. *Proofreader*: strumento di correzione dell'*output* OCR.

Un terzo contratto, ancora in fase di definizione, includerà la scansione di edizioni pubblicate tra il 1922 e il 1985 e che sono dunque ancora soggette ai vincoli del *copyright*. Parte del *workflow* della biblioteca responsabile (tedesca per questioni di logistica) comporterà la rimozione degli apparati critici e delle note, fornendo all'Università di Lipsia solo il testo latino o greco dell'opera curata dall'editore. Tali scansioni verranno aggiunte al sistema digitale bibliotecario tedesco per permettere a terzi di usufruirne. Qualora la biblioteca in questione non disponesse dei libri necessari, sarà cura della biblioteca dell'Università di Lipsia fornirgliene copia[18].

Un progetto che verrà avviato in futuro vedrà anche la partecipazione degli utenti al lavoro di digitalizzazione, come sta avvenendo ora mediante il coinvolgimento degli studenti dei corsi di filologia digitale organizzati presso l'Università di Lipsia e degli studenti Erasmus[19]. Il gruppo di lavoro di Lipsia sta infatti sviluppando un processo computazionale integrato con un sistema di pianificazione e notifica, che fornirà una visione sequenziale del progresso dei lavori dell'OGL e faciliterà i contributi esterni, per esempio da parte di ricercatori e studenti e di tutti coloro che sono interessati all'iniziativa. Vista la natura pubblica e aperta del progetto OGL, il *workflow* dell'OCR è stato progettato con interfacce che permettono agli utenti di partecipare

[18] La biblioteca di Lipsia, per quanto attrezzata in termini di digitalizzazione, non dispone ancora delle risorse umane necessarie per svolgere questo lavoro in tempi brevi.
[19] http://www.dh.uni-leipzig.de/wo/courses/

al lavoro di digitalizzazione. Sviluppato sulla base dell'Oracle Grid Engine, il *workflow* consiste di tre componenti principali: 1) un nucleo (*core*) intercambiabile di uno dei tre motori OCR (Gamera, Tesseract, OCRopus); 2) un livello di ottimizzazione sviluppato da Bruce Robertson e Federico Boschetti; 3) un modulo per allineare l'*output* dell'OCR a edizioni conosciute e consentire correzioni semi-automatiche.

I dati prodotti da questo processo vengono codificati secondo le specifiche EpiDoc, le quali forniscono un tipo di marcatura standardizzata, ma non ristretta, e compatibile con i testi dell'OGL. Tale codifica viene realizzata in parallelo alla conversione in EpiDoc dei *file* della Perseus Digital Library. La possibilità di disporre dei testi della Perseus DL e dell'OGL in formato EpiDoc faciliterà lo scambio e il collegamento dei dati con le collezioni di documenti epigrafici e papirologici che sono già stati codificati in questo modo e con tutte quelle altre banche dati che sono attualmente in fase di conversione, come per esempio EAGLE (*Europeana Network of Ancient Greek and Latin Epigraphy*)[20].

3. Historical Languages e-Learning Project

Un'ulteriore componente dell'OPP è rappresentata dall'*Historical Languages e-Learning Project*, il cui obiettivo è quello di realizzare un sistema per l'apprendimento delle lingue storiche in ambiente digitale. Questo sistema permette di selezionare frasi che abbiano una certa morfosintassi e/o un certo lessico, sulla base degli interessi specifici del discente o del docente che vuole impiegare questo sistema per insegnare le lingue classiche. Il testo selezionato per il *pilot* del progetto è una sezione del primo libro della *Guerra del Peloponneso* di Tucidide nota come *Pentecontaetia* (Thuc. 1.89-118). Lo scopo del *pilot* è quello di insegnare alcuni aspetti della lingua greca sia tramite l'uso della piattaforma *e-Learning* sia tramite un'attiva partecipazione di annotazione al testo. Dal successo del *pilot* dipenderà l'espansione del progetto per includere altri testi.

Per suscitare la curiosità del pubblico interessato e fidelizzare gli utenti, i creatori del *pilot* lavorano anche all'estetica della piattaforma e all'organizzazione del contenuto (cfr. fig. 2). L'intento è che il materiale sia disposto in modo intuitivo, divertente e incoraggiante, e che sia permesso all'utente di

[20] http://www.eagle-network.eu/

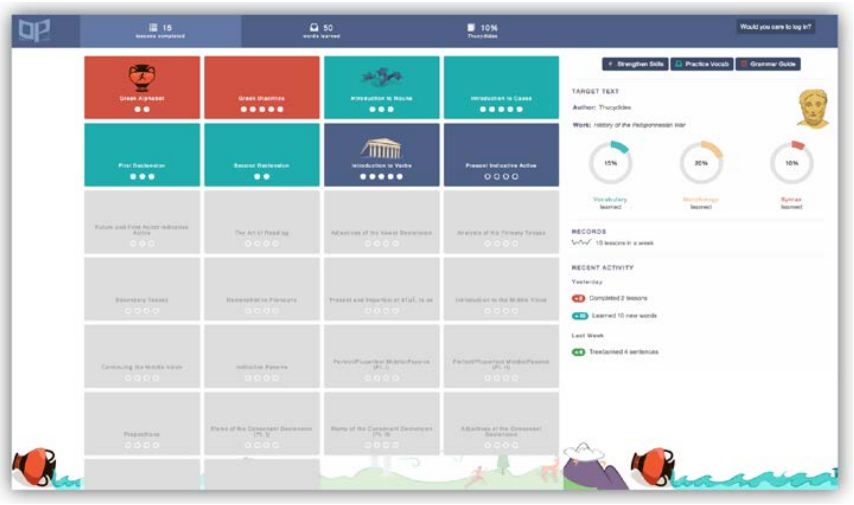


Fig. 2. *Historical Languages e-Learning Project*, Homepage.

imparare e partecipare, qualsiasi sia la sua conoscenza della lingua greca. La piattaforma *e-Learning* offrirà ulteriori vantaggi: la possibilità di iscriversi alla piattaforma (tramite registrazione e *log in*) e quindi la scelta di abbandonare e riprendere l'apprendimento a piacere, la possibilità di visualizzare il proprio percorso di apprendimento e una cronologia delle proprie annotazioni al testo, e infine la possibilità di esercitare la propria conoscenza linguistica tramite esercizi basati direttamente sul testo greco. La piattaforma offrirà inoltre la possibilità di scegliere la lingua moderna così che gli utenti possano imparare il greco con strumenti tradotti nella propria lingua madre (che non sia necessariamente l'inglese).

Motore del progetto è l'annotazione morfosintattica. I testi greci e latini sono annotati semi-automaticamente per la morfologia utilizzando il *tagger* Morpheus sviluppato dal Perseus Project, il quale restituisce un testo con l'analisi morfologica di ogni parola[21]. Nel caso di più analisi possibili, spetta all'annotatore decidere quale sia quella corretta sulla base del contesto. L'annotazione morfologica costituisce la base per l'annotazione sintattica che viene eseguita manualmente. Attraverso l'interfaccia grafica offerta da

[21] http://wiki.digitalclassicist.org/Morpheus

Alpheios[22], l'annotatore costruisce un albero sintattico secondo delle *guidelines* che si ispirano a quelle adottate per la *Prague Dependency Treebank* 2.0[23].

La *Ancient Greek and Latin Dependency Treebank* del Persues Project conta circa 400.000 parole[24]. Al momento è in corso una revisione tesa ad arricchire l'annotazione con l'aggiunta di glosse secondo lo schema delle *Leipzig Glossing Rules*[25], al fine di promuovere un tipo di analisi standard per la morfologia di ogni parola. L'annotazione conterrà inoltre riferimenti alla grammatica greca dello Smyth (1920) per coniugare il sapere della grammatica tradizionale con quello della *Functional Generative Description* della *treebank* di Praga.

4. Open Access Publishing

Uno degli obiettivi principali dell'Opp consiste nella creazione di un nuovo modello di edizioni scientifiche native digitali. Questo obiettivo è attualmente perseguito mediante l'implementazione di Perseids, che è una piattaforma collaborativa della Perseus Dl sviluppata mediante la personalizzazione di risorse *open source* create per annotare fonti classiche codificate secondo lo standard Tei Xml (per una descrizione della piattaforma e di diversi progetti ad essa connessi si veda Almas-Beaulieu 2013)[26]. Perseids è un ambiente condiviso dove gli utenti possono editare, tradurre e commentare diverse tipologie di fonti antiche, comprese le iscrizioni e i manoscritti. L'obiettivo di Perseids è duplice, perché mira sia alla pubblicazione di edizioni scientifiche che allo sviluppo di risorse didattiche per gli studenti dei corsi universitari:

1) Per quanto riguarda la comunità scientifica, uno dei principali modelli di pubblicazione all'interno di Perseids è il *Fragmentary Texts Editor* (Fte), che ha la funzione di produrre annotazioni complesse concernenti opere conservate solo attraverso citazioni e riusi in testi coevi o posteriori (Almas-Berti 2013a; Eadd. 2013b; Almas *et al*. 2013)[27]. A tal fine Perseids

[22] http://alpheios.net/
[23] http://ufal.mff.cuni.cz/pdt2.0/
[24] http://nlp.perseus.tufts.edu/syntax/treebank/index.html
[25] http://www.eva.mpg.de/lingua/resources/glossing-rules.php
[26] Perseids è disponibile al seguente indirizzo ed è liberamente accessibile http://sites.tufts.edu/perseids/
[27] Per una *demo* dell'Fte si veda http://perseids.org/sites/berti_demo/. Il codice sorgente è disponibile al seguente indirizzo https://github.com/PerseusDL/lci-demo

utilizza diversi metodi di *in-line* e *stand-off* markup combinando lo standard TEI XML e la CTS/CITE Architecture con altri *data model*, quali l'Open Annotation Collaboration (OAC), il Systematic Assertion Model (SAM) e il W3C Provenance Model (Almas *et al*. 2013). Parallelamente al *Fragmentary Texts Editor*, la cattedra di informatica umanistica dell'Università di Lipsia sta avviando il *Leipzig Open Fragmentary Texts Series* (LOFTS), il cui obiettivo è la realizzazione di nuove edizioni native digitali di autori frammentari[28]. Il progetto è supportato dal Perseus Project e avrà come sede di pubblicazione il Center for Hellenic Studies[29]. Il primo sforzo nell'ambito di questa iniziativa è la digitalizzazione dei cinque volumi dei *Fragmenta Historicorum Graecorum* pubblicati da Karl Müller tra il 1841 e il 1870 (progetto *Digital Fragmenta Historicorum Graecorum (DFHG)*), i quali costituiscono la prima opera monumentale di raccolta dei frammenti degli storici greci e rappresentano un ottimo punto di partenza per contribuire alla realizzazione di edizioni digitali in materia[30].

2) I risultati che Perseids mira a produrre non riguardano soltanto gli studiosi ma anche gli studenti, i quali hanno l'opportunità di lavorare direttamente sui documenti originali e contribuire ai risultati della comunità scientifica. Questo tipo di attività è svolto in parallelo presso la Tufts University e l'Università di Lipsia, la quale ha avviato una serie di corsi di filologia digitale. Attraverso questi corsi gli studenti apprendono come trattare diverse forme di organizzazione del sapere scientifico sviluppate dalla cultura della stampa, come le edizioni critiche, i lessici, le enciclopedie, i commentari, gli indici e le grammatiche. Gli studenti hanno inoltre l'opportunità di concentrarsi su temi particolarmente complessi, come l'annotazione linguistica delle fonti storiche, la rappresentazione delle fonti frammentarie e dei riusi testuali, o l'allineamento linguistico dei testi.

---

[28] http://www.dh.uni-leipzig.de/wo/open-philology-project/the-leipzig-open-fragmentary-texts-series-lofts/

[29] http://chs.harvard.edu/

[30] Per una descrizione del progetto si veda http://www.dh.uni-leipzig.de/wo/open-philology-project/the-leipzig-open-fragmentary-texts-series-lofts/digital-fragmenta-historicorum-graecorum-dfhg-project/. La pagina contiene un collegamento alle *guidelines* sviluppate dal gruppo di lavoro dell'Università di Lipsia per la codifica dei volumi secondo lo standard EpiDoc e un catalogo degli oltri 600 autori frammentari pubblicati dal Müller nei cinque volumi dei *FHG*. Le linee guida, oltre a fornire uno strumento per tutti coloro che collaborano al progetto, contribuiscono allo sviluppo generale delle *guidelines* di EpiDoc (http://www.stoa.org/epidoc/gl/latest/), mentre il catalogo degli autori contribuisce allo sviluppo e all'arricchimento del Perseus Catalog (http://catalog.perseus.org/).

5. Open Data Revenue Models e Open Philology Publishing

A supporto dell'attività scientifica sopra descritta, l'Opp intende sviluppare un *business plan* per creare strategie che permettano di sostenere economicamente il progetto e renderlo in futuro autonomo da investimenti esterni. Dato che l'Opp è per definizione basato su un modello di accesso libero e gratuito, la parte più complessa consiste nello sviluppare modelli che consentano il sostenimento di una piattaforma di apprendimento aperta e gratuita tramite l'aggiunta di servizi sofisticati a pagamento. Il principio base è quello di creare un'alternativa all'attuale monopolio della produzione del sapere, la cui fruizione è molto costosa per l'utente, favorendo un accesso gratuito, il quale sia però arricchito di servizi addizionali a basso costo per apprendere, analizzare e contribuire ad una massa di dati complessi in costante crescita. Il progetto intende fornire strumenti destinati a studiosi e studenti, oltre che alle scuole e in generale al pubblico interessato. I servizi offerti copriranno diverse aree, dai servizi informatici per l'*e-Learning*, ai libri di testo interattivi, ai sistemi di valutazione e di *ePortfolio*.

6. Bibliografia

Almas B., Beaulieu M.C. (2013). *Developing a New Integrated Editing Platform for Source Documents in Classics.* «Literary and Linguistic Computing», vol. 28, no 4, pp. 493-503. URL=http://llc.oxfordjournals.org/content/28/4/493.abstract. [ultima visita 3.3.2014].

Almas B., Berti M. (2013a). *Perseids Collaborative Platform for Annotating Text Re-Uses of Fragmentary Authors.* In F. Tomasi, F. Vitali, a. c. di, *DH-Case 2013. Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities,* ACM, art. no. 7. URL=http://dl.acm.org/citation.cfm?id=2517986. [ultima visita 3.3.2014].

Almas B., Berti M. (2013b). *The Linked Fragment: TEI and the Encoding of Text Re-uses of Lost Authors.* In F. Ciotti, A. Ciula, a. c. di, *The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013*. Università Roma La Sapienza, pp. 12-16. URL=http://digilab2.let.uniroma1.it/teiconf2013/wp-content/uploads/2013/09/book-abstracts.pdf. [ultima visita 3.3.2014].

Almas B., Berti M., Choudhury S., Dubin D., Senseney M., Wickett K.M. (2013). *Representing Humanities Research Data Using Complementary Provenance Models.* Poster presentato al *Building Global Partnerships - RDA Second Plenary Meeting in Washington DC, 16-18 September 2013*.

URL=http://www.fragmentarytexts.org/wp-content/uploads/2013/09/LTH_RDAPoster_2013.pdf. [ultima visita 3.3.2014].

Blackwell C., Crane C. (2009). *Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital Age*. «Digital Humanities Quarterly», vol. 3, no 1. URL=http://www.digitalhumanities.org/dhq/vol/003/1/000035/000035.html. [ultima visita 3.03.2014].

Böckh A. (1858). *Gesammelte kleine Schriften,* vol 1. Druck und Verlag von B.G. Teubner.

Böckh A. (1877). *Encyklopädie und Methodologie der philologischen Wissenschaften*. Druck und Verlag von B.G. Teubner.

Boschetti F., Romanello M., Babeu A., Bamman D., Crane G. (2009). *Improving OCR Accuracy for Classical Critical Editions.* In Agosti M. *et al*., a. c. di, *Research and Advanced Technology for Digital Libraries,* vol. 5714, Springer-Verlag, pp. 156-167.
URL=http://link.springer.com/chapter/10.1007%2F978-3-642-04346-8_17. [ultima visita 3.3.2014].

Bryant M., Blanke T., Hedges M., Palmer R. (2010). *Open Source Historical OCR: The OCRopodium Project*. In Lalmas M., Jose J., Rauber A., Sebastiani F., Frommholz I., a. c. di, *Research and Advanced Technology for Digital Libraries,* vol. 6273, Springer-Verlag, pp. 522-525.
URL=http://link.springer.com/chapter/10.1007%2F978-3-642-15464-5_72. [ultima visita 3.3.2014].

Manmatha R., Feng S. (2006). *A Hierarchical, HMM-Based Automatic Evaluation of OCR Accuracy for a Digital Library of Books*. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, pp. 109-118.
URL=http://dl.acm.org/citation.cfm?doid=1141753.1141776.
[ultima visita 3.3.2014].

Smyth H.W. (1920). *A Greek Grammar for Colleges.* American Book Company.

A collaborative tool for philological research: experiments on Ferdinand de Saussure's manuscripts*

Angelo Mario Del Grosso[1], Simone Marchi[1], Francesca Murano[2], Luca Pesini[2]
[1]Istituto di Linguistica Computazionale, CNR, Pisa, Italy
{angelo.delgrosso, simone.marchi}@ilc.cnr.it
[2]Dipartimento di Lettere e Filosofia, Università degli Studi di Firenze, Florence, Italy
francesca.murano@unifi.it, luca_pesini@yahoo.it

Abstract. The present paper describes a philological-computational tool developed by the Istituto di Linguistica Computazionale (ILC – CNR) of Pisa, aimed at creating a digital edition of Ferdinand de Saussure's unpublished manuscripts. Since the use of a digital edition and of the most modern computer technology allows a more in-depth research, the ILC is developing a set of digital tools in order to take advantage of both the documents and the related information added by the scientific community. The integration exploits the Java enterprise platform by organizing the different features in modules. Thus, the tool meets the following requirements: (i) converting legacy digital resources into valid XML documents (TEI compliant); (ii) parallel visualization among imported texts and related images; (iii) search and indexing; (iv) handling of variant readings; and (v) collaborative annotation.

Keywords: Computational and Digital Philology, Digital Scholarly Editing, Java Enterprise, Linguistics, XML Database, XML Text Encoding.

1. Introduction

As in many other fields, the philological work has undergone a process of digitalisation in the past decades, producing digital-born publications generated by software processing. Tools like the computational-philological tool described in the present article will allow to intervene in a collaborative environment and integrate existing publications, for example by adding missing information as an annotation (Bozzi 2013).

The scientific impact of the publication of many unpublished works is remarkable: the availability of new writings, all enriched with a scientific commentary and a computational lexical analysis, will concur to fill the

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

documentary gap. In the contemporaneous publishing scene, choosing a born-digital edition rather than a print-based one is particularly efficient to access and update all the inserted data, very easily and simultaneously (Buzzetti 2009); the use of web-based software components allows also to constantly update the software itself, by implementing new features and refining the current ones (Maguire 2013).

In the framework of the PRIN project funded by the Italian Ministry of Scientific Research (MIUR), the ILC developed a set of digital tools aimed to facilitate the research and take advantage of both the documents and the related information provided by the scientific community. In particular, ILC has developed a digital-philological tool that allows the philological and critical analysis of manuscripts.

The purpose of the PRIN project was to investigate the unpublished manuscripts of Ferdinand de Saussure, the father of European Structuralism, and to create technological tools supporting this research. The ILC tool supports the editorial work and the interpretation of the texts, allowing a more efficient and thorough research on Saussure's thought.

The ILC developed also a structured computational thesaurus (Ruimy *et al.* 2013),[1] to promote the semantic research and contribute to the understanding of Saussure's terminology. This work, in fact, is part of a broad action carried out by the ILC for the development of collaborative research infrastructures (stand-alone and web-based) for 'computational philology', for managing critical apparata – both of ancient and modern texts – and integrating them in morphological analysers. The importance of philological collaborative infrastructure is testified, for instance, by the Open Philology Project[2].

Although data computerisation is decisive in the study and valorisation of cultural heritage, it is not enough: today, it is essential to work through collaborative platforms and share the results globally.

It is important to note that computational philology is not to be confused with computational linguistics, even if they employ the same methodology. These disciplines exhibit a different approach to the texts: from the perspective of computational linguistics, "texts are serial sequences of textual units, whereas from the perspective of computational philology, texts (with variant readings) are parallel sequences of textual units that insist on the same tex-

[1] See http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=917/vers=ing
[2] Created by the Humboldt Chair of Digital Humanities at Leipzig.

tual positions" (Boschetti 2010, 1-2). In this context, we offer the case study of Saussure's writings.

2. The Ferdinand de Saussure's manuscripts: philological requirements and case-study

The theory of language formulated by Saussure was developed at the beginning of the 20th century and, since then, it has determined several implications in various fields of knowledge. The interest in the studies of Saussure, in his work as a linguist scholar on Indo-European languages and as a theoretician of philosophy of the language, has always been alive. Saussure was the enactor of revolutionary theories and methodologies during his lifetime and the posthumous publication of the *Cours de linguistique générale* determined new research lines in all Humanities.

The richness of Saussurean studies demands the acquisition of new unpublished manuscripts. Although Saussure's theories have received considerable attention, the number of unpublished works is still considerable: we count about 17000 folia, located in libraries worldwide (Bibliothèque de Genève, Swiss; Harvard's Houghton Library, USA; Bibliothèque Nationale de Paris, France) and only a low percentage of these manuscripts has been published.

During the PRIN collaborative work performed by the University of Florence unit, responsible for the processing of texts, and the ILC, the tool was customised to be adapted to the specific needs of the critical edition of Saussurean unpublished writings (Murano and Pesini 2013).

The PRIN case study was conducted on the manuscript about the sonorant in Indo-European, *Théorie des sonantes*, published by Marchese in 2002 (Marchese 2002), which allowed us to compare the differences between a traditional and electronic edition.

The ILC tool facilitates the study of manuscripts, being designed to perform multiple functions: to explore simultaneously different data (e.g. text and images), create a multilevel structured critical apparatus, generate accurate lexical indexes, perform a multilevel analysis of the text. In order to improve the efficiency of the tool, the first step was the digitalisation of the Marchese's edition, which was already in digital format, provided as a set of Microsoft Word files (.doc).

Saussurean writings are heterogeneous in many aspects (chronology, topic, etc.) and, therefore, it is mandatory to have information about the manuscripts (pressmark, description, intellectual content, etc.), so that scholars can retrieve every kind of information through queries and identify only specific sets of texts (e.g. those related to a time span or to a specific topic, etc.).

Philological accuracy is the essential requirement for a critical approach to the Saussurean manuscripts, which require careful and detailed analysis. Particularly, it is very important to consider all the editorial phases (corrections, marginal additions, erasures and re-writings) that brought Saussure to formulate his linguistic theories.

Another important requirement concerns the visualisation of the photographic reproductions of the manuscripts. The tool allows the simultaneous presentation of many informative layers, including the digital image and the relative transcription and annotations, so that scholars can perform an autoptic examination with the support of digital pictures.

Moreover, editors are enabled to comment and add critical and philological notes about different aspects of the intellectual content of the manuscript. Unlike the printed edition, where all comments are indiscriminately included in the footnotes, the digital edition allows to make categorised annotations and comments, so that the user can select the typology he needs. We decided to include: *notes théoriques* (critical reflections of linguistics and philosophy), *personne* and *bibliographie* (respectively, named-entities and bibliography in the text), *glossaire* (relevant terminology), *notes critiques au texte par l'auteur* and *par l'éditeur* (philological notes referring to interventions respectively by Saussure and by the editors), *notes supplémentaires* (free comment).

Philological notes constitute the critical apparatus; the *notes théoriques* contain the hermeneutical comment to the text; the other typologies allow the analysis of the text and the creation of indexes. The note *bibliographie* contains the full-normalised citation and references to the bibliographic source (independently from Saussure's way of citation, which was quite incomplete or abbreviated); there are two external links, to the Italian Catalogo Unico OPAC[3] and to Internet Archive Project,[4] chosen to take advantage of open-source projects created by the community. The annotation *personne* is useful to disambiguate bibliographic items when Saussure uses the scholar's name

[3] http://www.iccu.sbn.it/opencms/opencms/it

[4] https://archive.org

as a metonymy for his work: in the note the names occur in normalised form and are followed by the birth and death dates and by a link to Wikipedia, yet in a collaborative perspective. So we can identify the works and scholars cited by Saussure and retrieve the total number of occurrences.

Given that, Saussure, in *Théorie des sonantes*, cites words in many languages, we created a lexical index with the various forms used in the text, listed by each language. For each word we have both the total number of occurrences and the textual context. The tool scans the texts and builds a list of search terms; thus, the user is enabled to perform specific queries, for instance, to analyse the evolution of Saussure's meta-language.

Moreover, it is possible to visualise the text in different ways: per folium, per groups of folia, or the manuscript full text. On the one hand, scholars interested in the editor's annotations can select the folium/folia and the relative annotations, while, on the other hand, non-specialist users, not interested in critical apparatus, can access directly the plain text.

3. The ILC philological-computational tool

Computational philology, on the one hand, attempts to shift the attention from the classic means (printed edition), through which a text is compiled and exploited, to more advanced and digital technologies. On the other hand, the broader goal aims to handle digital data and electronic information, by taking into account the overall scholarly editing process (Bozzi 2006).

The underlying model (fig. 1) describes a digital philological entity as a complex and structured object that can evolve in space and in time, creating variants. Furthermore, the entity is interconnected with objects either of the same type or of different nature (linked data). Hence, our efforts have led to the design of a software environment able to manipulate objects, which are composed of four basic components: a) version; b) interpretation; c) granularity; d) position.

The cornerstone of the model is the handling of primary sources through a parallel management and visualization, as well as the possibility to annotate, comment, and analyse them. The functionalities available for the PRIN project are extensions of the basic core functions and can be instantiated for the needs of a specific project.

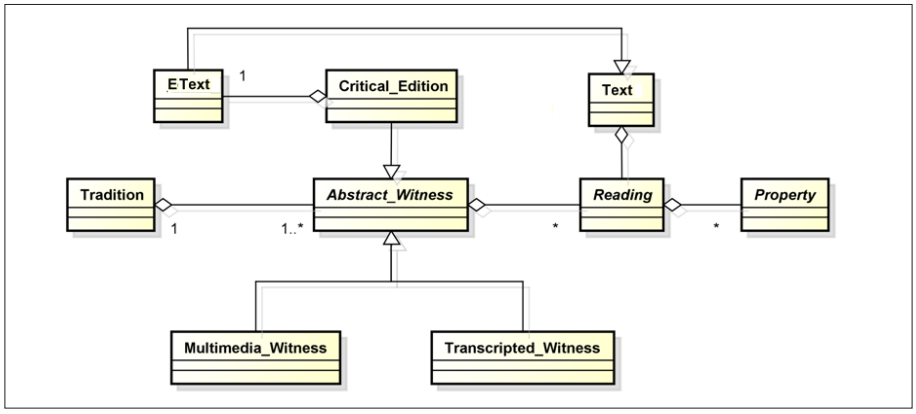


Fig. 1. The image describes the basic object-oriented model, which uses the Unified Model Language. The schema formalizes the textual tradition, taking into account potential witnesses (text transcription and/or facsimile images) and variant readings. Variant readings (Reading) have specific properties (Property) which intent to describe them.

The system, therefore, ought to be made up of modular and flexible MVC components (Model View Controller) for the sake of extendibility, adaptability and usability.

The process of software design and development of such an environment must be dynamic and should pay great attention to both defining the model, by using a formal/semi-formal language (i.e., UML), and applying the most effective design patterns (e.g., composite component, factory, observer, strategy, etc.). Drawing a basic class diagram is important, in order to realize an object oriented and technology independent view of the domain. Similarly, users play a fundamental role, especially for what concerns the system testing and use case refinement. Given all this, we can adopt the most efficient methodologies in software engineering promoted by the progressive enhancement strategy (i.e., 'agile' design) (Martin 2003).

The framework is designed and developed as a web-based application, able to provide modules for the use of standardised mark-up schemes and guidelines for text encoding purposes (e.g., XML-TEI P5[5]) (Burnard and Baumann 2008). It includes and uses open source libraries, in order

[5] http://www.tei-c.org/

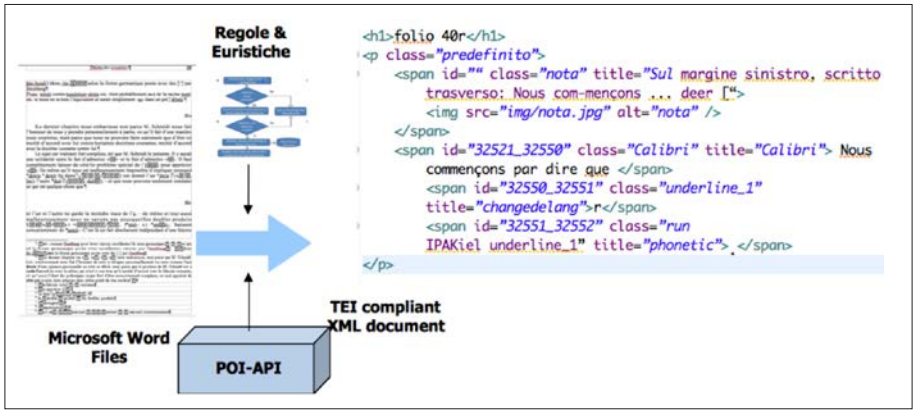


Fig. 2. Importer. The electronic version of the text (e.g. old Microsoft .doc format) is converted by the importer component by using the Apache POI open source library alongside heuristics and regular expressions. The result of this process is a TEI compliant encoded file stored in the eXist XML database.

to enhance capabilities (such as, Apache-Lucene, JDom, IBM-ICU, jQuery, OpenCV, etc.), and will be released under the GPL license.

The Saussure project case study gave us the possibility to deploy the system to comply with real philological requirements and the opportunity to test its flexibility and modularity. The most important components of the system implemented for the Saussure project are the following (Del Grosso and Marchi 2013):

- *Importer*: this module allows us to obtain an electronic well-formed edition of a text published and edited in a printed format, or in a legacy digital encoding. The module, for the specific needs of the project, parses and transforms old files, edited by way of an old processing editor, in order to recover the content. Its functionalities seek to decode and encode the text and return a well-formed Unicode encoding of the characters, which are out of the ASCII range. Moreover, the module preserves the format and the styles by using the XML mark-up language (fig. 2).
- *Text-Image viewer*: the viewer aims to present the content of a manuscript or the content of a page in parallel with its representation (facsimile) (fig. 3).

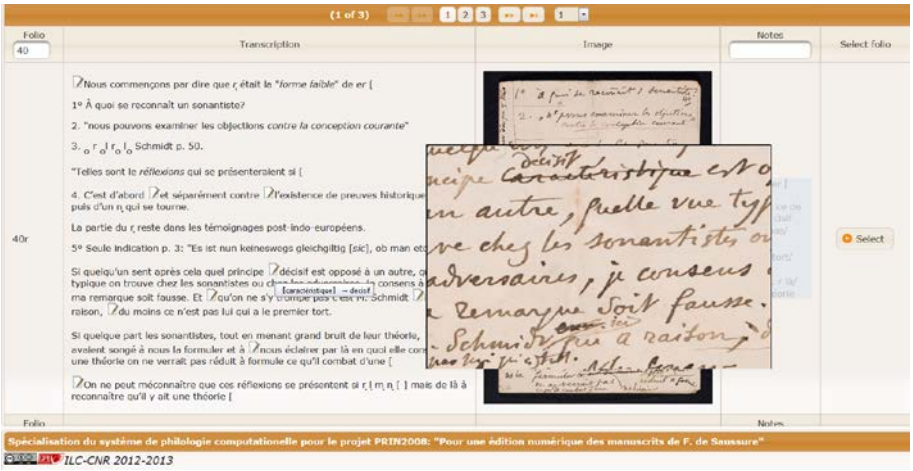


Fig. 3. Parallel view. Text and image are shown together in order to provide different kinds of view to the scholars. Philological notes in the text can be expanded simply using mouse rollover. The image can be zoomed pointing the mouse over it.

- *Digital book viewer*: this viewer presents the content of a critical edition in a 'mise en page' look and feel.
- *Multilingual indexes*: different categories of words are provided in order to handle the great lexical variety used by Saussure (fig. 4).
- *Search component*: both the lexical data and the data added by the user, are searchable in order to retrieve the right manuscript passage, the related concordances or annotations (fig. 4).
- *Variant handling*: the history of the text edited by the author is preserved and integrated in the content as para-textual information. Thanks to this module, users can search the variants and view them as a tooltip while reading the folia along side their own images (figg. 2-3).
- *Linguistic analysis*: this module performs linguistic analysis on the text and provides morphological analysis, lemmatization and classification. Since the Saussure's manuscript includes many languages (e.g. Greek, Latin, etc.), the system customized for this project does not provide any automatic linguistic analysis tools. We performed, instead, a semi-automatic language classification of each word extracted from the text by using some online linguistic resources (i.e. Wiktionary).[6]

[6] http://www.wiktionary.org/

Fig. 4. Search. Some indexes are available for aiding the scholars during the search operations. A semi-automatic indexing process is made up for each recognized language, in order to ensure higher accuracy.

- *Collaborative annotation*: this module is the core of the collaborative workflow. Scholars can select chunks of text at different granularities and add their personal annotations to them by using a natural language input area or by tagging by means of a list of categories (fig. 5).

The technological environment is based on the Java Enterprise Platform.[7] It oversees the integration of the data and the user experience (UX). The overall system has been developed following the Server Faces Framework (JSF2)[8] and the Model View Controller (MVC) architectural pattern (Glenn *et al.* 1988). TEI-Compliant encoding documents are stored in an eXist-db[9] (XML oriented database) and the platform is synchronized with it for the data management. Therefore, the 'view tier' is the web, the 'business logic tier' consists of the objects managed by the java beans, and the 'data/integration tier' is achieved by the XML native database.

[7] http://www.oracle.com/technetwork/java/javaee/overview
[8] https://jcp.org/aboutJava/communityprocess/final/jsr314
[9] http://exist-db.org/

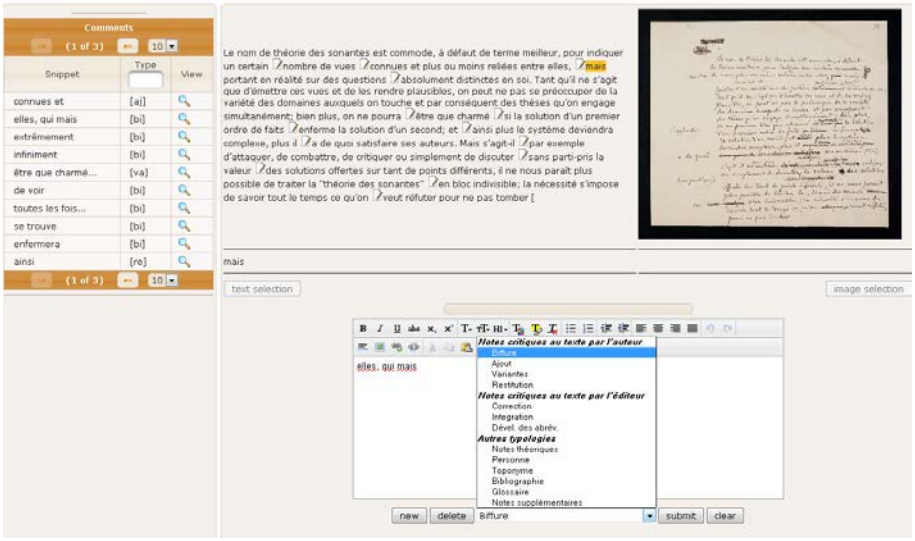


Fig. 5. Comments. The annotation component allows to add comments and editorial annotations, categorized by the scholars. A searchable list summarizes the collaborative work on the text.

As mentioned before, the application is designed as a collaborative multi-layered application and handles the presentation logic by making use of the World Wide Web. We have used two complementary java enterprise technologies: a) FACELETS and b) Primefaces.[10] The first one is a component-oriented technology for web templating; its benefits are represented by efficient writing code and in effective software reusing. The second one is a rich and friendly Ajax taglib that allows the developing of flexible user interface (GUI).

4. Conclusions and Further works

The synergy between the various components developed in this project guarantees, already, a large and simultaneous accessibility to the information concerning Saussure's manuscripts, facilitating both the scientific research

[10] http://primefaces.org/

and the exploit of all data and related information by non-experts. This work contributes to extend the use of innovative digital editions and digital management functions, through a set of specific technologies and research infrastructures, tuned for the study of Saussure's unpublished works.

The legacy left by Saussure in the field of Linguistics is limitless. Saussure is mainly known for his theories of language. Particularly, the importance of his studies on Indo-European languages is immeasurable, since they still offer scholars very valuable suggestions. Furthermore, Saussurean studies stimulate new debates and are the starting point for new interesting researches not only in the field of Linguistics, but in the Human Sciences as a whole. Thus, the complexity of Saussure's thought and its importance in the development of Western Linguistic thought make crucial an in-depth study of Saussurean unpublished manuscripts.

From the perspective of a collaborative research, we will implement multiple levels of theoretical notes, and enable other researchers to comment the texts, always distinguishing between notes left by the first publisher and notes produced by later editors.

Acknowledging the importance of open sources in modern research, we will publish our work as open data, in order to promote the circulation of knowledge concerning Saussurean texts and of philological contributions to these studies.

During the last working phases of the project we have tracked some important improvements. First of all, the most desirable upgrade is to develop a software component able to connect the system to the Saussure lexical resource created in the PRIN project.

At the moment, the images are stored directly into the file-system. In the future, we plan to improve the application by using an advanced high-performance feature-rich image server system for web-based streamed viewing and zooming of ultra high-resolution images (e.g. IIP server).[11] The relation between texts and images has to be enhanced to allow users to create a finer grain link among a chunk of text and a selection of the related image, using open source libraries such as those of the DigiPal[12] palaeographic project.

Linguistic tools can be embedded to add morphological and syntactical information to the indexed texts. The information derived from this process allows the user to perform more powerful searches.

[11] http://iipimage.sourceforge.net/
[12] http://www.digipal.eu/

Finally, the graphical user interface usability can be improved. For example, the visualization of the results can be enhanced by highlighting the image selections as well as the texts. Furthermore, a dynamic resource upload and suitable management of new documents have been planned.

In conclusion, the challenge of our research is to model and develop a general solution for the automatic management of the critical apparatus. Thus, an editing component for scholars could be plugged into the system in order to create different edition of the same text[13].

5. References

Boschetti F. (2010). *A Corpus-based Approach to Philological Issues*, thesis submitted for the degree of Philosophiae Doctor, Università di Trento. URL= http://eprints-phd.biblio.unitn.it/185/ [last visited 02.03.2014]

Bozzi A. (2006). *Edizione elettronica dei testi e filologia computazionale*. In A. Stussi, a c. di, *Fondamenti di critica testuale*, Il Mulino, pp. 207-232.

Bozzi A. (2013). *G2A: a Web application to study, annotate and scholarly edit ancient texts and their aligned translations. Part I. General model of the computational philology application*. «Studia graeco-arabica. The Journal of the Project Greek into Arabic Philosophical Concepts and Linguistic Bridges», vol. 3, no 1, pp. 159-171. URL=http://www.greekintoarabic.eu/uploads/media/BOZZI_SGA_3-2013.pdf [last visited 10.03.2014]

Burnard L., Bauman S. (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Oxford. URL=http://www.tei-c.org/Guidelines/P5/ [last visited 02.03.2014]

Buzzetti D. (2009). *Digital Editions and Text Processing*. In M. Deegan, K. Sutherland, eds., *Text editiong print and the digital world*, Ashgate, pp. 45-61.

Del Grosso A., Marchi S., *Una applicazione web per la filologia computazionale. Un esperimento su alcuni scritti autografi di Ferdinand de Saussure*. In D. Gambarara, M. P. Marchese, eds., *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*, Dell'Orso, pp. 131-157.

Krasner G. E., Pope S. T. (1988). *A cookbook for using the model-view controller user interface paradigm in Smalltalk-80*. « J. Object Oriented Programming». 1, 3 (August 1988), pp. 26-49.

Maguire M. (2013). *Using human factors standards to support user experience and agile design*. In *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, Springer, pp. 185-194.

[13] We would like to thank our colleague Alessia Belusci for proof-reviewing the paper.

Marchese M. P., ed. (2002). *Ferdinand de Saussure. Théorie des sonantes: il manoscritto di Ginevra BPU Ms. fr. 3955/1*, Unipress.

Martin R. C. (2003). *Agile Software Development: Principles, Patterns, and Practices*, Prentice Hall PTR.

Murano F., Pesini L. (2013). *L'edizione digitale dei manoscritti di Ferdinand De Saussure: l'analisi dei requisiti e il caso di* Théorie des sonantes. In D. Gambarara, M. P. Marchese, a c. di, *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*, Dell'Orso, pp. 121-129.

Ruimy N., Piccini S., Giovannetti E., Bellandi A. (2013). *Lessicografia computazionale e terminologia saussuriana*. In D. Gambarara, M. P. Marchese, a c. di, *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*, Dell'Orso, pp. 161-179.

Edition Visualization Technology: a tool to publish digital editions*

Raffaele Masotti, Julia Kenny
Università di Pisa, Pisa, Italia
{raffaele.masotti, julia.kenny90}@gmail.com

Abstract: EVT (Edition Visualization Technology) is a lightweight software for the creation of image-based web editions of TEI P5 encoded texts. It is built on open and standard web technologies, such as HTML, CSS, JavaScript, etc., and works as a client-only infrastructure. The final web application provides instruments to fully take advantage of manuscripts' scans and transcripts (such as a magnifying lens, a general zoom, hot spots, image-text link).

Keywords: XML, TEI, XSLT, digital edition, digital publishing, web publication, image based edition, diplomatic edition, user studies.

1. Introduction

The EVT (Edition Visualization Technology)[1] project was born as part of the VBD (Digital Vercelli Book)[2] project in order to allow the creation of a digital edition of the Vercelli Book, a parchment codex of the late tenth century, now preserved in the Archivio e Biblioteca Capitolare of Vercelli and regarded as one of the four most important manuscripts of the Anglo-Saxon period as regards the transmission of poetic texts in the Old English language.

In order to realize a digital edition of a manuscript, the choice – or the eventual development – of a viewer is the last step in a long process that also includes the digitization of the manuscript folios and the transcription of the texts. The choice of the encoding scheme to be used for the transcription and mark-up of the manuscript texts is a crucial point of this long process

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.
[1] http://sourceforge.net/projects/evt-project/
[2] http://vbd.humnet.unipi.it/

and it influences the choice of the visualization tool to use. In the specific case of the Vercelli Book, it was decided to adopt the TEI encoding schemes for the transcription of texts.

There are several tools that allow to convert XML files encoded according to the TEI schemes into web pages ready to be published on the web. Unfortunately, none of the tools that we researched has been able to provide a satisfactory answer to the task of creating an image-based edition. For this reason, we decided to create a new tool designed so to be generally compatible with files encoded in TEI P5 format. To ensure a high level of modularity, we decided to logically separate the part of code used for the generation of the web application from the part which includes the style sheets used for TEI elements transformation.

2. Basic principles

The idea behind the project is to simplify the production of digital editions, freeing the scholar from the burden of dealing with the web programming required to build the site hosting the edition. This project started with having a few fundamental goals, briefly summarized below:

- to allow the visualization of the manuscript images, providing a set of tools for a more accurate investigation of the latter;
- to allow the comparison between the digital scans of the manuscript and the corresponding text of the edition;
- to allow the comparison between different edition levels of the text.

The set of principles that has been traced in order to achieve these goals is still the foundation of the project. Furthermore, we decided to use non-proprietary formats and languages, both to ensure the durability of the resulting web-based edition and to ensure the possibility to modify and redistribute the software according to the open source model of development.

3. Content preparation and initial configuration

In the Digital Vercelli Book project, as it is usually the case for similar initiatives of image-based diplomatic editions of medieval manuscripts, the edition data consist of TEI XML documents, which may include more than one edition level, and digitized images of the manuscript. Before you can

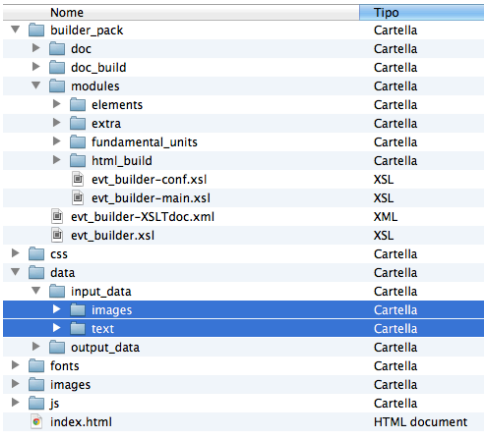


Fig. 1. Project folders.

use EVT it is necessary to copy such content within the appropriate folders (fig. 1.): there are different sub-folders for text (text) and images (images) inside the data/input_data folder.

The system also provides the possibility of customizing the digital edition by configuring some parameters and variables. The most interesting customizations are as follows:

- the personalization of the file path, through which the user can specify absolute paths while maintaining consistent references within the web application. For example, the user might want to indicate that the image files are located on a different server than the one hosting the web application;
- the presence or absence of the image frame, for cases in which the scans are not yet available for the edition. When disabling the image frame the system will prepare the skeleton HTML so that it displays only the text of the transcript and not the images;
- the generation of different types of edition text, in which you can both customize each edition level name and enable or disable the generation of that level. The system has been developed to allow the generation of an arbitrary number of edition levels and the web interface is designed to simultaneously manage different levels of edition.

Changing these parameters allows the user to define the shape of the final output without requiring modifications of the project source code. Of

course further customization is possible, in particular the user can modify the current edition-related style sheets or even add new ones for specific purposes (see below).

4. How it works

After all the edition data has been copied in the correct locations and the configuration parameters have been set, starting EVT is quite simple: the TEI P5 document containing the edition texts is opened in an XML editor such as Oxygen and an XSLT stylesheet is applied to it; this starts a chain of transformations, sequentially calling other style sheets, and the final result is a web application (fig. 2) composed of a mix of HTML, CSS and Javascript.

Through the application of the above mentioned XSLT style sheets, an HTML home page and a set of HTML files – each corresponding to an individual folio of the manuscript – are automatically generated and hierarchically organized into subfolders according to their edition level. The index home page dynamically calls the other HTML files produced.

Each of the HTML files contains references to other external files, in particular JavaScript functions and CSS style sheets. The JavaScript functions are implemented by means of the jQuery framework and handle the interactions with the user or provide appropriate tools for the visualization and the navigation of the manuscript, while the CSS style sheets define the layout and allow to attach the Junicode[3] font to each page, ensuring the visualization of non-standard characters.

The use of plug-ins has been limited as much as possible in order to reduce dependencies on third-party components, except in cases of functions that were not implementable otherwise, such as plug-ins for image tools. The general structure of EVT is designed to be as modular as possible, so to ease project maintenance and ensure further expansion.

Hence, on a practical level, those who use the system are exempted from directly dealing with the programming, what they only have to do is to perform the XSLT transformation with an XML editor or any other software that includes an XSLT 2.0 processor.

Of course this is not the only possible approach. There are some other interesting alternatives to manage XML data, such as native XML databases

[3] http://junicode.sourceforge.net/

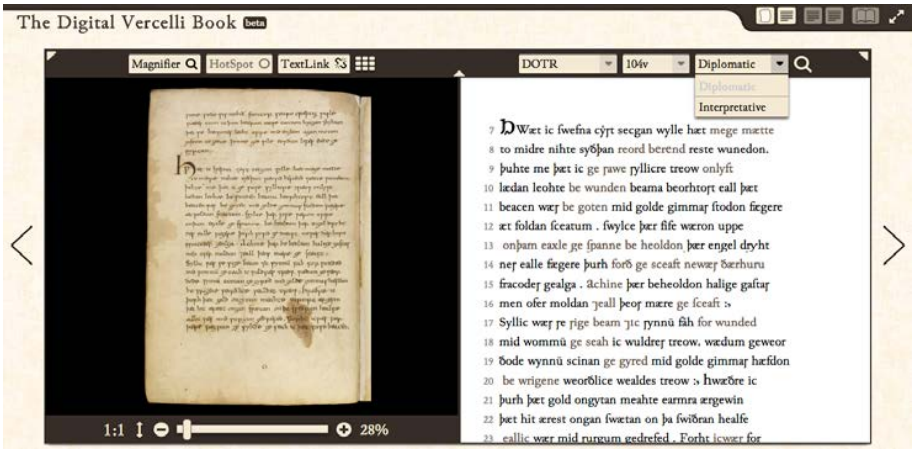


Fig. 2. Main view of the web application.

(e.g. eXist-db[4]). In other words, systems that use documents themselves as fundamental storage units, exploiting the structure of files to extrapolate relationships and data. Nonetheless, such systems are not suitable for immediate use, for they require a specific proficiency in information technology, due to their high level of complexity.

5. The XSLT transformation

For the purpose of a diplomatic or semi-diplomatic edition it is essential that the text of the transcription be divided into smaller parts as to recreate the physical structure of the manuscript. Therefore, paginated XML documents must be marked using a TEI <pb/> (page break) element at the start of each new page, so that the transformation system is able to recognize and handle everything that stands between a <pb/> element and the next one as the content of a single page.

The EVT builder's transformation system is composed by a modular collection of XSLT 2.0 style sheets: these modules are designed to permit the scholar to freely add his own style sheets and to manage all the desired levels

[4] http://exist-db.org/exist/apps/homepage/index.html

of the edition without influencing other parts of the system, for instance the generation of the index home page (index.html).

The transformation chain has two main purposes: generating the HTML files containing the edition and creating the home page which will dynamically recall the other files.

```
<xsl:template name="page">
    <xsl:variable name="pb_n" select="self::tei:pb/@n"/>
    <xsl:for-each select="$edition_array">
        <xsl:if test=".!=''">
            <xsl:variable name="edition_current" select="lower-case(.)"/>
            <xsl:result-document method="html" encoding="UTF-8" media-type="text/plain"
                byte-order-mark="yes" href="{$filePrefix}/data/output_data/{$edition_current}/
                page_{$pb_n}_{$edition_current}.html" indent="yes">
                <xsl:call-template name="data_structure">
                    <xsl:with-param name="output" select="$edition_current"/>
                    <xsl:with-param name="pb_n" select="$pb_n"/>
                    <xsl:with-param name="edition_pos" select="position()"></xsl:with-param>
                </xsl:call-template>
            </xsl:result-document>
        </xsl:if>
    </xsl:for-each>
</xsl:template>
```

Fig. 3. Example of a XSLT template.

EVT builder's transformation system uses a collection of style sheets to divide the TEI P5 XML file containing transcription into smaller portions, each one corresponding to the content of a folio, recto or verso, of the manuscript. For each of these text fragments it creates as many output files as requested by the file settings, that is one for each edition level. Then, in order to create the contents of these files, templates required to handle the desired edition level are selected by using the value of their mode attribute.

For example, by associating the transformation for the interpretative edition level to the int mode, the correct content of the pages is obtained applying all the templates that have int as value for @mode to the text of that page. This requires that each one of the pages selected by the system is processed by an <xsl:apply-template select="current-group" mode="int"/> instruction before the text of the page is inserted in a interpretative output file.

As already stated above, the editor will be able to freely add his own style sheets in order to personally manage the different levels of the desired edition. More in detail, what is required from the user is to:

- personalize the edition generation parameter;
- copy his own XSLT files containing the template rules to generate the desired edition levels in the directory that contains the style sheets used for TEI elements transformation;
- include the new style sheets in the file used to start the transformation chain;
- associate a mode value to the new edition level transformation;
- add a mode attribute with the new transformation value to all the template rules that are used for that transformation.

6. Tools and Features of the web-based edition

On the images side, the system includes a ready to use set of tools (zoom, magnifier, thumbnail navigation, etc.) for a full fruition of the scans of the manuscript.

One of the most important features is the image-text link (fig. 4.): if the XML file contains the elements that are necessary for the activation of this tool, clickable areas corresponding to the lines, both of the image and of the text, are automatically produced during the transformation. When the image-text link is activated, you can browse with the mouse the content of

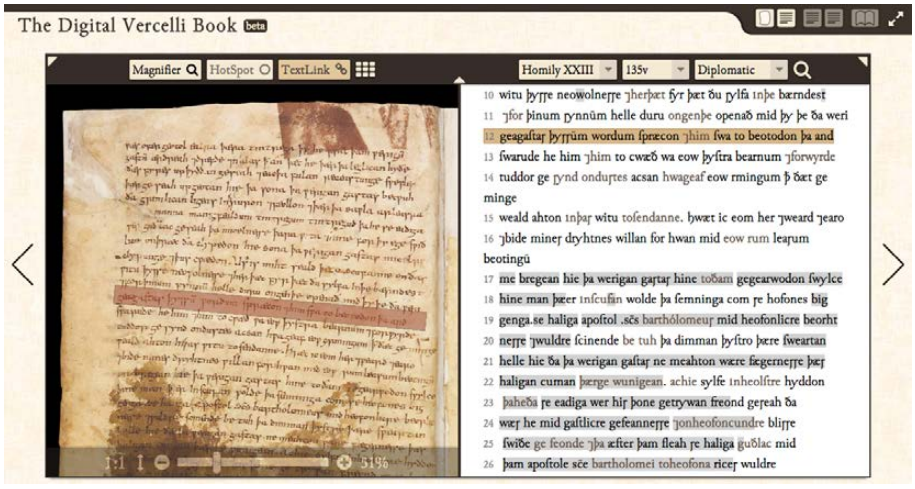


Fig. 4. Image-Text link.

an edition line by line, highlighting in parallel its original form on the scan of the folio, and vice versa; this feature is particularly useful for teaching purposes.

In order to create the image-text link, the original XML document must contain information about areas of the image corresponding to each line and each line of text (marked with <lb/>) must have a reference to the corresponding area. This means that each line of the image must be represented by a <zone> element with the following attributes:

- an identifier (@xml:id);
- an attribute indicating that it represents a line (@renditition = "Line");
- the coordinates of that area;
- if necessary, the angle of inclination of that area (@rotate).

The corresponding <lb/> element must contain a reference to that zone by means of the attributes @facs or @corresp with the value of the @xml:id of the zone.

Note that to give provide coordinates for a rectangular area it is sufficient to indicate the coordinates of the upper left and the lower right corners angle, from which position and size of the area can be derived.

The hotspot tool allows to display additional information (notes, restored images, etc.) regarding specific details of the images. In order to create the hotspots, the additional information must be marked in the XML document and the corresponding areas of the image, that is the <zone> elements, must have "hotspot" as value of the @renditition attribute.

At the first loading, the edition is presented in an interface that shows the folio image on the left side of the screen and the text on the right side, but the user can switch the visualization mode choosing between the following:

- Image-Text view (default view): it shows a manuscript folio image and the corresponding text, the user can choose from the drop-down menu in the text frame toolbar the edition level to visualize;
- Text-Text view: conceived to compare different edition levels;
- Bookreader: this view expands the image frame to show double side images of the manuscript.

The user can change the visualization mode using the appropriate buttons at the top right corner of the interface; the transition from one mode to another is done in a convenient way for the user as the structure of the interface is changed keeping the content position (folio number).

7. Future Developments

The project, which at the moment of this writing is about to enter a beta phase, has several features currently in development; among these, the most important one is the introduction of a text search engine. The goal of the search engine is to offer a free-text and also XML-based searching functionality which the user can use directly into the locally generated web pages without the need of complex server side installations. This feature is already available in the developer release and allows both to quickly search through the texts of all the edition levels generated and to highlight the searched word(s) within the pages, making it easier to identify them. A medium-term objective is to extend these functions also to data stored on other servers, thus allowing cross-searches on different texts.

EVT was born in the context of a specific use case, characterized by the presence of a single witness and the need to realize a diplomatic-interpretative edition. However, our plan is to make it compatible with a growing number of case of studies. The system is continuously tested with other texts (e.g. the corpus NZTEC[5]) and we are currently working on critical edition and embedded transcription support. While we are technically facing these problems, we are also asking ourselves about the possible choices to provide an innovative layout that allows the user to enjoy the content in the most effective way.

8. References

Editions and digital facsimiles

Biblioteca Apostolica Vaticana: http://www.vaticanlibrary.va/home.php [accessed on April 2014].

Codex Sinaiticus: http://www.codex-sinaiticus.net/en/manuscript.aspx [accessed on April 2014].

Foys M. K. (2003). The Bayeux Tapestry: Digital edition [CD-ROM]. Leicester: SDE.

Kiernan K. S. (2011). Electronic Beowulf [CD-ROM]. Third edition. London: British Library.

The Dead Sea Scrolls: http://www.deadseascrolls.org.il/ [accessed on April 2014].

[5] http://nzetc.victoria.ac.nz/

New Zealand Electronic Text Collection: http://nzetc.victoria.ac.nz/ [accessed on April 2014].
Vercelli Book Digitale: http://vbd.humnet.unipi.it/ [accessed on April 2014].

Software tools

DFG Viewer: http://dfg-viewer.de/en/regarding-the-project/ [accessed on April 2014].
DM Tools: http://dm.drew.edu/dmproject/ [accessed on April 2014].
TEIBoilerplate: http://teiboilerplate.org/ [accessed on April 2014].
tei2html Jawalsh: https://github.com/jawalsh/tei2html [accessed on April 2014].
Das intellektuelle Berlin um 1800: https://sites.google.com/site/annebaillot/digitale-edition [accessed on April 2014].
TEICHI: http://www.teichi.org/ [accessed on April 2014].
The TEIViewer project: http://teiviewer.org/ [accessed on April 2014].
eXist-db: http://exist-db.org/exist/apps/homepage/index.html [accessed on April 2014].

Essays and other references

Buzzetti D. (2009). *Digital Editions and Text Processing*. In M. Deegan, K. Sutherland, eds., *Text Editing, Print, and the Digital World*, pp. 45-62. Digital Research in the Arts and Humanities, Ashgate.
Foys M.K., Bradshaw S. (2011). *Developing Digital Mappaemundi: An Agile Mode for Annotating Medieval Maps*. «Digital Medievalist» vol. 7. URL=http://www.digitalmedievalist.org/journal/7/foys/ [accessed on April 2014].
O'Donnell D.P. (2007). *Disciplinary impact and technological obsolescence in digital medieval studies*. In S. Schreibman, R. Siemens, eds, *A companion to digital literary studies*, Blackwell, pp. 65-81. URL=http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-4-2 [accessed on April 2014].
Rosselli Del Turco, R. (2006). *La digitalizzazione di testi letterari di area germanica: problemi e proposte*. Atti del Seminario internazionale 'Digital philology and medieval texts' (Arezzo, 19 – 21 Gennaio 2006), Sismel.
Rosselli Del Turco, R. (2011). *After the editing is done: designing a Graphic User Interface for Digital Editions.* «Digital Medievalist» vol. 7. URL=http://www.digitalmedievalist.org/journal/7/rosselliDelTurco/ [accessed on April 2014].
TEI Consortium, eds. *Guidelines for Electronic Text Encoding and Interchange*. V. P5 (31 January 2013). URL=http://www.tei-c.org/P5/ [accessed on April 2014].

Codifying the codex. The digital edition of the *Becerro Galicano* of San Millán[*]

David Peterson[1]
Department of Medieval History, University of the Basque Country, Vitoria-Gasteiz, Spain
vpeterson@euskaltel.net

Abstract. The creation of a Digital Edition of the 12th century monastic cartulary known as the *Becerro Galicano* of San Millán has allowed us to develop an innovative range of search tools and associated functionality while resolving problems that have traditionally dogged the paper edition of such volumes. Most notable, a lemmatised index of the cartulary's contents allows the scholar to overcome the orthographic irregularities and morphological variety that limit the usefulness of traditional indices. Cartographical functionality has also been developed, allowing the user to visualise the geography of any given document, group of documents, person, word or lemma appearing in the codex. And the traditional problem of whether to order the contents of such volumes chronologically or codicologically has now been resolved, as both sequences are made available to the user.

Keywords: Digital Edition, Cartulary, Becerro Galicano, Medieval History, Philology, Lemmatisation, Automatic map generation.

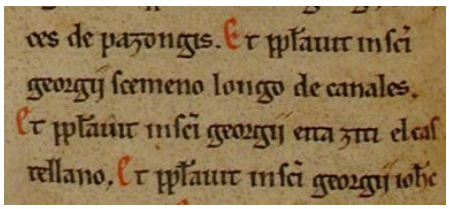
1. Introduction

San Millán de la Cogolla is a monastery in the Rioja region of northern Spain. The importance of its documentation is recognised by historians and philologists alike, as it is fundamental to the early medieval history of the regions of Castile, Navarre and the Basque Country, while witnessing the emergence of both Castilian and Basque as written languages. The single most significant codex in this sense is the *Becerro Galicano.* Composed

[*] M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. Cleup, Padova, 2014.
[1] This paper has been developed within the "Grupo de Investigación Consolidado *Alta y Plena Edad Media"* (Gobierno Vasco IT536-10) and more specifically as part of the research project "De los cartularios al territorio, la iglesia y la sociedad: edición digital y estudio crítico del *Becerro Galicano* de San Millán de la Cogolla" Micinn (HAR2010-16368).

around 1195, it is a cartulary, in other words, a volume containing copies of the abbey's charters, many of which date from centuries before, the earliest from the year 759. Covering some 250 parchment folios, written in a single elegant and regular Caroline hand, the cartulary contains 770 documents, and in total some 200,000 words, including 3,000 different place-names. The contents are ordered not chronologically, but geographically.

Although the codex is basically in Latin, it is linguistically rather complex, as fragments of both Romance vernacular and Basque are recorded, the latter generally in place-names, as well as some Arab vocabulary absorbed into the Romance. To illustrate this we reproduce below a brief but rich extract from the Becerro of a text dated to the period 1022-1076, in which a Latin verb (*populavit*) is mixed with early Spanish (use of prepositions, the definite article), an Arab personal name (*Ziti*), a Basque word (*eita*, meaning 'father'), and an early reference to the emergent county of Castile. The example, hopefully, serves to qualify the idea of a monolingual Latin text, illustrating the complexity and wealth of the material.



… de Pazongis. Et populavit in Sancti Georgii Scemeno Longo de Canales. Et populavit in Sancti Georgii eita Ziti el castellano. Et populavit in Sancti Georgii Iohannes …

Fig. 1. Extract from *Becerro Galicano Digital* #12 (f. 6v) illustrating the linguistic complexity of the codex.

This convergence of interests of philologists and historians was the origin of the project presented here as, concurrently, a team of medieval historians from the University of the Basque Country and one of philologists from the University of The Rioja, frustrated by the inadequacies of previous paper editions of this material, decided to re-edit the *Becerro Galicano*, resulting in this digital edition: http://www.ehu.es/galicano/?l=en. In this paper I will focus on three different aspects of the project that illustrate how digital edition can resolve problems that have dogged paper editions for centuries and offer tools and facilities beyond the scope of such traditional editions, while also observing how some existing technologies and resources have proved of frustratingly limited use to us.

2. Ordering and reordering the cartulary's contents

The cartulary genre poses a particular problem. Such volumes work on two levels: the codex as a whole, with a date of composition, authorship, and a logic behind its composition and structure; and its component parts – in our case 770 of them – each regarded as a document in its own right, with an original date of composition, author etc. The subdivision of a codex into different component parts is of course not in itself unusual, what distinguishes cartularies from other codices is the question of how to order the component parts when editing the material. As already mentioned, generally such volumes are not structured chronologically, and yet when they have come to be edited in the past, their contents were traditionally re-ordered by date. Over the last two decades, scholars of the genre have come to question this approach, arguing that it ignores the logic behind the structure of these volumes and thus makes their contents harder to understand. Thus there are two competing ways to present the material contained in a cartulary: chronologically or codicologically.

To give an example, we reproduce another folio (f. 4r) from the *Becerro* (fig. 2), which on just one page documents the acquisition at different times of some 21 vineyards. Clearly, in each case the transaction is only briefly summarised, and the only thing that allows us to understand such minimal notes is their mutual inter-contextualisation as the same people and the same topographical features recur in different transactions. However, since the chronologies differ, what is presented in the codex as a unitary text was split up and thus decontextualized by traditional editorial practice and this is the meaning of the references in blue to the old edition of Ubieto and Ledesma (Ubieto 1976; Ledesma 1989) followed by numbers indicating where these notes were published in different parts of that edition, demonstrating how a unitary text was broken up and scattered. The alternative methodology of simply following the codex structure is not however entirely satisfactory (hence the ascendancy of the chronological methodology), and chronological contextualisation undoubtedly remains important to both historian and philologist. What digital edition allows is for us to resolve this tension, indeed to eliminate the problem, by allowing the user to choose in which mode the material is accessed, and thus order and reorder the material according to their interests.

In our digital edition, the contents of the cartulary can be ordered by folio or with one click reordered chronologically, although, in fact, that

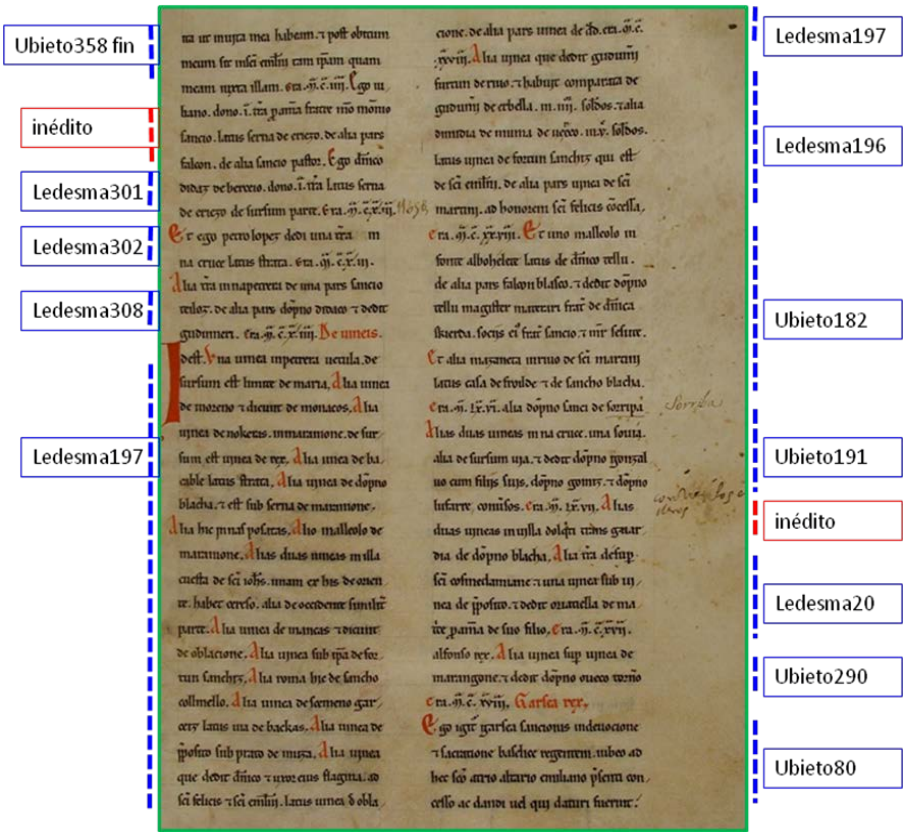


Fig. 2. Extract from *Becerro Galicano Digital* #4 (f. 4r) illustrating the fragmentation of interrelated texts as a result of chronological reordering.

'simple' click was the result of several months work, as the chronology of many of these texts is often problematical. There are also dozens that lack dates and thus inevitably have been decontextualised until now, languishing as an appendix at the end of a traditional chronological sequence. To address these problems, as well as the dates that appear in the codex, we introduced the concept of critical dates – providing precise or approximate chronologies for both undated texts, and for those with problematical dates (anachronisms, scribal error, etc.) – and hence a third way of ordering the cartulary's contents.

3. Development of a lemmatised index

Our digital edition consists of a transcription with accompanying critical apparatus that, for each individual text, provides a regesta, references other extant copies and previous editions, and offers notes addressing questions of chronology and authenticity. The transcription can be visualised with or without notes referring to structure and layout, and the corresponding facsimile images are consultable, either folio by folio or line by line mapping directly to the document's layout on the page. In addition, a comprehensive range of search tools and indices has been developed. Any literal sequence of letters, including wildcards, can be searched for, and there are separate place – & personal – name indices that can be both scrolled through and searched for literal sequences. All such searches can be further restricted by date, document and/or section of the codex, and furthermore the output can be tailored by the user in various ways: altering the size of the text segment that appears in the results in order to achieve either a compact list or fully contextualised extracts; and/or ordering the results by folio or by date.

The most innovative of the tools developed, however, is a lemmatised index of the cartulary's contents, with each word appearing in the cartulary related to a standardised form, i.e. a lemma. For example, *amas* would lead back to *amo*, *servi* to *servus*. This is useful as it allows a particular concept or word to be studied diachronically regardless of the form adopted, overcoming the morphological and orthographic variations and irregularities that undermine traditional indices, and that are even more problematical in such a linguistically complex case as the *Becerro Galicano*, with its mixture of irregular Latin and early Spanish. Thus the philologist can trace the changing form of a given word over time, while the historian can study the chronology and geography of a word's use, irrespective of its form. For example, the lemma *aedifico* (the verb 'build') appears 27 times in the Becerro, assuming 15 different forms in accordance with both Latin morphology and the irregular orthography of our scribes who frequently though not always incorporate a non-etymological initial h- characteristic of early Castilian (e.g., *hedificare*).

Starting from a transcription of the cartulary in a *Word* file, prepared by Fernando García Andreva as his doctoral thesis at the University of the Rioja (García Andreva 2011), the lemmatisation was initially done automatically. All words beginning with capital letters and not immediately preceded by full-stops were excluded – i.e. numerals and proper names –, the lat-

ter set aside to become the basis for the construction of the personal and place-name indices[2]. The remaining almost 10,000 different non-capitalised words, were then tested, using the *Latin Word Study Tool* developed at Tufts University, to see if matching Latin lemmas could be found.

In theory, this process would overcome the problem of the morphological variation in Latin, however in practice the results were disappointing, as a result of the very irregularity of the Latin which was what had made such a tool so desirable in the first place. In quantitative terms, discrete lemmas were found for less than half of the different word forms left after the separation of the named-entities and numerals (Table I). If this is in itself disappointing, we ought to bear in mind that this figure includes a number of false positives, where the lemma found is grammatically plausible but semantically unacceptable, i.e. the codex form coincides with an unrelated Latin lemma. Identifying such cases means scrolling through a list of supposedly lemmatised words looking for erroneous matches, sometimes obvious, but often not, and it is inevitable that behind a superficially satisfying 99,9% of words now lemmatised an unknown number of errors persist. In another 16% of cases the codex form has more than one grammatically plausible lemma, and we had to decide which was the correct one. Finally, we were left with some 3,700 different words to be lemmatised one by one. Often the missing lemma would be obvious, it not having been found the result of the orthographical irregularity of the text. If we go back to the *aedifico* example, we can appreciate not only the usefulness of the tool being developed, but also why the automated detection of the lemmas proved difficult, though in such a case the manual lemmatisation was relatively straightforward. A few isolated forms have proved much more intractable, and indeed some opaque cases are still unresolved, though in practical terms this is not a great problem since the point of the exercise is to produce a tool that relates disparate but cognate forms.

[2] The *Becerro's* total of 197,713 words take 16,291 different forms, of which 6,628 correspond to named-entities & numerals, leaving 9,663 different forms of 'normal' vocabulary.

Tab. 1. Automatic lemmatisation using the Latin Word Study Tool - http://www.perseus.tufts.edu/hopper/morph.

Different forms (excluding numerals and named entities)	9,663	
o Single lemmas	4,444	46%
• [False positives]	????	??
o Multiple lemmas	1,507	16%
o No lemma	3,712	38%

4. Cartography

The final aspect of the digital edition to be discussed has to do with the geographical breadth of the monastery's domains and how we can express this wealth of information cartographically. The prestige of San Millán led to it being massively endowed by monarchs from both sides of the Navarre-Castile border it straddles, and the result was a vast domain ranging over some 120,000 km^2 of northern-central Spain. Consequently, the monastery's cartulary detailing these possessions is exceptionally wide-ranging and contains some 3,000 different place-names.

The identification of many of them is complicated by the parochial nature of many of the texts copied into the cartulary at the end of the twelfth-century which were originally prepared not in the mother house itself but by local scribes, and furthermore by the problem of homonymy, with dozens of different settlements named identically. For example, if we search our place-name index for *Villa nova*, we encounter 17 different and distinguishable places with that name. Conversely, the same place-name can appear under a variety of spellings: for example, the name of the village of *Bolivar* in Álava is recorded with five different spellings in its ten appearances.

One possibility we explored to help in the identification of the different place-names was the use of a digital gazetteer such as the *Nomenclátor Geográfico Básico de España* (http://www.ign.es/ign/layoutIn/actividadesToponimia.do) developed by the the *Instituto Geográfico Nacional.* However, even though it lists almost 800,000 place-names in Spain, it proved of little use. With so many names repeated and with so many of San Millán's possessions being located in tiny settlements that subsequently came to be abandoned the risk of false positives is clear. For example, of our 17 *Villanuevas*, only five

have survived with the same name, and any automated routine to match them to the text necessitates human input to say which *Villanueva* is which.

Where Information Technology has helped us, and enormously, is in the free availability online of cartographical resources previously unheard of. Though frustratingly, these are rather unevenly spread across Spain, conditioned by the fractured and variegated nature of local administration. Nonetheless, we have identified the vast majority of these places, and almost all the settlements as well as hundreds of field-names have been assigned geographical coordinates. Now it is a question of putting to use this ground work, and this is being done on three levels.

We have a gazetteer style place-name index, pointing homonymous occurrences to their different identifications/locations, and linking each place-name directly to the cartography of the *Instituto Geográfico Nacional*. Moreover, a map can be generated for each of the 870 component documents, situating cartographically the significant place-names contained there-in. Finally, from these significant place-names, each with an intrinsic spatial value, a median value has been generated for the document as a whole, which is then inherited by all the words and people that appear in said document. This allows the user to generate maps automatically: for a group of documents, for a given person, or even for a given word or lemma, in short, mapping the results of any user-defined search.

On reflection, I think the technology and methodology employed have served us well to attain our principal objectives of a reorderable and fully searchable edition which is readily accessible. However, it has been a very labour-intensive process, and attempts to use existing digital resources have on occasions proved frustrating, mainly, I believe, because of the idiosyncratic nature of our source material, with its irregular Latin and lost villages. While the edition as it stands has proved a great success with historians and philologists, our target audiences, it remains to be seen to what extent our work can be usefully and economically recycled for other projects.

5. References

García Andreva F. (2011). *El Becerro Galicano de San Millán de la Cogolla. Edición y estudio*, Cilengua.

Ubieto A. (1976). *Cartulario de San Millán de la Cogolla, 759-1076*, Anubar.

Ledesma M. L. (1989). *Cartulario de San Millán de la Cogolla, 1076-1200*, Anubar.

Papers

Digital Cultural Heritage / Patrimonio culturale digitale

ASIt: Atlante Sintattico d'Italia
A linked open data geolinguistic web application*

Giorgio Maria Di Nunzio[1], Jacopo Garzonio[2], Diego Pescarini[3]
[1]Dept. of Information Engineering, University of Padua, Padova, Italy
giorgiomaria.dinunzio@unipd.it
[2]Department of Linguistic and Literary Studies, University of Padua, Padova, Italy
diego.pescarini@unipd.it
[2]Dept. of Comparative Linguistic and Cultural Studies, Ca' Foscari University of Venice, Venice, Italy
garzonio@unive.it

Abstract. Digital Geolinguistic systems encourage collaboration between linguists, historians, archaeologists, ethnographers, as they explore the relationship between language and cultural adaptation and change. These systems can be used as instructional tools, presenting complex data and relationships in a way accessible to all educational levels. However, the heterogeneity of geolinguistic projects has been recognized as a key problem limiting the reusability of linguistic tools and data collections. In this paper, we propose a Linked Open Data (LOD) approach to increasing the level of interoperability of geolinguistic applications and the reuse of the data. We present the ongoing research of the project "Un'inchiesta grammaticale sui dialetti italiani: ricerca sul campo, gestione dei dati, analisi linguistica"; in particular, we present an open source geolinguistic Web application built on top of the Atlante Sintattico d'Italia (ASIt) database.

Keywords: Digital Geolinguistics, Syntactic Databases, Linked Open Data.

1. Introduction

The research field of linguistics studies all aspects of human language, including morphology, syntax and phonology (Akmajian *et al.* 2010). Research in language variation allows linguists to understand the fundamental principles that underlie language systems and grammatical changes in time and space. Geolinguistics is an interdisciplinary field that aims at mapping the geographical distribution of linguistic phenomena which are mainly due

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

to processes of grammatical principled changes (Hoch and Hayes 2010). In this context, the linguistic atlas has proved to be a fundamental tool and product of geolinguistics since the earliest stages of the field; moreover, it has provided a stage for the incorporation of modern GIS. In the last two decades, several large-scale databases of linguistic material of various types have been developed worldwide. The Open Language Archives Community is an example of a world-wide network dedicated to collecting information on language resources and developing standard protocols for interoperability.

The preparation of a linguistics resource of high quality requires several steps: crawling, downloading, cleaning, normalizing, and annotating the data are some of the actions that need to be performed to produce valuable content (Kilgarriff 2007). Some of these steps do require human intervention to achieve the highest quality possible for a resource of usable scientific data. There are three important points about the design and distribution of language resources (Bird, Klein and Loper 2009):

- How do we design a new language resource and ensure that its coverage, balance, and documentation support a wide range of uses?
- When existing data is in the wrong format for some analysis tool, how can we convert it to a suitable format?
- What is a good way to document the existence of a resource we have created so that others can easily find it?

In this paper, we present the ongoing research of the project "Un'inchiesta grammaticale sui dialetti italiani: ricerca sul campo, gestione dei dati, analisi linguistica" (Bando FIRB - Futuro in ricerca 2008); in particular, we present an open source geolinguistic Web application built on top of the ASIt Linked Open Dataset based on the LOD paradigm with the aim of enabling interoperability at a data-level. The LOD paradigm refers to a set of best practices for publishing data on the Web[1] and it is based on a standardized data model, the Resource Description Framework (RDF).

2. The ASIt Enterprise

The ASIt enterprise builds on a long standing tradition of collecting and analyzing linguistic data, which has originated different efforts and projects over the years (Agosti *et al.* 2010, 2011, 2012). Linguistic data stored in

[1] http://www.w3.org/DesignIssues/LinkedData.htm

the ASIt were gathered during a twenty-year-long survey investigating the distribution of several grammatical phenomena across the dialects of Italy (Benincà and Poletto 2007). Research on the syntax of Italo-Romance varieties is of great interest to several important lines of research in linguistics: it allows comparison between closely related varieties (the dialects), formulation of hypotheses about the nature of cross-linguistic parameterization. Furthermore, it allows to single out contact phenomena between Romance and Germanic varieties, in those areas where Germanic dialects are spoken within Italian borders.

At present, there are eight different questionnaires written in Italian and almost 500 translations in more than 240 different dialects, for a total of more than 54,000 sentences and more than 40,000 tags stored in the data resource managed by the ASIt digital library system.

In order to efficiently store and manage the amount of data recorded in the questionnaires, the interviews and the tagged sentences, we have realized curated linguistic database organised into three main conceptual areas containing information about:

- The *geographical area*, which is the place where a given dialect is spoken and where a speaker is born;
- the *speaker area*, which focuses on the background of the speaker: the level of knowledge of the dialect, the particular variety of the dialect, the birthplace, the ancestors, the document that she/he translated;
- the *tagging area*, which is how the document is structured and how it has been tagged (at a sentence level and at a word level).

Fig. 1. A screenshot of the ASIt tagging interface. In this pages, the linguist can split the sentence into words and make spelling corrections. In this figure, the sentence refers to a document written in Cimbrian.

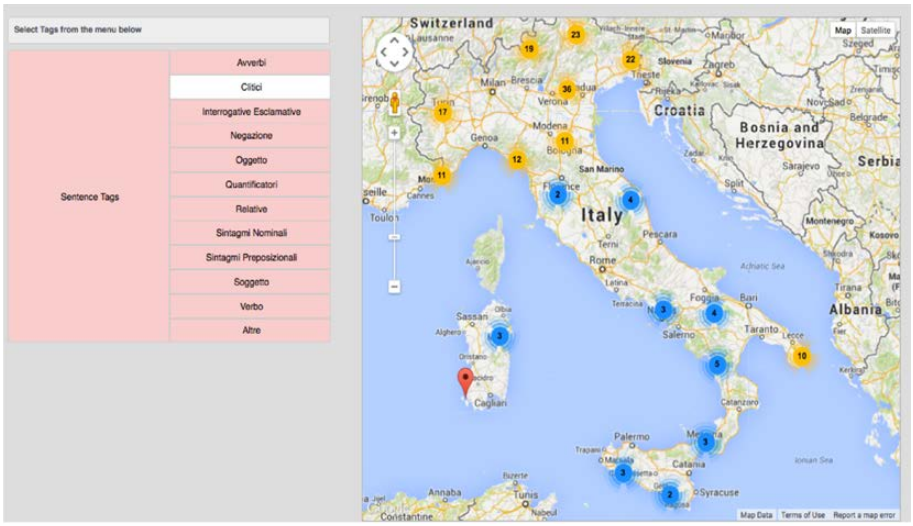


Fig. 2. A screenshot of the ASIr RDF GeoSearch interface.

A relevant aspect of the ASIt curated database is that it explicitly models sentence level tagging, which is not modelled by any other of the presented linguistic projects. Furthermore, we have developed a language-specific set of POS tags in order to link the ASIt data to other databases of dialect syntax. We can therefore imagine the creation of a language-specific tagset including a universal core, shared by all languages, and a language-specific periphery capturing language-specific structures.

3. A Geolinguistic Linked Open Data Web Application

The research direction we want to pursue in this project is to move the focus from the systems handling the linguistic data to the data themselves. For this purpose the LOD paradigm is very promising, because it eases interoperability between different systems by allowing the definition of data-driven models and applications (Heath and Bizer 2011).

The objective of a geolinguistic Web application is to provide linguists with an interactive tool for investigating variations among closely related languages (Di Buccio, Di Nunzio and Silvello 2013a, 2013b). To this purpose, we developed a graphical user interface on the top of the ASIt system

that dynamically produces maps on the basis of the user request.[2] By means of this Web application, linguists are able to annotate sentences and words with tags and perform tag-based searches. In Figure 1, we show a portion of the interface which is used to check spelling and split sentences into words prior tagging. After splitting the sentence, the researcher can start tagging each sentence and each word of a sentence by selecting tags from a predefined hierarchy. Tags of a given type are hierarchically organised in a tag tree. For instance, the root node of sentence tag three is the "Sentence Tags" node; this node has twelve children that correspond to the twelve subsets in which the complete sentence tag set was divided by the linguists. Figure 2 shows the first level of this hierarchy. The same tags used in the annotation phase can be used to search the database and show results on a map. Two different types of search are currently supported. The first search type returns the list of sentences tagged by the requested sentence tags and that satisfy the specified Boolean constraint. The second type of search is performed when the user clicks on the "Search on Map" button. This type of search aims at satisfying the information need of a user searching for the geographic distribution of linguistic resources. Figure 2 shows an example of this type of search. In this case, the user was interested in retrieving all the sentences tagged with "Clitici" and analysing the distribution of this tag on a geographical scale. These results are dynamically populated by performing a query to the SPARQL end point. When the user clicks on the marker corresponding to a specific location, information on the geographical location is displayed. The interface exploits the Leaflet javascript[3] library to visualise the results on a map based on OpenStreetMap data.[4]

4. Conclusions

Digital Geolinguistic systems encourage collaboration between linguists, historians, archaeologists, ethnographers, as they explore the relationship between language and cultural adaptation and change. These systems can be used as instructional tools, presenting complex data and relationships in a way accessible to all educational levels. However, the heterogeneity of

[2] http://purl.org/asit/rdf/search (The Web application is optimised for Firefox browser).
[3] http://leafletjs.com/
[4] http://www.openstreetdata.org/

geolinguistic projects has been recognized as a key problem limiting the reusability of linguistic tools and data collections. In this paper, we propose a LOD approach to increasing the level of interoperability of geolinguistic applications and the reuse of the data. One of the key points of this approach is the decoupling between the system which manages the data and the one which provides services over those data. In fact, we imagine the use of the Geolinguistic Linked Open Dataset by third-party linguistic projects in order to enrich the data and build-up new services over them. In this context, we studied the use case of a linguistic project named ASIt. By exploiting the LOD approach, the ASIt Geolinguistic Linked Open Dataset grows proportionally to the size of the ASIt database. As a concrete example, we presented a geolinguistic Web application build upon the ASIt dataset which provides linguists with a system for investigating variations among closely related languages. Finally, we also developed a graphical user interface on top of this application that dynamically produces maps on the basis of the user requests.

5. Acknowledgments

This work has been supported by the projectt "Un'inchiesta grammaticale sui dialetti italiani: ricerca sul campo, gestione dei dati, analisi linguistica" (Bando FIRB - Futuro in ricerca 2008) and the PROMISE network of excellence (contract n. 258191) project, as part of the 7th Framework Program of the European Commission. The authors want to thank Maristella Agosti, Emanuele Di Buccio and Gianmaria Silvello from the Department of Information Engineering of the University of Padua, Mariachiara Berizzi from the Department of Linguistic and Literary Studies of the University of Padua, and Silvia Rossi from the Department of Comparative Linguistic and Cultural Studies of the Ca' Foscari University of Venice.

6. References

Agosti M. *et al.* (2011). *A Digital Library of Grammatical Resources for European Dialects*. In *Digital Libraries and Archives*, 7th Italian Research Conference, IRCDL 2011 Revised Papers. Communications in Computer and Information Science, vol. 249, Springer-Verlag, pp. 61-74.

Agosti M. *et al.* (2012). *A Curated Database for Linguistic Research: The Test Case of Cimbrian Varieties*. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 2230-2236.

Agosti M. *et al.* (2010). *A Digital Library Effort to Support the Building of Grammatical Resources for Italian Dialects*. In *Digital Libraries,* 6th Italian Research Conference, IRCDL 2010 Revised Selected Papers. Communications in Computer and Information Science, vol. 91, Springer-Verlag, pp. 89-100.

Akmajian A. *et al.* (2010). *Linguistics: An Introduction to Language and Communication*. The MIT Press, Sixth edition, 2010.

Benincà P., Poletto C. (2007). *The ASIS Enterprise: A View on the Construction of a Syntactic Atlas for the Northern Italian Dialects*. In *Nordlyd. Monographic issue on Scandinavian Dialects Syntax*, vol. 34, pp. 35-52.

Bird S., Klein E., Loper E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Di Buccio E., Di Nunzio G.M., Silvello G. (2013a). *An Open Source System Architecture for Digital Gelinguistic Linked Open Data*. In Proceedings of Research and Advanced Technology for Digital Libraries - International Conference on Theory and Practice of Digital Libraries, TPDL 2013 Lecture Notes in Computer Science, vol. 8092, Springer, pp. 438-441.

Di Buccio E., Di Nunzio G.M., Silvello G. (2013b). *A Curated and Evolving Linguistic Linked Dataset*. «Semantic Web», vol. 4, no 3, pp. 265-270.

Heath T., Bizer C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.

Hoch S., Hayes J.J. (2010). *Geolinguistics: The Incorporation of Geographic Information Systems and Science*. «The Geographical Bulletin», vol. 51, no 1, pp. 23-36.

Kilgarriff A. (2007). *Googleology is Bad Science*. «Computational Linguistics», vol. 33, no 1, pp. 147-151.

The "Verbo-Visual Virtual" Platform for Digitizing and Navigating Cultural Heritage Collections*

Alessandro Marchetti[1], Sara Tonelli[1], Rachele Sprugnoli[1-2]
[1]Fondazione Bruno Kessler, Povo (TN), Italy
[2]University of Trento, Povo (TN), Italy
{amarchetti, satonelli, sprugnoli}@fbk.eu

Abstract. Linked collections and archives in the field of contemporary art are usually less common than in other fields of the humanities such as in historical and literary research. The "Verbo-Visual Virtual" (Vvv) project aims at developing a new web portal which unifies the Verbo-Visual collections of two museums, namely the "Archivio Nuova Scrittura" (Ans) owned in part by the Museum of Modern and Contemporary Art of Trento and Rovereto (Mart)[1] and in part by the Museum for Modern and Contemporary Art of Bolzano (Museion)[2]. Through the Vvv web portal, it will be possible for users to search for pieces of art and the corresponding information records now stored in two different museums and accessible using two different catalogues. It will also be possible to discover new information about the collection by performing complex queries over the underlying database and the links manually set among the records.

Keywords: web platform, crowdsourcing, digitization, cultural heritage.

1. "Archivio di Nuova Scrittura" (Ans)

The Archivio di Nuova Scrittura (Ans) is a specialized collection that privileges the connections between writing and image, between art and literature starting from the '60s.

The collection, besides its internationality, is centered around the artworks of Italian artists, and finds its origin in the collecting activity of Paolo Della Grazia, an entrepreneur with a passion for interdisciplinary forms between art and poetry, who started the collection in 1988.

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. Cleup, Padova, 2014.
[1] http://www.mart.trento.it/
[2] http://www.museion.it/

Della Grazia's focus on verbo-visual poetry was influenced by the collaboration with Ugo Carrega, one of the most representative artists of visual poetry in Italy in the '60s. He directed for many years the cultural center "Mercato del Sale", which was open between 1974 and 1990 in Milan and hosted several exhibitions of verbo-visual artworks.

Towards the end of the nineties, Della Grazia decided to donate to a public institution the archive, which had been steadily growing and needed an appropriate site. He decided first to contact Museion, the Museum of Contemporary Art in Bolzano. However, since the museum did not have enough space available, they decided to split the collection into two parts, giving to MUSEION through a long-term loan around 2,000 verbo-visual artworks, and to MART, the Museum of Modern and Contemporary Art in Rovereto, the archive with more than 15,000 volumes. For this reason, a collection which was originally conceived as a single archive is now divided in two and hosted by two different institutions. ANS's fulcrum is represented by works linked to concrete poetry, visual poetry, Fluxus, and conceptual art, to which numerous individual positions have to be added.

The VERBO-VISUAL-VIRTUAL proposal (henceforth VVV) originates from the need to virtually re-create the original version of the archive, at least in digital format.

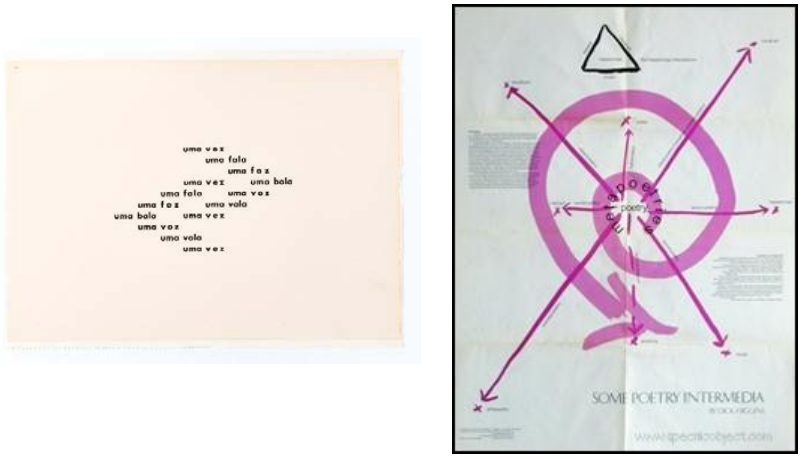


Fig. 1. An example of concrete poetry (left): Augusto de Campos, *Uma Vez*, 1957, and an example of Fluxus (right): Dick Higgins, *Some poetry intermedia*, 1976. Both from MUSEION collection. Further details on the history of ANS can be found in Ferrari 2012.

2. Past related projects: VEKTOR

In the recent years MUSEION participated in the **Vektor project**, which took place from 1999 to 2003 and was sponsored by the 'Culture 2000' programme of the European Commission.

Several European museums archives and institutions were part of the project, sharing their collections or giving their knowledge and competences about digital archives.

The Vektor project aimed at the linking of archives and databases from different institutions through the adoption of standard archive methodologies to develop a centralized interface of access to remote databases.

During the project the records from the different database were linked, and the applicability to contemporary art data of metadata standards such as the Dublin Core and the Getty Museum Thesaurus was evaluated.

As a result of the research on metadata standards, a series of recommendations on cataloguing, and on the structuring of database systems according to standardized data processing were collected in a manual (Reddeker 2006).

During the project, a centralized platform was also developed, which acts as an interface to the database of the partner institutions[3]. The interface allows to query all the database of the museums and archives participating in the project. The platform contains only the metadata of the various artworks, while the complete information about the pieces of art can be accessed through a link to the originating institutions.

MUSEION took part in the Vektor project with its artworks from the "Archivio Nuova Scrittura" (ANS). The museum was able to catalogue and photograph 1200 artworks of ANS, which were then included in the Vektor platform and also made available through MUSEION's online catalogue .

Each item of the ANS collection was catalogued using the elements required by the Vektor platform: *date* of the artwork; *author* of the artwork; *title* of the artwork; *material* of the artwork; *description* of the artwork.

The major challenges that the Vektor project faced were partly the same of the VVV project, as they constitute the basic needs for the creation of a shared catalogue infrastructure.

We will take advantage of the results obtained in the Vektor project reusing the records from ANS already catalogued by MUSEION. However,

[3] http://www.european-art.net/

Vvv will extend the number of artworks in the database and it will also go one step further by integrating different techniques to enrich the existing records (e.g. by art experts, through crowdsourcing). Besides, the exploration platform will provide sophisticated search and visualization functionalities that were not part of Vektor.

In the next paragraphs, we will show in detail how the Vvv will go beyond a traditional digital catalogue like the one proposed by Vektor, augmenting the information of the catalogues and using state-of-the-art natural language processing (Nlp) techniques to allow advanced search functionalities.

3. The Verbo-Visual-Virtual project (Vvv)

Given the possibility offered by current digital technologies to make available art collections to the wide audience, the Verbo-Visual-Virtual project was proposed by Mart, Museion and Fondazione Bruno Kessler, with the goal of implementing an online platform for navigating and querying the Ans collection. The project consortium wanted to virtually re-unify the original Ans and to put in connection the proper artworks, mainly hosted by Museion, with the documental part, currently owned by Mart. In fact, the history of many artworks can be tracked looking at the documents present in the archive, including art magazines where a given work was mentioned, notes by collectors, exhibition reports, etc.

Therefore, the aim of Vvv is an extension and an improvement of the Vektor project, which was exclusively a digitization activity and a meta-catalogue initiative. Within the Verbo-Visual-Virtual project, the analyses performed by two art historians and the connections they manually set among the different parts of Ans will be made available online. Besides, advanced search functionalities aimed not only at *searching* a database but also at *discovering* new connections and information will be implemented. The project was funded by Cassa di Risparmio di Trento e Rovereto. It started in November 2013 and will last 24 months.

Three project phases are foreseen, two of which have already started:

– **Enrichment and interlinking of material inside the Ans collection**: Currently, Mart and Museion collect information on Ans using two differ-

ent data management systems, i.e. MuseumPlus[4] and AdLib[5] respectively. In order to make the unified ANS data available through a web interface, the entries of the two systems must be matched, so as to create a mapping between the same information stored under different naming conventions.

The two systems contain different fields (see screenshot in Fig. 2) and the information stored by each museum is not always consistent. Therefore a selection of the mandatory fields was first performed, and then a mapping between the naming conventions used for such restricted set of fields was established. The partners agreed to make available for each artwork at least the following information: author, title, year, technique, material, dimensions, credit line, link to the originating database and inventory id. Additional information can be added when creating the repository but is not mandatory. Particularly relevant for the development of the final exploration platform is the transcription of words/alphanumeric characters reported in the artwork, if any, and the possibility to add free text comments, which may be converted into new fields at a later stage of the project. Details on this are given in Section 4.

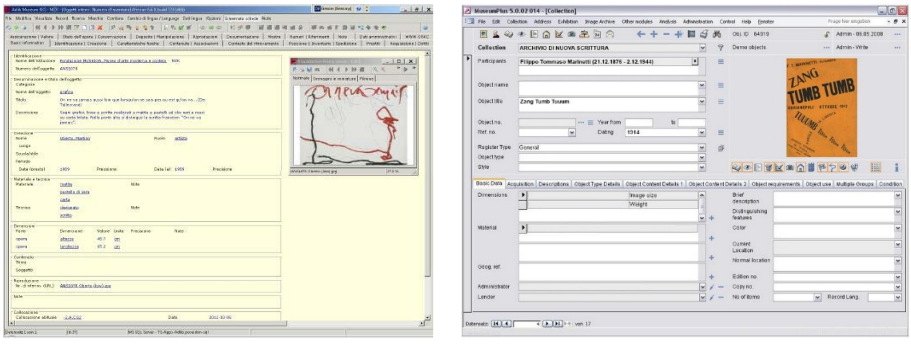


Fig. 2. Screenshots of Ad-Lib (left) used by MUSEION and Museum Plus (right) used by MART.

According to the project plan of activities, the mandatory fields will be made available online through the VVV platform for all 2,000 verbo-visual artworks currently present in the collection. However, additional research and enrichment activities will be carried out for a subset of works in ANS,

[4] http://www.zetcom.com/products/collection-management-software-museumplus/
[5] http://www.adlibsoft.com/

grouped according to some specific criteria (yet to be defined). For these selected entries, two art historians will look for supplementary documental material in the section of the ANS archive hosted at MART, in order to explicitly link an artwork with all documents referring to it or witnessing its history. Due to time constraints, this enrichment will not cover all artworks, but it will represent a first feasibility study for future extensions.

Another issue related to the preparation of the repository is the *choice of the tool* to be used by the art historians to encode such additional information, which will later be stored in the database behind the exploration platform. A preliminary analysis of the requirements for data collection highlighted that art historians needed to have a tool to store and link documents in different formats (not only documents but also pictures and audio/video files) that could be accessed through the Internet from different locations, given that their activity can be carried out in different archives. Besides, easy-to-use export functionalities should be available to minimize the effort to convert and store the data in the final database. In the light of these requirements, FilemakerServer was adopted[6].

– Implementation of the exploration platform: This second activity has started right after the partners have agreed on the skeleton of the data to be stored. This phase includes a set of technical steps for the creation of a database containing all information stored on the verbo-visual collection, and the implementation of a web-based platform for navigating such data. Some functionalities will be available only to expert users through authentication. This means that the system administrator will provide trusted experts (for instance, the two art historians enriching the data) with access credentials. These additional functionalities will include *i)* the possibility to annotate the data by adding notes or tags, and *ii)* the option to mark records as "verified". This feature was added because the final platform will display all artworks exported from the museums' data management systems, but only few of them will be manually checked and enriched with additional information. Although the information is likely to be mostly correct, there may be inconsistencies. The "verified" tag will ensure that the displayed information was judged as correct by an expert.

All users without "expert" credentials will be able in any case to access and navigate through the collection. *Basic* search functionalities will be offered based on the available metadata (e.g. search by author's name, date,

[6] http://www.filemaker.com/it/products/filemaker-server/

artwork title, etc.). Other functionalities will be implemented depending on the additional information entered for each work of art. For instance, if some works are transcribed, it will be possible to search by concept, i.e. the system will automatically recognize synonymous words in the transcriptions and cluster the corresponding occurrences. It will also be possible to retrieve the entries that are most similar to a given one, based on the number of matches in the metadata and – possibly – in the transcribed text. If information is annotated e.g. on the colour of the work, this feature will also be included in the search and be taken into account in the similarity-based search. The goal of the search interface is not only to retrieve artworks, but also to discover connections that were not explicitly stated in the collection, involving artworks, artists, temporal and spatial relations, etc. Specific software libraries for visualizing the search outcome will make the output very intuitive and easy to understand.

The VVV platform will take advantage of the experience gained by the Digital Humanities group at Fondazione Bruno Kessler in the development of the A.L.C.I.D.E. (Analysis of Language and Content In a Digital Environment) system[7] for exploring historical documents and combining different search functionalities. A particular emphasis will be put on the implementation of an intuitive, easy-to-use interface, taking into account past research in human-computer interaction and system usability (Caviglia *et al.* 2012).

– **Evaluation of the platform**: The platform will be by evaluated by inexperienced and expert users in a real context based on a user-centered design approach (Norman 2002). Expert users will be chosen by MART and MUSEION, while inexperienced users will be selected among non-expert volunteers such as visitors of the two museums or students. Evaluation will include collecting feedback from the users through specific questionnaires and focus groups.

After the various cycles of test and revision of the platform, the interface will then be available online and accessible to the public. A final exhibition is planned in the two museums to present the outcome of the project at the end of the 24 months.

[7] http://dh.fbk.eu/projects/alcide-analysis-language-and-content-digital-environment

4. Research challenges

Recently, crowdsourcing has emerged in many domains as a popular paradigm used to accomplish a variety of tasks such as collect or classify data, correct content and raise funds to sustain an initiative. A comprehensive definition of crowdsourcing is proposed in Estellés-Arolas and González-Ladrón-de-Guevara (2012): "Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task."

In the last few years, several crowdsourcing projects have been undertaken by galleries, libraries, archives and museums (see Carletti *et al.* 2013 for a survey). Moreover, a growing number of studies have analyzed opportunities and challenges given by engaging the wisdom of the crowd in the cultural heritage domain (see Oomen and Aroyo 2012 and Ridge 2013 among others).

Within the VVV project, we will exploit crowdsourcing methodologies to enrich institutional metadata (i.e. the five mandatory fields for each artwork), adding value to the existing resources. The output of crowdsourcing tasks will be integrated into the search functionalities so to incrementally improve navigation of the collection. Besides, this will allow the consortium to perform cutting-edge research in the field of digital humanities, which is one of the objectives of the VVV grant. In particular, we will use the collected data to perform different experiments, for instance to do similarity-based clustering of the artworks using visual and textual features, to perform authorship attribution, to assign the artworks to different categories (e.g. temporal span) in a supervised classification setting, or to study the correlation between given features and the emotion evoked by the artwork.

At the moment, two tasks have been planned, namely the transcription of the textual content contained in the artworks and the tagging of emotions. Figure 3 shows the mock-ups of web interfaces for both tasks with the CrowdFlower platform[8].

It is estimated that 50% of artworks in the ANS collection contains textual content: transcribing this content will give us the possibility to apply text analysis techniques to transcriptions. The use of automatic Natural Lan-

[8] http://www.crowdflower.com/

guage Processing tools will allow us to normalize word forms (e.g. plurals, inflected verbs) and to recognize and consequently cluster similar words and concepts. Users will thus be able to perform a detailed search on the collection: for example, the search word "vento/wind" will not only retrieve the artworks that contain the exact occurrence of that word (e.g. "Spleen 2" by Ugo Carrega) but also similar words such as "brezza/breeze" (e.g. "Spleen 1" by Ugo Carrega). Similarly, searching for "pensiero/thought" will retrieve all artworks containing the word both in singular and plural form (e.g. "Lirica Logica" and "Oltre la Materia" by Ugo Carrega).

As for emotion annotation, we refer to Ekman's model of human basic emotions: i.e. anger, disgust, fear, happiness, sadness and surprise (Ekman 1972). The output of this annotation will be useful for analyzing the emotional impact of verbo-visual artworks (Yanulevskaya *et al.* 2012; Bertola and Patti 2013) and for enriching the search functionalities of the platform: in particular, these emotions will be used as keywords to search for artworks and navigate the collection.

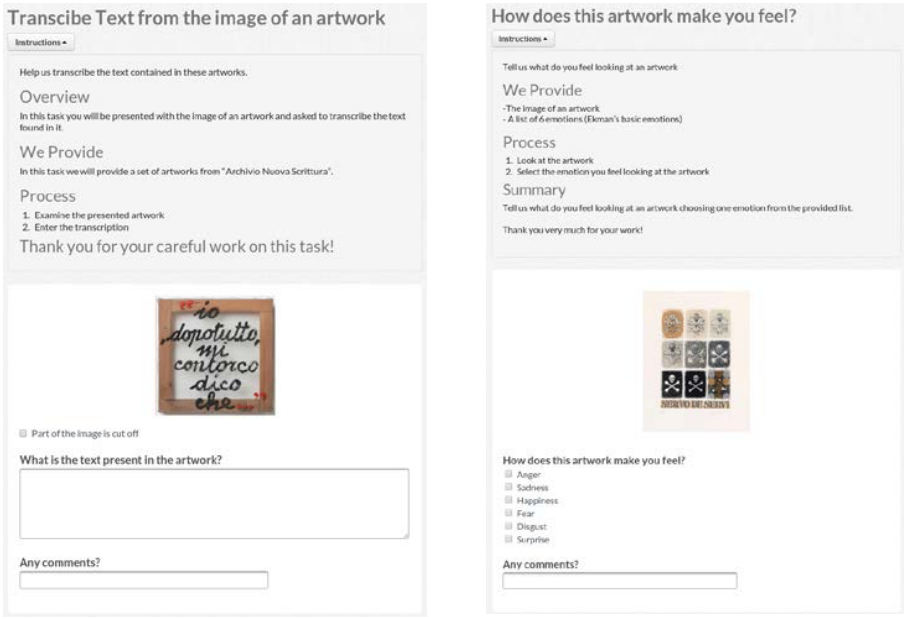


Fig. 3. Mock up of the transcription (left) and emotion annotation (right) interfaces.

The challenges to be addressed in order to maximise the use of crowdsourcing and successfully improve the outcome of this initiative are mainly two and strongly interrelated. The first issue is about how to assure the high quality of the content gathered from the crowd and integrate it with institutional information provided by museums. The second challenge is related to how to engage the public. This latter aspect is connected to the former one: promote the active engagement of users can substantially improve the quality of collected data because motivated contributors tend to be more reliable. For this reason we are exploring the possibility of engage users through serious games following the "Games With A Purpose" paradigm proposed in Von Ahn 2006. The aim is to rely on a combination of social motivations (e.g. will of helping museums, love for contemporary art), fun and competition to foster user participation.

5. Expected project outcome

The main outcome of the VVV web portal will be the promotion of verbo-visual art to a large number of users and researchers. The new catalogue will make possible an in-depth analysis of verbo-visual art, investigating its diffusion, the network of the main artists involved in the movement, and the complex interrelations between verbal and visual art.

Another important outcome of the project is the enhancement of the visibility of the two museums in the project (MART and MUSEION). These institutions will be involved not only as owners and curators of the collections, but also as providers of metadata information, which will be shared through the Europeana network[9] following the Linked Open Data in the Europeana Data Model (EDM)[10].

The mapping between the data of the museums and the EDM format will be performed using tools such as KARMA[11] and taking advantage of insights from other linked open data conversions in the cultural heritage domain (Szekely *et al.* 2013).

[9] http://pro.europeana.eu/web/europeana-project/home
[10] http://pro.europeana.eu/edm-documentation
[11] http://www.isi.edu/integration/karma/

The compliance to EDM will guarantee that the standard used in the project is shared by many other museums and archives such as the Rijksmuseum of Amsterdam, the British Library, the Prague National Museum, etc. The adoption of internationally recognized standards will also enlarge the audience of scholars and facilitate the connection between researchers interested in verbo-visual art.

After being tested on the ANS collection, the exploration platform and the methodology for gathering and converting data could be extended to other collections owned by the two museums. From a methodological perspective, this project will promote good practices for the extraction of metadata from art collections, their enrichment and interlinking using different approaches (experts, crowdsourcing) and their integration in the Europeana network.

6. Novelty of the project

This project presents several features of novelty with respect to the object of study, the methodologies used and the interaction between the partners of the project.

The verbo-visual artworks that will be included in the project repository are part of a collection which has been only recently studied by art historians, since the cataloguing is still in progress. The project will enhance the availability of these works and create a virtual collection with homogeneous information.

Another novelty is represented by the collaboration between researchers with different backgrounds and the subsequent multidisciplinary effort. The cooperation between experts in the area of information technology and experts in the field of history of art is an essential part of the project. This collaboration will allow the two museums to participate in the Europeana network, to get confident with semantic web issues and to get involved in Linked open data initiatives (Haslhofer and Isaac 2011) aimed at the diffusion of metadata.

For both MART and MUSEION, the VVVweb portal will also be the chance to extend their activity beyond the realization of paper catalogues and exhibitions and to contribute to the sharing of knowledge about verbo-visual art.

From a research point of view, the characteristics of the data stored in the VVV database (i.e. quite a large amount of homogeneous data enriched with

diverse, high-quality information) make it an ideal data set for classification experiments exploring the visual and textual features of the artworks.

7. Conclusions

In this paper, we presented an ongoing interdisciplinary effort aimed at the virtual unification of two existing sections of the ANS archive through the creation of a web platform.

We detailed the workflow to implement the system and we discussed the challenges to be addressed during the project to match existing standards and be able to release the artworks' metadata. We also presented the research activities foreseen to transcribe the verbal content of the artworks and to tag them with emotion labels, using crowdsourcing techniques.

After the first six months of activity, the project has already achieved some intermediate results, i.e. the definition of the data fields to be imported in the database and of the mapping between the data management systems of MART and MUSEION, the choice of the tool and the procedure to be used by the art historians to enrich such data, and the definition of the database structure underlying the final navigation system. In the near future the database will be populated with the ANS dumps from the museums' data management systems and the first tests will be run to check the information consistency.

8. Acknowledgements

Images of artworks in Figures 1 and 3 are reproduced by permission of Fondazione MUSEION, Museo d'arte moderna e contemporanea Bolzano, Collezione Archivio di Nuova Scrittura.

9. References

Bertola F., Patti V. (2013*). Emotional Responses to Artworks in Online Collections*. In Proceedings of PATCH.

Carletti L., MCauley D., Price D., Giannachi G. (2013). *Digital Humanities and Crowdsourcing: an Exploration*. In Proceedings of Museum and the Web 2013, Portland (Oregon, USA), pp. 223-236.

Caviglia G., Ciuccarelli P., Coleman N. (2012). *Communication Design and the Digital Humanities. Visualizations and Interfaces for Humanities Research*. In Proceedings of the 4th International Forum of Design as a Process.

Ekman P. (1971). *Universals and cultural differences in facial expressions of emotion*. In *Nebraska symposium on motivation*, University of Nebraska Press.

Estellés-Arolas E., González-Ladrón-de-Guevara F. (2012). *Towards an integrated crowdsourcing definition.* «Journal of Information science», 38, 2, pp. 189-200.

Ferrari D. (2012), *Archivio di Nuova Scrittura Paolo della Grazia. Storia di una Collezione / Geschichte einer Sammlun*g, Silvana Editoriale.

Haslhofer B., Isaac A. (2011). *data. europeana. eu: The Europeana Linked Open Data Pilot*. In International Conference on Dublin Core and Metadata Applications, pp. 94-104.

Norman D. A. (2002). *The design of everyday things*, Basic books.

Oomen J., Aroyo L. (2011). *Crowdsourcing in the cultural heritage domain: opportunities and challenges.* In Proceedings of the 5th International Conference on Communities and Technologies, ACM, pp. 138-149.

Oomen J., Belice Baltussen L., Limonard S., van Ees A., Brinkerink M., Aroyo L., Vervaart J., Afsar K., Gligorov R. (2010). *Emerging practices in the cultural heritage domain-social tagging of audiovisual heritage*. In Proceedings of WebSci10: Extending the Frontiers of Society On-Line, Web Science Trust, 2010.

Reddeker L., ed. (2006). *Archiving The Present. Manual on Cataloguing Modern and Contemporary Art in Archives and Databases*, Wien.

Ridge M. (2013). *From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing*. Curator: «The Museum Journal», 56, 4, pp. 435-450.

Szekely P., Knoblock C.A., Yang F., Zhu X., Fink E.E., Allen R., Goodlander G. (2013). *Connecting the Smithsonian American Art Museum to the Linked Data Cloud*. In *The Semantic Web: Semantics and Big Data*, Springer, pp. 593-607.

Von Ahn L. (2006). *Games with a purpose*, «Computer», 39, 6, pp. 92-94.

Yanulevskaya V., Uijlings J., Bruni E., Sartori A., Zamboni E., Bacci F., Melcher D., Sebe N. (2012). *In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings*. In Proceedings of the 20th ACM international conference on Multimedia, ACM, pp. 349-358.

Dante. A Web Application for the History of Art*

Chiara Ponchia
Department of Cultural Heritage, University of Padua, Padova, Italy
ponchiachiara1@gmail.com

Abstract. The paper presents a web-application that was developed by a multidisciplinary team of IT researchers and art historians at the University of Padua to study illuminated scientific manuscripts, and that was recently re-used to support a PhD research study on the first illuminated manuscripts of the *Divine Comedy*. The experience proved to be successful on both sides: it was a valuable occasion to verify the web-application usefulness with a different collection, and the web-application tools effectively helped in managing and studying the illuminations selected for the PhD research. The paper focuses on the creation of the manuscripts and illuminations catalogue metadata; particularly the second process requires an in-depth image analysis that can bring about new reflections on the image itself and finally can help researchers to find out and establish new connections among images.

Keywords: information technology tools, history of illumination, iconography.

1. Introduction

Information technology (IT) tools can support humanities research in many ways. These include helping researchers to deal with a large amount of data, supporting the retrieval and organization of information, and keeping track of the work done.

In the History of Art field, researchers need to manage a large number of images, especially for the purposes of comparison: in fact, comparison lies at the heart of art-historical research, because it is through comparison that scholars can disclose new knowledge. For example, comparing two pictures realized in different moments by the same painter can help to trace his career, or, if we compare his pictures with contemporary works by other artists, we can find out if the painter was influenced by other art

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

movements or not. These are just a few examples of how art-historians work with images.

This paper presents the result of a long joint research project between IT researchers and art historians that led to the creation of a web-application to foster and support research in History of Art. The research has particularly helped professional users to manage a great number of images, to compare them and to find out new connections between them, thus creating new knowledge on different subjects.

2. IPSA

IPSA (*Imaginum Patavinae Scientiae Archivum*)[1] is a digital archive which gathers illuminated scientific manuscripts produced mainly, but not only, in Italy and in the Veneto region during the 14th and 15th centuries. It was created specifically for professional researchers in History of Art and History of Illumination to allow them to compare the illuminations held in the digital collection and to verify the development and the spread of a new scientific frame of mind in the 14th century at the University of Padua and a new realistic way of painting. IPSA is not only a digital archive, but also a web-application that enables users to work with images in different ways; above all, they can compare images and create links between them, specifying the kind of link they have identified. In fact, they can choose from a drop-down menu among five different labels:

- "Copied in", if the subject of the older image is quite faithfully re-proposed in the newer image;
- "Not related to", if the two illuminations show subjects belonging to different iconographic traditions;
- "Same tradition of", if the two illuminations show subjects belonging to the same iconographic tradition; this kind of relation is valid both for images markedly distant in time and for images closer in time;
- "Siblings" if the two illuminations were copied from the same model;
- "Similar to" if the two illuminations show some analogies, but it is not possible to further specify the kind of relation existing between them.

In addition to this, users can annotate the links they create, and share their annotations with other users (Agosti 2003, 253-264). These function-

[1] http://www.ipsa-project.org/; informative site http://ipsa.dei.unipd.it/

alities are the result of a multiyear close collaboration between IT researchers and art-historians: firstly, IPSA was realized in 2001 at the University of Padua in the context of the project *Da Padova all'Europa. Per un corpus informatizzato dell'illustrazione scientifica nello Studio padovano dal Medioevo al Rinascimento: astronomia-astrologia, botanica, medicina*, and secondly was refined three years later in the context of another project that also involved the University of Naples Federico II and the Second University of Naples. In the last three years, in the context of the European project CULTURA[2], IPSA was further improved and the collection was increased. Thanks to this long-standing interdisciplinary cooperation, IPSA was tailored on professional users requirements and as a result it perfectly addresses very important research needs of art-historians, so that its precious tools appear to be useful not only for the current collection of scientific manuscripts, but they can be effectively used for different types of collections as well.

Thanks to a new collaboration between the same IT researchers that created IPSA and researchers in History of Art, it was decided to use IPSA to support part of a PhD thesis in History of Illumination, the topic of which is the illustration of the *Divine Comedy* in Northern Italy during the XIV century (Ponchia 2014). It was considered a very valuable occasion to prove the usefulness of IPSA with another collection and to create a fruitful synergy between different disciplines.

3. Dante

The aim of the PhD research is to study the most important illuminated manuscripts of the *Divine Comedy* produced in Northern Italy in the XIV century, especially the first manuscripts realized after Dante's death (1321), when the poem started to become known. While in Florence an almost serial production of illuminated manuscripts took place, with very simple and schematic illuminations only in the initial pages of the three *canticas*, in Northern Italy the production was numerically inferior, but some of the manuscripts from this area have a stunning number of beautiful images. Every one of these books is a unique masterpiece that presents different solutions to visualize Dante's pilgrimage in the Hereafter.

[2] http://www.cultura-strep.eu/

One of the key-issues of the PhD research was to find and highlight differences and analogies between *Divine Comedy* illuminations in Northern Italian manuscripts, and try to find out what influenced the choices made by illuminators when they had to face the difficult challenge of illustrating a text with no previous iconographic tradition.

To fulfil this goal it was necessary to compare a large amount of images, and in this respect using a tool such as IPSA was crucial, as it can effectively support researchers in managing great quantities of digital reproductions. Hence a new instance of IPSA, called *Dante*[3], was created. Working with *Dante* immediately proved to be useful and operational, thanks to the previous experience with IPSA.

The first step was to create the catalogue metadata of each manuscript chosen for the research. To do this, the user must select the function "Works-Add" in the upper part of the page. Then the user is presented with a form containing different fields, where they can insert the main information about the manuscript, thus creating an accurate codicological description (fig. 1).

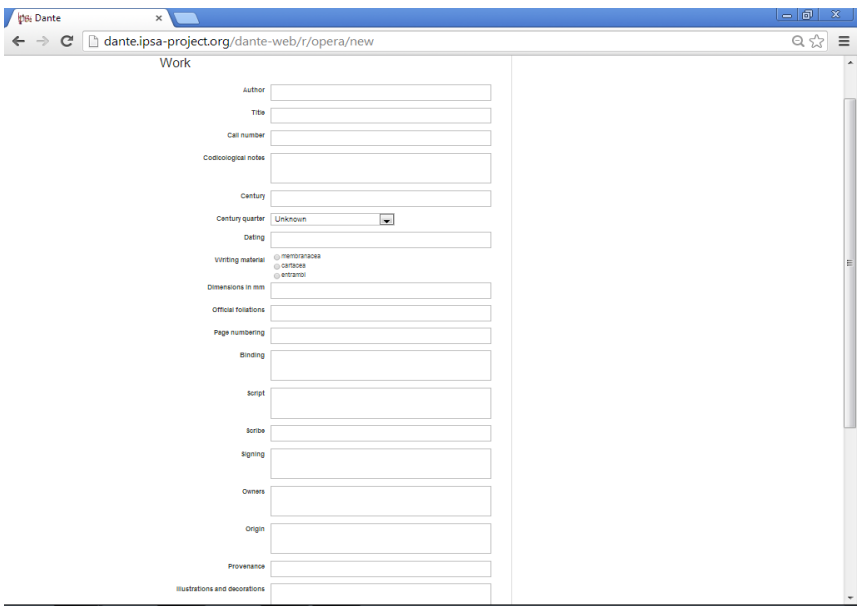


Fig. 1. Manuscript catalogue metadata form.

[3] http://dante.ipsa-project.org/dante-web/

Because a standard codicological description for illuminated manuscripts does not exist yet, the form in *Dante* gathers all the fields that can be found in the current main kinds of codicological description, aiming at completeness. Hence, the user will insert not only basic information such as "Author" and "Title" or concerning images, such as "Illustrations and decorations" and "Iconographic tradition", but they will also be able to report more details about the physical aspect of the manuscript, e.g. "Writing material" and "Dimensions in mm". The result is a page that presents all the information inserted by the user in the upper part, while in the lower part there is a button marked "Add a new illustration" that allows the upload of digital reproductions of the illuminated pages in that manuscript. All the uploaded images appear as a wall of thumbnails in the lower part of the manuscript catalogue metadata; clicking on a thumbnail, the user can access a page with the selected illumination and its catalogue metadata, a button that allows the image to be zoomed, and a button that allows the original image to be downloaded.

To create the illumination catalogue metadata, the user is presented with a form similar to the manuscript form, where they can insert information such as: "Subject", "Illuminator", etc (fig. 2).

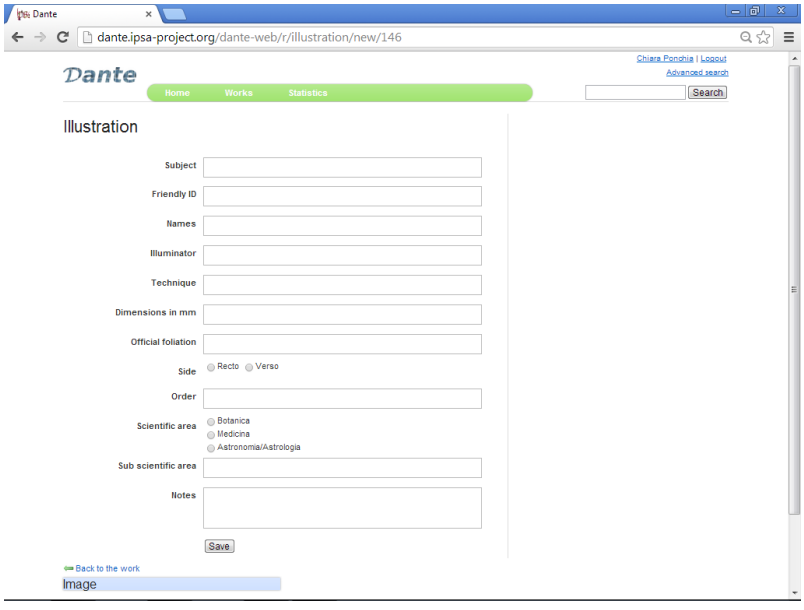


Fig. 2. Illumination catalogue metadata form.

Also for illuminated images cataloguing a standard form does not exist yet, so the form in *Dante* aims at collecting all the fundamental data about illuminations, especially concerning their iconography. Particular attention must be given to the field "Names", as it is pivotal for all the subsequent research that will be done in the digital archive. While in the field "Subject" the title of the illumination must be reported, e.g. *The encounter with Minos*, in the field "Names" the user should insert the names of all the characters represented in the illumination, in this case Minos, Dante, and Virgil. This is a key-passage because if the user omits to insert all the names, important elements can be missed in the advanced research. For example, if the user describes *The encounter with Paolo and Francesca*, an illumination referred to in the second part of the 5th *canto*, they should not only report the names of Dante, Virgil, Paolo and Francesca, but also the cardinal sin the two lovers belong to, i.e. lust. If the user fails to pay attention to this, a search with the term "Lust" will show just illuminations referring to the first part of the 5th *canto*, while it would be correct to see all the illuminations, both those referring to the first part of the *canto* and those referring to the second part. Creating the image catalogue metadata clearly requires a comprehensive overview of the digital archive and of the possible searches users could be interested in doing. In any case, the catalogue metadata, both of manuscripts and illuminations, are editable, so it is possible to add details that may be needed in the future.

Once the creation of all the manuscript catalogue metadata and illumination catalogue metadata was completed, it was possible to easily analyze and compare a large amount of images, finding differences between the illuminations held in the digital archive and new research paths. A good example is the study of Charon. By inserting his name in the web-application search engine, the user will see as a result all the representations of this character in the manuscripts selected for the PhD research (fig. 3).

The value of this must be underlined, because at a glance the user can see illuminations that are held in different parts of Europe and that otherwise would be impossible to bring together in the same place. In fact, because of their small size, people at the time could easily bring manuscripts with them while travelling; manuscripts could be given as a present to important people, like bishops or princes, or they could be stolen because of their value. As a result, few of the medieval manuscripts are still preserved in the same place they were produced: more often, they are held in countries other than their original one. In the case of Italian *Divine Comedy* manuscripts,

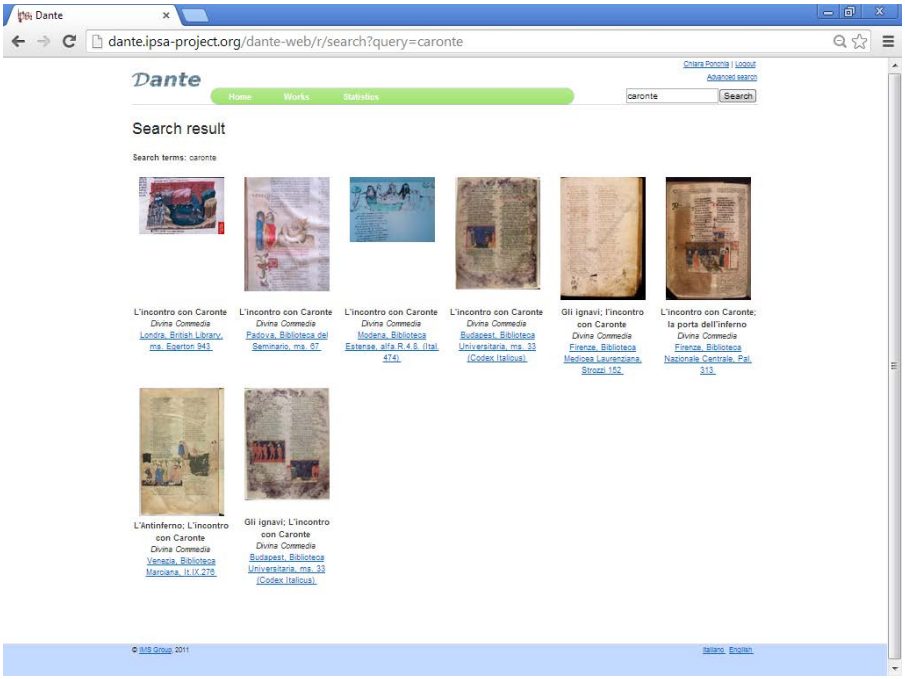


Fig. 3. "Caronte" search results.

today they are held in many different European libraries, such as the British Library in London or the Bibliothéque Nationale de France in Paris. The beauty of *Dante* is that by simply looking at a computer screen it is possible to compare all the illuminations needed and point out differences and similarities. Such an instrument is evidently a major help for art-historical research. Furthermore, the image search procedure allows the user the certainty of retrieving all the illuminations containing the elements they desire to study: doing the research manually would be much more difficult and would increase the possibility of missing something. Moreover, it is possible to link images that are of particular interest and annotate them, so the researcher can keep track of the connections they discover and takes note of the main passages of their research process. In the case of Charon, only two illuminations respect Dante's words, which describe Charon as an old man with white hair, while all the other illuminators painted Charon as a medieval devil. This suggests the idea that just a few illuminators were interested in

being faithful to Dante's poetry, while other artists preferred to follow more common sources, such as devil representations in frescos or in the sculpted portals of the gothic cathedrals. Therefore, the user may want to link the two correct illuminations and highlight their accuracy (fig. 4).

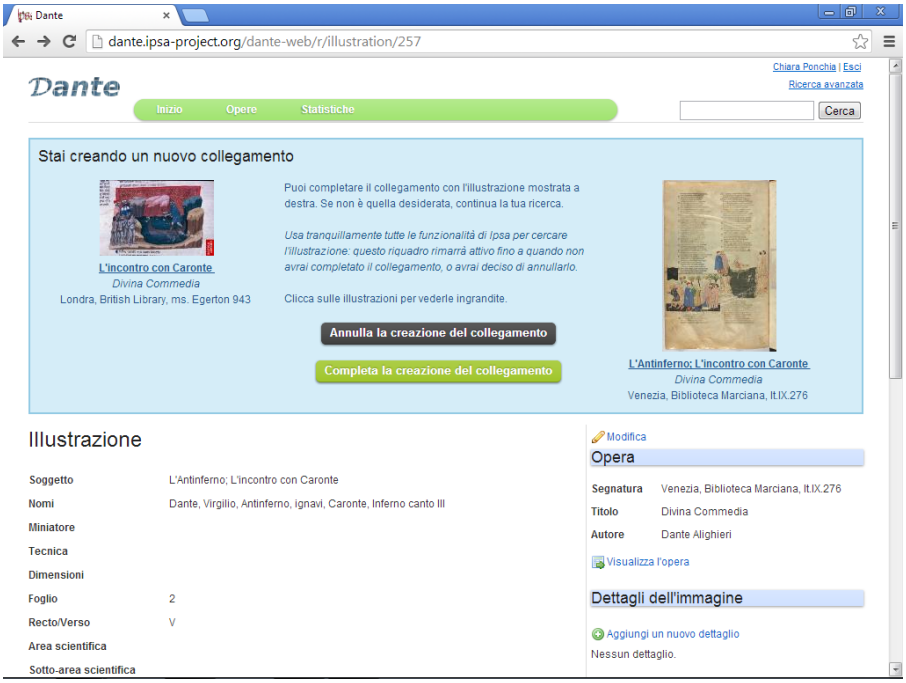


Fig. 4. Linking two images.

Creating a link between two images is a very intuitive process that was realized during the elaboration of Ipsa to address art-historians requirements, and was refined thanks to the evaluations carried out in the following years (Agosti 2012, 89-94). By clicking on the thumbnail of one of the two illuminations, the user enters its catalogue metadata where they find the button "Start a new link". After clicking on this button, the user can start a new search to find the second image they want to link the first image to. During this process, a box at the top of the page shows the status of the operation, including the thumbnail of the starting image and a short text explaining how to complete the process or cancel it. Also at this stage it is possible to

zoom in on the image, if the user needs to better analyze some details. When the user finds the second image, they can select it simply by clicking on its thumbnail. Afterwards, the thumbnail of the second image will appear in the box that is showing the status of the process, and the user will be able to see the two images together, compare them and then decide whether to complete the link creation or cancel it. Clicking on the button "Complete the link creation", the user will be showed a page with the selected images, the older on the left, and between them a free-text field where they can annotate the link with their thoughts and impressions. Also a drop-down menu is showed, where the user can choose the type of link existing between the two images: "Copied to", "Unrelated to", "Same tradition of", "Sibling of", "Similar to". As a result, the thumbnail of the image it has been linked to will appear in the catalogue metadata of each of the two illuminations, with the call number of the manuscript the image belongs to and the annotations of the user. Hitherto it is possible to annotate only the link between two images; the option of annotating a single image is currently being considered and will possibly be included in the next version of *Dante*.

4. Conclusions and future work

The usefulness of such an instrument is clear, as it enables researchers in History of Art and History of Illumination to easily browse a collection of images, retrieve the needed digital reproductions, analyze them and compare them.

In the future, *Dante* could be fruitfully used also for other purposes, especially as a collaborative research environment where professional users, not only art-historians but experts in codicology, philology and Italian literature as well, can share their opinions through annotations. It would be a valuable tool for the international research community, as people could study manuscripts held in different parts of Europe without travelling, and could discuss different issues related to the *Dante* collection without needing to meet in person.

Dante could also be used as a teaching tool in different disciplines related to manuscripts and their history. In the case of the History of Illumination field, *Dante*'s collection would be perfectly suitable for showing students some of the different trends of Northern Italy illumination during the XIV century. In fact the collection gathers illuminated manuscripts from impor-

tant cities of Northern Italian regions, such as Venice, Padua and Bologna, and can display the peculiarities of their artistic productions and the way illuminators influenced each other. In addition, the variability of the manuscripts and the illuminations held in the collection would offer students a useful overview of the many solutions chosen by calligraphers and illuminators concerning the page lay-out, the relation between text and images, image style and iconography, and so on.

Another valuable application would be to ask students to carry out small tasks in *Dante*, like analyzing the style of an illumination or setting new links between images; to this purpose, annotations left by scholars would serve as a guide for students, increasing the teaching potential of this web application.

5. Acknowledgements

The work reported has been partially supported by the CULTURA project as part of the Seventh Framework Programme of the European Commission, Area "Digital Libraries and Digital Preservation" (ICT-2009.4.1), grant agreement No 269973.

6. References

Agosti M., Benfante L., Orio N. (2003). *IPSA: A digital archive of herbals to support scientific research*. In Sembock T.M.T., Zaman H.B., Chen H., Urs S.R., Myaeng. S.M., eds., *Digital Libraries: Technology and Management of Indigenous Knowledge*, 6th International Conference on Asian Digital Libraries, ICADL 2003, Lectures Notes in Computer Science (LNCS) 2911, Springer, 2003, pp. 253-264.

Agosti M., Benfante L., Manfioletti M., Orio N., Ponchia C. (2012). *Issues to Be Addressed for Transforming a Digital Library Application for Experts into one for Final Users*. In Ioannides M., Fritsch D., Leissner J., David R., Remondino F., Caffo R., eds., *Progress in Cultural Heritage Preservation*, Euromed 2012, 4th International Conference, Short Papers, Multi-Science Publishing Co Ltd, pp. 89-94.

Ponchia C. (2014). *Frammenti dell'Aldilà. Immagini nella* Commedia *nell'Italia settentrionale del Trecento*, PhD thesis, supervisor prof. Toniolo F., Cultural Heritage Department, University of Padua, PhD School in History and Critique of Visual Arts, Music and Performing Arts, XXVI cycle.

Digital Lightbox: a web-based visualization framework applied to paleographical research*

Giancarlo Buomprisco
King's College London, London, United Kingdom
giancarlo.1.buomprisco@kcl.ac.uk

Abstract. Previous digital approaches to Paleography have mostly been interested in the automatic classification and recognition of paleographic samples using quantitative and objective features. The tool proposed is called Digital Lightbox, a web visualization framework to support the work of paleographers in analyzing and studying paleographical elements. Contrary to most of the previous studies in the field of digital paleography, the goal of Digital Lightbox is to support for scholars qualitative interpretations, a task which has had much less interest in the field of the digital humanities when concerning digitized images. An important factor of the project is the particular collaboration-oriented approach, which aims to aid the collaborative work among researchers by importing and exporting working sessions, and an easy reuse of XML catalogues in order to load lists of images in the application.

Keywords: digital paleography, virtual lightbox, visualization framework, images processing.

1. Introduction

During the last years, the development of web technologies has brought new possibilities in creating computing applications of any kind. This sort of resources is of particular importance for the Digital Humanities world, since the improvement of these applications has been crucial for two important factors of the field: cooperation and sharing.

The advantages deriving from collaborative work and knowledge sharing are factors of primary importance for a discipline which, in the last years, has exponentially increased the number of employees and professional figures, as well as study programs and research labs with their own practical and theoretical methodologies, which can take advantage of the collaboration from the experience and the results of multiple research projects.

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

Speaking instead of Digital Paleography, considered a sub-set of the Digital Humanities, where while the projects developed in the past have tried to elaborate and process quantitative and objective results to problems related to the classification and the description of paleographical elements, those who instead have tried to back qualitative analysis and to support researchers' interpretations have been few, put apart and often abandoned. This is the case of the so called "Virtual Lightboxes", which due to not mature technologies, have never been able to express an important potential for the study of digital manuscripts. Nonetheless, today more than ever before, this kind of tools could be necessary for managing all the data and the digital cultural heritage which academies, but also public institutions, have been developing during the last 15 years. Indeed, the digitization of images is a ongoing process which is thought to be soaring in the next years as a result of the efforts of government institutions in order to digitize national cultural heritage (Niutta 2010), but which has recently also involved private enterprises such as Google, which is currently working on the *Dead Sea Scrolls* project.[1]

The world wide web has brought us the possibility to access to any kind of data instantly: as already described above, this fact has a particular meaning about manuscripts because the difficulties to access to manuscripts. Now it is possible to freely access a large number of digital manuscripts from every place and at every time. Various tools have been developed by libraries and universities: because digitized manuscripts often have a high resolution with the possibility of zooming images, also on a modest hardware, researchers have had the possibility to make revolutionary software to aid paleographers in manuscripts analysis. This is one of the reasons why Digital Lightbox has been created: exploit all the images available through the WWW with an images management system which could improve scholars research and the way they could share their work.

2. Digital Lightbox

Developed at King's College London, and previously at University of Pisa, Digital Lightbox is a web application that aims to help paleographers,

[1] A Google project with the Israel Antiquities Authority to scan and preserve ancient texts. For further information visit the website: Dead Sea Scrolls, Digitizing the biblical manuscripts, http://www.google.com/culturalinstitute/about/deadseascroll.html

historians, art historians and philologists in doing qualitative research of digitized resources, such as manuscripts, paintings, etc., and also tries to fill the gap in the field of the Digital Humanities due to the lack of such tool.

Being inspired by the DigiPal project, Digital Lightbox has been mainly thought for paleographical research: indeed, this application allows a detailed analysis of one or more images through tools of manipulation, management, comparison (automatic and not) and transformation of images. The development of the project is currently being followed by the DigiPal Project team at King's College London, where it is being used as support for the analysis and the collection of paleographical samples, and it is being tested for new features and general debugging.

What makes Digital Lightbox not just an image processing application is the collaboration factor: researchers, indeed, have the possibility to share and reuse their "working sessions". Another meaningful consideration is that this application is web based, and therefore it is potentially working on any system which has at disposition a modern web browser, without the need of a modern hardware nor the purchase of any license.

The research methodology that Digital Lightbox proposes is that of studying paleographical elements analysing their details thanks to the tools provided to researchers. This methodology is of qualitative type and tends to help subjective and qualitative interpretations, rather than relying on quantitative methodologies such as machine learning and pattern recognition, which are used for extracting objective outputs from large sets of paleographical images by number of projects. Even though this kind of projects are certainly promising, and are a new way to aid paleographers and give them a quantitative point of view regarding handwriting features, the results obtained have not yet proved enough efficacy to replace human interpretation. A first problem in extracting objective features thanks to automatic tools is how to interpret the results returned by such software; furthermore, ancient handwriting is particularly difficult to be analysed compared to modern handwriting, where results obtained by forensic experts are instead significantly successful: indeed, it is necessary to consider the fact that the analysis on medieval manuscripts is rarely done in optimal conditions due to images' resolution, to the physical state of the manuscript analysed (which are in most cases very damaged), to the natural variety of handwriting which often makes even difficult the human interpretation.

Because of the reasons explained above, it is difficult to speak of state of the art in applications like Digital Lightbox. All those projects which were

created in the past have been abandoned: for example, can be mentioned projects such as "Virtual Lightbox" developed at MITH and "Virtual Lightbox for Museums and Archives (VLMA)" developed at the University of Reading. A quite similar modern project, which is possible to be compared to Digital Lightbox, is *Shared Canvas.* This project is being developed at Stanford University, and is mainly thought for the collaborative annotation of digital resources using an interface for managing, in the same time, a multiple number of imagesl, and using the *Open Annotation* data model.

2.1 Practical tools and applications

The features provided by the project involve sets of features which may be typically found in any image editor software, with the advantage of doing it on a browser and without any particular hardware requirements.

It is possible to use a set of filters applicable to any kind of image, such as opacity, brightness, contrast, greyscale, and color inversion. Used in combination, the filters could be a valid help in identifying and comparing whole images or part of them. It is possible, for instance, to use filters such as brightness and contrast for improving the visualization of damaged manuscripts, which could be of primary importance in the tasks of transcription and identification (see fig. 1).

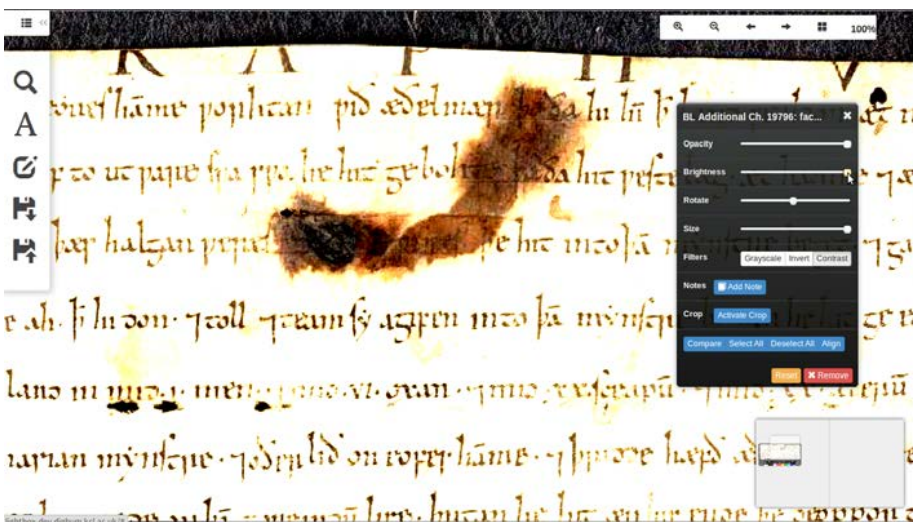


Fig. 1. Damaged manuscript recovering.

Furthermore, the application can be used to crop regions from any image, and collect them through folders which represent the manuscript where the region was cropped from. This is a quick method to create collections of letters, which, afterwards, can be automatically compared thanks to a tool for displaying the differences between two images (see fig. 2).

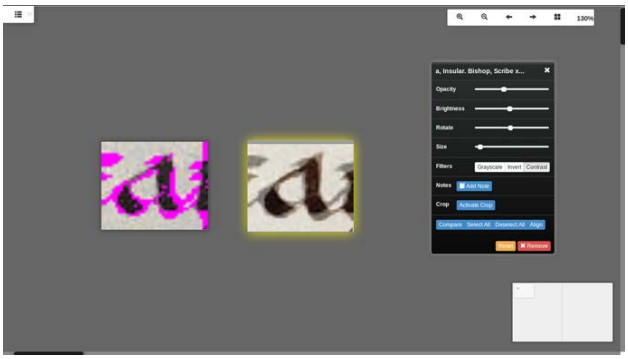


Fig. 2. Comparison made through the automatic tool.

A first application that has been particularly useful by applying the tools provided by the framework to digital manuscripts images is the application of the filters to damaged manuscripts. Another possible application is the disambiguation of different allographs[2]: as displayed in fig. 3, it is possible to see an example of working session where two ore more allographs could be compared in order to disambiguate them.

.

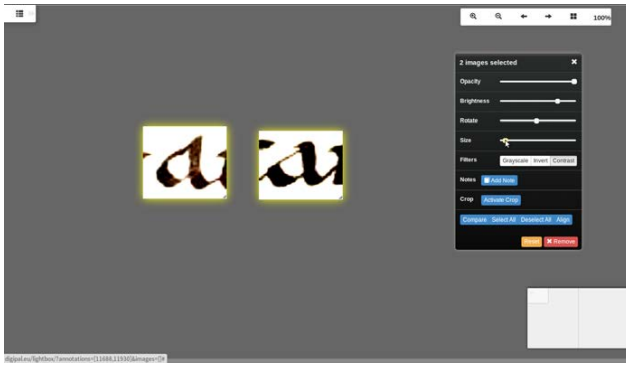


Fig. 3. Comparison between two allographs.

[2] A recognized variant of the same character (ex. Caroline a and Square a, or Insular d), according to the terminology used and formalized by the Digipal Project: http://digipal.eu/glossary/

2.2 Not just Paleography

As mentioned above, Digital Lightbox has been inspired and was born during my work on the DigiPal project, and therefore has been mainly thought for being oriented to paleographical analysis. Nevertheless, the project has been able to attract the interest of a number of different applications. Indeed, it could be possible to use this tool for any type of image, even thanks to the possibility of loading files from local disks. Some disciplines that may use the tools provided could be, for instance, art history, archeology, medieval studies, etc.

Another application of interest is the use of the project as support for teaching: indeed, the ease of showing the differences between different letters from manuscripts may be a useful application during paleography classes.

Being also a complete Javascript application, I worked for making it as much modular as possible: this means, basically, that the project could be used by taking parts of it, instead of using the whole project. Indeed, as a modular project, the code could be reused by taking single classes and functions, and creating, therefore, small instances of the lightbox for other purposes or projects.

2.3 Collaborative work

The possibility of conducting collaborative work among scholars is of crucial importance for this project: indeed, there are available a number of tools which make possible to save, export and import "working sessions", i.e. the current status of any task done by using the lightbox, such as all the images loaded, the regions criopped, notes written, or any filter applied. The session will be, therefore, loaded in the same way the user had left it, with the consequent possibility of sharing or re-editing it.

Another possibility is to export only the regions cropped in the current session: this can be done in both HTML and TEI XML formats, and potentially redistributable to scholars or students.

The advantage of using the TEI XML format as input file is dual: the task of writing an XML document is nowadays largely known in the humanities community (especially in the Digital Humanities), also thanks to the support of the TEI guidelines which facilitate text encoding; on the other hand, it is

certainly valuable the fact of reusing existing data: a good number of manuscripts and collections are already encoded in XML, or they can be easily converted from other formats, entailing, therefore, that it could be possible to load existing collections of manuscripts in the application without any need of setting up any database.

The characteristics and the features of the project try to contribute to humanities research not only thanks to the practical tools provided for managing and analyzing images, but especially thanks to the strong orientation to the collaboration among scholars and students in the field of the Digital Humanities.

2.4 Technical details

The web application has been build using the most powerful and recent technologies for web development: the back-end has been written in Python, and it is based on the popular web framework Django, while the front-end is based on JQuery, Bootstrap and some various Javascript third-party plugins. The project makes massive use of the recent technologies HTML 5, CSS 3 and the version 6 of ECMAScript, all necessary for building a modern web application. Compared to the projects I analyzed before starting my project, Digital Lightbox has not only a more modern User Interface, but even a technical architecture which makes it available for a various range of scopes. This is, for example, a prominent feature for making the application accessible, reusable and improvable by third parties: to achieve this, the application has been developed using a Django App architecture, which has the goal to make easier the installation and the deployment to already existing applications.

The Digital Lightbox project has been released under the open source license GPL3, whose code is hosted on GitHub and is freely available to be downloaded, edited and shared with anyone. Because of this, Digital Lightbox is also a potential future implementation for important projects such as DigiPal (Digital Resource and Database for Paleography, Manuscript Studies and Diplomatic), which is being developed at King's College London, and EVT (Edition Visualization Technology), at University of Pisa, Italy.

3. Future development

The next objectives that I aim to reach in the next future mostly concern the task of annotating images: it would be possible, therefore, to create, import and export annotations by using the standard *Open Annotation* (used, for instance, by the above mentioned Shared Canvas project). Others interesting developments which would be helpful to reach are a better cross-browser compatibility, and the quick configuration of the application for people which are not necessarily able to set up the Digital Lightbox, which could be helpful to those humanists willing to set up a project instance using their own content.

The last, and much more ambitious prospective, would be a real-time collaboration for one or more users working on the same session.

4. References

Stokes P. (2012). *Computer-Aided Palaeography, Present and Future*. «Dagstuhl Reports», vol. 2, issue 9, pp. 184-199.

Stokes P. (2007). *Palaeography and Image-Processing: Some Solutions and Problems*. «Digital Medievalist», vol. 3.

Niutta F. (2010). *Manoscritti nella rete*. «DigItalia», year V, num. 2.

Towards a shared methodology for audio preservation: Luciano Berio's private collection of sound recordings*

Federica Bressan, Sergio Canazza
Sound and Music Computing Group - Centro di Sonologia Computazionale (CSC)
Department of Information Engineering, University of Padua, Padova, Italy
{federica.bressan, canazza}@dei.unipd.it

Abstract. The article presents the key concepts of the scientific methodology for the preservation of audio documents defined and adopted at the Centro di Sonlogia Computazionale in Padova. These include: "accurate, verifiable, and objective" procedures, measurements based on an ideally objective knowledge, modern playback equipment, and the reversibility of each action ensured by a careful documentation. The entire process of preservation is characterized by a computer-science approach and it is carried out within a working environment especially equipped. The methodology is currently being applied to a financed research project aimed at the preservation of the entire private collection of sound recordings by Luciano Berio, which is also presented in the article.

Keywords: digital humanities, audio preservation, digital philology.

1. Introduction

Audio recordings represent an irreplaceable documentary source for studies in the field of linguistics, sociology, ethno-musicology and others. The awareness about the short life expectancy that characterizes all audio and video media dates back in the 1980s, when a lively international debate started in the archival community (for an overview see Bressan and Canazza 2013). It soon became clear that the paradigm of "preserving the original", still valid for other tangible cultural materials, fails to apply to audio media. So the concept of "preserving the content, not the carrier" was introduced, putting the emphasis on the acoustic information rather than on the physical object that stores it. And to preserve the acoustic information mean to incessantly transfer it onto new media, considered that short life expectancy (months or years, hardly decades) characterizes all types of existing media,

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

with no exceptions. However, "copying is not a value-neutral act; a series of technical judgments and physical acts (such as manual repair) determine the quality and nature of the resulting copy. It is possible, in effect, to distort, lose or manipulate history through the judgments made and the choice and quality of the work performed. Documenting the processes involved and choices made in copying from generation to generation is essential to preserving the integrity of the work" (Edmonson 2004).

The impossibility to achieve neutrality in the process of copying an audio document raises a number of problems that need to be addressed first from a theoretical point of view and only then from the technical-operative level. The Guide commissioned by UNESCO (Boston 1998) reports the philosophical approach "save history, not rewrite it", drastically limiting the freedom of choice reserved to those who perform the copy for preservation purposes, and assigning them a great responsibility. A couple of years earlier, Dietrich Schüller, former Director of the Phonogrammarchiv of the Austrian Academy of Sciences in Vienna, had illustrated a methodological approach that consists in analysing "what the original carrier represents, technically and artistically, and to start from that analysis in defining what the various aims of re-recording may be" (Schüller 1991). According to Schüller, there are multiple possible re-recordings, differing in some aspects that depend on the final purpose of the re-recording. Here are some approaches defined by the authors (Canazza 2012, 125):

- Conservative approach: it is the object of this article, and the approach that is being applied to Luciano Berio's audio tapes (for details, refer to the next paragraphs);
- Documental approach: it considers the recording for its cultural value in historical perspective, with all its relations with the context where it was produced. The playback equipment and the eventual restoration should not exceed the technological level of the era;
- Aesthetical approach: it pursues a sound quality that matches the current expectations of the public (note that this may vary over time);
- Sociological approach: it has the purpose of obtaining a historical reconstruction of the recording as it was listened to by the people of the era (Storm 1980);
- Reconstructive approach: it has the objective of preserving the original (assumed? documented?) intention of the author (Storm 1980).

The dichotomy between carrier and content (i.e., artifact and information) distinguishes audio recordings from other cultural materials such as

sculptures and paintings (see Figure 1): in these cases, preservation and restoration are addressed to the object representing the cultural good, the meaning of which cannot be separated from the physical expression (Brandi 1963). The fact that the musical artwork or the acoustic information do not coincide with the artifact (the medium) allows for restoration to be carried out on a digital copy of the original recording, preventing any destructive incidental damage and allowing at the same time multiple solutions (interpretations) which cannot coexist in the traditional restoration school addressing sculptures and paintings. In other words, the value of an audio recording is represented by its *content*, while the medium is only the *container*. This is a general statement that can be applied to all audio documents, with one important exception: electro-acoustic and experimental musical compositions (see section 3).

The preservation of audio documents is divided in *passive* and *active*: the first aims at defending the medium from external agents without altering its structure, while the second involves data transfer onto new media (*re-mediation*). Passive preservation is further divided into indirect and direct: indirect passive preservation does not physically involve the audio document, but it involves preventative strategies that span from fire detection/suppression systems to disaster preparedness, including personnel training, environmental monitoring, and even political/legislative actions for the safeguard of cultural heritage. Indirect passive preservation is to be intended in a very broad sense, and the reason is that in a general perspective it becomes clear that documentary heritage is part of life and part of society, and it is influenced in various degree from many factors that are apparently far from the archives walls. Also the education of the general public, starting from the school system, influences the preservation of documentary heritage in the long run, because it modifies the perception that the value of the cultural patrimony holds in society, which is translated in economical support for preservation, access and fruition. Direct passive preservation is concretely focused on the documents, and it is generally carried out inside the archive. The documents are treated (cleaning, repairing, restoration) without altering their structure and composition. This article mainly focuses on active preservation, that is on the transfer of the acoustic information onto another media, namely a digital non-audio carrier (such as a redundant array of independent disks or an LTO[1]).

[1] Linear Tape-Open: a digital magnetic tape storage system.

Fig. 1. The dichotomy between carrier and content (i.e., artifact and information) distinguishes audio recordings from other cultural materials such as paintings (on the left, the Mona Lisa at the Louvre Museum in Paris). On the right side of the figure, the dashed lines separate the acoustical information as emitted by the source (top), the physical media where the information is stored for playback (middle), and the acoustical information as extracted from the media in (bottom). The methodology described in this article aims at the long-term preservation of the audio signal (bottom) accompanied by useful information about the physical media.

2. The Luciano Berio project

The research project dedicated to the private collection of audio tapes by Luciano Berio, and stored by the Centro Studi Luciano Berio in Florence, started in mid-2013 and will last until 2017. The funding is entirely provided by the Paul Sacher Foundation in Basel, Switzerland, and the entire digitisation process takes place at the Centro di Sonologia Computazionale in Padova, Italy. The importance of the project lies in: (i) the value that the recordings hold for the world-wide research community of musicologists and musicians; (ii) the complexity that such an audio collection raises at a scientific-technological level due to the obsolescence of the media and of the formats. Additional complexity is given by the very nature of the recordings,

which include electronic compositions for magnetic tape, and rehearsals or live takes in acoustic scenarios where the distinction between the desired signal and noise is often ambiguous. The expected output of the project is a digital audio collection of preservation masters that meet the requirements of accuracy, reliability and authenticity needed to consider them a valid documentary source for scholarly studies.

Fig. 2. Luciano Berio working with some tapes recorders at the Studio di Fonologia at the RAI Milano.

In the Introduction it was mentioned that the value of an audio recording is represented by its *content*, while the medium is only the *container*. It was also mentioned that electro-acoustic and experimental music on magnetic tape are an important exception to this statement, otherwise true. Figure 2 shows Luciano Berio at the Studio di Fonologia di Milano with some tape recorders. This picture shows a typical and necessary working method of the composers coeval with Berio, eager to explore sound manipulation with magnetic tapes. Composers used the tapes for their experiments by cutting and joining them in unusual ways, and often they wrote notes directly on the back side of the tape. As a consequence, the tapes bear witness of the creative process; in the case of finished compositions, the tapes even coincide with the artwork. Especially if there is no score – which is frequent in experimental music – the tapes are the only testimony of the composer's work, and if the tape is damaged or lost, so is the work.

2.1 Luciano Berio's tapes collection

Luciano Berio (1925-2003) has been an authoritative exponent of the new generation of the musical avant-garde since the 1950s, experimenting with complex combinations of timbres and with the expressive resources of the female voice. In December 1954, Luciano Berio and Bruno Maderna created the first Italian studio of electronic music at the RAI Milano headquarters, inaugurated the following year as the Studio di Fonologia Musicale. There he was able to experiment with the interaction of acoustic instruments and electronically produced sounds. Berio's musical research is characterised by his attainment of an equilibrium between a keen awareness of tradition and a propensity to experiment with new forms of musical communication, and his commitment to music extended to other activities including conducting, the conception of concert series and the promotion of contemporary music. In 1994 he was appointed "Cavaliere di gran croce dell'Ordine al merito della Repubblica italiana", and in 1998 he received the "Medaglia d'oro ai benemeriti della cultura e dell'arte".

Fig. 3. Tape from Luciano Berio's collection during playback on a tape recorder STUDER A810, exhibiting severe magnetic coating shedding.

The audio collection of the Centro Studi Luciano Berio comprises nearly four hundred open-reel tapes, which are currently stored at the Centro di Sonologia Computazionale in Padova in a controlled environment (temperature, humidity, light). The assessment of the physical condition of the tapes is one of the first steps provided by the operative protocol for the preservation of the audio documents defined at the Centro di Sonologia Computazionale (Bressan and Canazza 2013). The priority codes determining the project roadmap and the treatment that is going to be applied to the tapes depend on this assessment, which is carried out with the customary visual/olfactory inspection by trained staff, as well as with specific physical-chemical analyses conducted in collaboration with the Department of Industrial Engineering - Chemical sector, of the University of Padova. In particular, innovative experiments are currently being performed in order to define a scientific methodology for the physical recovery of magnetic tapes suffering from a common condition known as SBS-SSS (Soft Binder Syndrome - Sticky Shed Syndrome) (Bressan *et al.* 2013). Figure 3 shows a close-up of a Berio's damaged tape, namely exhibiting severe magnetic coating shedding.

3. Foundations of the methodology for the preservation of audio documents

This section describes the foundations of the methodology defined and adopted at the Centro di Sonologia Computazionale in Padova for the preservation and the restoration of Luciano Berio's private tapes collection, which is then implemented in an operational protocol characterized by a computer-science approach. The expertise required to carry out the workflow vary from the history of technologies for sound recording to aesthetics, to the history of electro-acoustic music to signal processing and programming languages, making this field a highly multidisciplinary one, and requiring that the professional figures involved have a strong cultural and operative training.

3.1 Conservative approach

Reliability, accuracy and authenticity need to be a primary concern in long-term preservation (Allen 2013), especially when adopting a conserva-

tive approach. This type of approach aims at minimizing the manipulation of the document and of the recording, in order to avoid the introduction of any arbitrary element due to the choices that the operator bases on her hearing or experience. The training of the preservation staff is crucial in order to ensure that the protocol is applied correctly, but in this type of approach it is strictly forbidden to manipulate the audio signal according to a presumed personal preference or, in other words, to "enhance" the signal. Enhancement is not an objective action, but it depends on what one means by noise, by "a good sound", by *the* sound and many other terms that are usually employed when discussing audio quality. Aware that copying is not a neutral act (Edmonson 2004), the authors believe that in this approach the operator should try to transfer as much as possible of the original document into the digital preservation master, minimizing the loss of information (Bressan and Canazza 2013) and thoroughly documenting each action, in order to ensure reversibility (Schüller 2001). This approach answers the questions: how is the document? How was is transmitted to us? Documenting the process that generated the preservation digital master is particularly important in the audio field, because the medium from which the signal is extracted might be irrecoverable – in case of advanced degradation, with subsequent impossibility of future comparisons to determine a document's authenticity.

Despite the large attention that digitization and audio archives have received in the last decades, little attention has been paid to quality control and shared procedures, as the authors first pointed out in (Bressan *et al.* 2011). The foundations of the methodology defined and adopted at the Centro di Sonologia Computazionale in Padova derives from the study of the international debate started by the article "Proposal for the establishment of international re-recording standards" published by William Storm in 1980 and from the experience matured during a number of financed research projects carried out or coordinated by the authors[2]. The following is

[2] 2011-2012: GRAFO (GRammoFOni - Le soffitte della voce). Institution: Scuola Normale Superiore di Pisa. Topic: creation of digital archive of speech corpora stored on obsolete or endangered analog and digital audio carriers. 2009-2011: REVIVAL (REstoration of the VIcentini archive in Verona and its accessibility as an Audio e-Library). Institutions: Fondazione Arena di Verona and the Department of Computer Science of the University of Verona. Topic: preservation and restoration of the audio documents stored in the archive of the Arena di Verona Foundation. 2005-2006: POFADEAM (Preservation and Online Fruition of the Audio Documents from the European Archives of ethnic Music). Institution: University of Udine. Topic: preservation and restoration of different typologies of documents of Ethnic music,

a summary of the founding principles of the methodology for preservation with a conservative approach:

- procedures are "accurate, verifiable, and objective" (Storm 1980);
- measurements are based on an ideally objective knowledge;
- the playback equipment is modern, fully compliant with the format specific parameters of the recordings;
- ensure reversibility: a careful documentation of all measures employed and of each manipulation applied is produced and attached to the digital preservation master (Schüller 2001).

All of these actions are aimed at fighting a common enemy: the *falsification of history*, which is the problem of "authenticity" by another name. An interesting definition of authenticity was given by (Factor *et al.* 2009): according to this definition, authenticity cannot be evaluated by means of a Boolean flag, but it is rather the *result of a process*, and never limited to the resource itself but extended to the information/document/record system (CASPAR Consortium 2008).

3.2 Preparation for access

"Permanent access is the goal of preservation: without this, preservation has no purpose except as an end in itself" (Edmonson 2004). Although active preservation is a challenging endeavor, in order to achieve a set of deliverable audio resources additional steps must be planned. In particular, a shift in the approach is required: after ensuring that the historical faithfulness of the document is preserved (with the creation of the digital preservation master), it is necessary to consider what the documents mean within their context, why is the community interested in them. Competences other than the technical-audio-related are necessary, and they depend precisely on the content of the recordings.

The cataloguing staff is in charge of the description of the contents, which in practical terms involves *a re-organization of the audio materials*, producing new audio files that which derive from the digital preservation masters but are not directly related to them. The output of the cataloguing is an archive of access or deliverable copies. Only at the end of this procedure it is possi-

destined to remote fruition of an online archive. For more projects and details, please visit the institutional web-pages of the authors.

ble to build the applications for (remote) access and to develop more or less sophisticated access strategies.

The scheme in Figure 4 summarizes the main steps involved in the process of preservation of audio documents: the preliminary activities, finalized at setting the digitization roadmap and most importantly the methodological and operative rules; the re-mediation of the documents, the core of the scheme; and the preparation for access, which is the sum of the steps required before the realization of the hardware infrastructure and the software services for access by the final users.

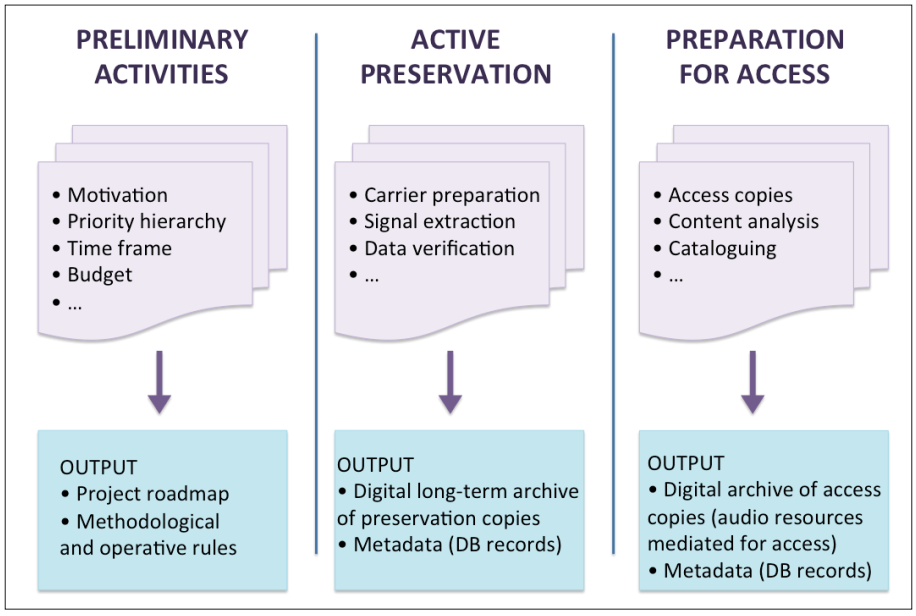


Fig. 4. The scheme summarizes the main steps involved in the process of preservation of audio documents according to the methodology presented in this article.

4. Acknowledgement

This work has been supported by the Paul Sacher Stiftung of Basel, Switzerland, within the financed research project (2013-2017) aimed at preserving the private collection of Luciano Berio's audio recordings, with the scientific support of Talia Pecker Berio and the Centro Studi Luciano Berio in Florence, Italy.

5. References

Allen A. a c. di (2013). *Interpares3 - team Canada final report,* University of British Columbia, Tech. Rep.

Boston G. (1998). *Safeguarding the Documentary Heritage. A guide to Standards, Recommended Practices and Reference Literature Related to the Preservation of Documents of all kinds*, UNESCO.

Brandi C. (1963). *Teoria del restauro*, serie Arte e restauro, Nardini, rist. 2005.

Bressan F., Canazza S. (2013). *A Systemic Approach to the Preservation of Audio Documents: Methodology and Software Tools*. «Journal of Electrical and Computer Engineering», vol. 2013, pages 21. URL=http://www.hindawi.com/journals/jece/2013/489515/. [last visit 30.3.2014].

Bressan F. *et al.* (2013). *Pavarotti sings again: A multidisciplinary approach to active preservation of the audio collection at the Arena di Verona*. «Journal of New Music Research», vol. 42, no 4, pp. 364-380. URL=http://www.tandfonline.com/doi/abs/10.1080/09298215.2013.840317#.UzgEFZTO58s. [last visit 30.3.2014].

Bressan F. *et al.* (2011). *Toward an informed procedural approach to the preservation of audio documents: The case of the "Fondazione Arena di Verona" archive*. In Proceedings of Sharing Cultures 2011 - 2nd International Conference on Intangible Heritage, pp. 177-185, Green Lines Institute, Tomar (Portugal).

Canazza S. (2012). *The Digital Curation of Ethnic Music Audio Archives: From Preservation to Restoration*. «Journal of Digital Libraries», vol. 12, no 2-3, pp. 121-135.

CASPAR Consortium (2008). *Report on OAIS - access model*, Centre national de la recherche scientifique (CNRS) and Université de Technologie de Compiègne (UTC), Tech. Rep.

Edmonson R. (2004). *Audiovisual Archiving: Philosophy and Principles*, UNESCO.

Factor M. *et al.* (2009). *Authenticity and provenance in long term digital preservation: Modeling and implementation in preservation aware storage*, in First Workshop on the Theory and Practice of Provenance, San Francisco, CA.

Schüller D. (1991). *The ethics of preservation, restoration, and re-issues of historical sound recordings*. «Journal of Audio Engineering Society», vol. 39, no 12, pp. 1014-1017.

Schüller D. (2001). *Preserving the facts for the future: Principles and practices for the transfer of analog audio documents into the digital domain*. «Journal of Audio Engineering Society», vol. 49, no 7-8, pp. 618-621.

Storm W. D. (1980). *The establishment of international re-recording standards*, in Phonographic Bulletin, vol. 27, pp. 5-12.

Knowledge objects and bodies of knowledge: knowledge sharing platforms applied to international relations*

Giuseppe Vitiello
Library and Knowledge Centre/NATO Defense College, Rome, Italy
g.vitiello@ndc.nato.int

Abstract. A reference service, including digital reference, is a standard library service. In the 19th century, library rooms were lined with monumental series of bound bibliographies. Later, libraries organized reference services based on printed sources. With the advent of online technology, digital reference became the buzzword. Now, what happens if library users are themselves first-hand sources, as in the case of Members of Parliament, researchers, analysts, in a word, "thought leaders"? Crowdsourcing used for knowledge management and discovery may be the answer. This contribution aims to show how a library can turn a major setback – the loss of its monopoly to the benefit of search engines – into an opportunity. It shows how the NDC Library and Knowledge Centre departed from the standard four-fold library organizational module – acquisition, cataloguing, reference, distribution – to adopt a multi-layer organizational structure, based on library, knowledge-management and publishing concepts. "A library is conversation", David Lankes maintains. Today, at LKC, a recorded item or a library transaction is relevant only if it relates to an on-going conversation, in the form of a link, an annex, or an object of reference. The advent of LKC thus marks a shift in focus, from objects to people. Stored knowledge has been superseded by "bodies" of knowledge, in the literal sense of the expression, the hearts and minds of NDC thought leaders. This is how we interpret crowdsourcing for internal knowledge capture and distribution. For centuries, books and articles in printed form have triggered conversations. Crowdsourcing for reference purposes can be implemented in an effective way through collaborative portals and shared resources, provided that these are adequately embedded into knowledge work flows and into the institutional roles of the actors involved.

Keywords: Knowledge Management, Crowdsourcing, Digital Libraries, Digital reference.

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

1. Introduction

First of all, let me thank you for your kind invitation to come to contribute to the 2nd AIUCD Conference. Digital humanities are a fascinating field, close to – and yet quite different from the traditional library world in which I have spent the bulk of my career. However, in speaking here, I am not changing my cap – libraries today provide the most important basis for the progress of digital humanities.

I have been given only ten minutes for my presentation; therefore, I will go straight to my main topic: the transformation of the NDC Library into a Knowledge Centre. Let me expand a few seconds upon the NATO agency I work for – the NATO Defense College. I am not doing this for marketing purposes, but because the description of the communities present at the NDC is instrumental to what I am going to say.

The NDC is a military-academic structure that provides for a variety of subject-specific programmes.[1] The flagship course at NDC – and its largest academic undertaking in terms of time to completion and scope – is the Senior Course. This five-month programme is run twice a year, with each edition usually catering for about 80 participants. It prepares officers at the rank of colonel and equivalent-ranking civilian officials, diplomats, and civil servants for senior appointments in NATO or NATO-related duties. Before joining the College, Course Members have often been involved in military operations within theatres of war.

An environment of this kind requires a special configuration of traditional reference services. As you know, reference is a library activity which consists in advising library users on materials and sources that are relevant to their information needs. These sources are normally books and journals – what we call secondary literature. How relevant is this kind of literature for NDC analysts who are often considered first-hand sources for military operations and strategic thinking? And what can a library offer to NDC specialists, to people who are, rightly or wrongly, deemed to be 'thought leaders'? Crowdsourcing may be one of the answers.

Please let me offer you now a history of reference services, from bibliography to crowdsourcing, in… just one slide! Reference books were traditionally – and still are – encyclopaedias, dictionaries and bibliographies, the monumental series of thick, bound books you often see lined in the loftiest

[1] http://www.ndc.nato.int/

spaces of a library. Two centuries ago, these reference collections started to be offered in conjunction with the individual support provided by some qualified staff known as Subject Specialists or, in German, Fachreferenten. Reference soon became one of the most qualified library services until ready availability of information on the Internet emerged as a challenge to its usefulness. This change of paradigm occurred despite of the fact that both reference sources and specialist staff were also going online, thanks to online databases and interactive tools, respectively. Today, crowdsourcing is the new frontier.

Crowdsourcing is the on-line contribution offered by groups of people to the creation or the enhancement of a product or a service. This approach has been applied in many fields, particularly in areas related to research and innovation. It has boomed in the media sphere, where amateur films or reports by improvised journalists are often used to record facts and events, thus creating the myth of the Internet as the ubiquitous medium.

Crowdsourcing applied to reference services is the natural extension of collectively generated encyclopaedias and dictionaries. At a general level, Wikipedia is the most popular and widespread application of crowdsourcing in the reference field. In the international relations arena, where sources are fragmented and local knowledge is indispensable to understand political change, crowdsourcing is an effective complement to secondary literature. I will illustrate how and why by making reference to the specific NDC case.

A landmark reference source for research in International Organizations (IOs) is the *Yearbook of International Organizations*. Released for over a century, the *Yearbook* changed its nature in 1953 and became a fully-fledged bibliography, enriched with resources published by, and on concerns of, IOs. The *Yearbook* reviews an impressive list of periodicals, reports and books published by IOs and NGOs; every year, some 24,000 reference items are incorporated in this excellent reference tool.

The *Yearbook* went online in the year 2000. Meanwhile, ad hoc portals and data bases took off, thus giving birth to the age of online reference services. The *WWWVirtual Library: International Affairs Resources* is one of them – a rich repository of resources of interest to the international community. *IAR* includes more than 2,000 annotated links to high-quality English-language sources of current information and analysis in a wide range of international affairs, international relations, international studies, global

studies, and global education topics.[2] Many other products add to the sources I have just mentioned and feed the current reference offer in the sphere of international relations.

A connected library, however, cannot merely rely on these sources. It must go further and delve into the very essence of reference. Books, articles and other kinds of academic materials help spur new ideas through intellectual exchange. Therefore, knowledge dissemination has its starting point in source materials; knowledge in action begins when content is interpreted and consolidated through conversations triggered by interested people. It may be useful to introduce now a typically Knowledge Management concept – the community of practices.

1.1 Communities of practices and workflows

Communities of practice are "groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and their expertise by interacting on an ongoing basis."[3] Two distinctive communities have a dominant role in the NDC educational workflows - Faculty Advisers and Course Members.

Faculty Advisers are sent by national governments to work for a period of approximately three years before leaving the College for further appointments. Unlike other colleges and academic institutions, the NDC does not have tenured teaching staff: Faculty Advisers are facilitators in charge of ensuring that the right inputs are provided for the smooth, effective running of the Senior Course and of the other courses of the College. They are responsible for organizing theme-based Study Periods, inviting lecturers, facilitating discussion within Committees, as well as running occasional inter-Committee debates and exchanges.

Course Members are divided into Committees, each with about 10 participants. In addition to preparatory reading for lectures, Course Members discuss course material and working on their Study Projects within their respective Committees. The course is organized in six distinct Study Periods, lasting a total of 23 weeks, during which lectures are given by visiting experts

[2] http://www2.etown.edu/vl/

[3] Wenger E., McDermott R., Snyder W. M. (2002). *Cultivating communities of practice.* Boston (MA), Harvard Business School Press, p. 4.

on specific topics. In their short stay at the College, Course Members carry out independent research to identify, locate and consult readings useful for individual and Committee-based work. Course Members' working practices are thus strictly regulated in a tight agenda of lecture attendance, individual work, presentations, and collective work on Study Projects.

There are not only Course Members and Faculty Advisers involved in the Senior Course. Other communities of practices also play a role, albeit a limited one. For instance, lecturers are fundamental actors in the education of Course Members. Their contribution to the educational work flows, however, is limited to the lecture they deliver followed by a Question and Answer session. The same goes for resident Researchers. They are the authors of NATO-related analyses normally published as NDC Research papers. Nevertheless, their participation in the Senior Course is limited to the mentorship they provide for the Study Projects carried out by Course Members.

Therefore, technologies at NDC mainly support conversations between Course Members and Faculty Advisers and among Course Members. The nature and quality of these interchanges are reinforced through "(1) exchange of knowledge among individuals; (2) exchange between individuals and knowledge repositories […]; and (3) exchange among existing knowledge repositories."[4] This three-fold concept underpins the design of the educational platforms at the NDC. 'Study Periods' and 'Committees' pivot the breakdown of portal sections and the related division of content for the Senior Course Curriculum and the Study Projects, respectively.

The following slide is an example of how activities linked to the Academic Curriculum have been structured. You can readily appreciate how central the notion of Study Period is:

The following slide relates to Study Project progress. It includes not only activities that are linked with the content of a Study Project, but also the steps devised to reach consensus within the Committee and the various drafts submitted to the Mentors and Faculty Advisers:

[4] Alavi M., Denford J. S. (2001). *Knowledge Management: Process, practice and Web 2.0*. In M. Esterby-Smith, M. A. Lyles, eds., *Handbook of organizational learning & knowledge management*, Wiley, pp. 105-124. Quotation at p. 106.

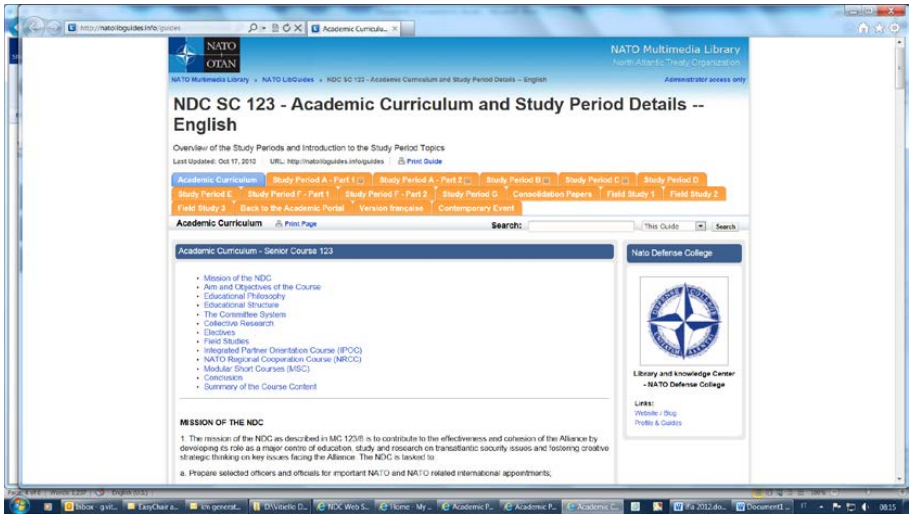


Fig. 1. NCD SC 123 - Academic Curriculum and Study Period Details.

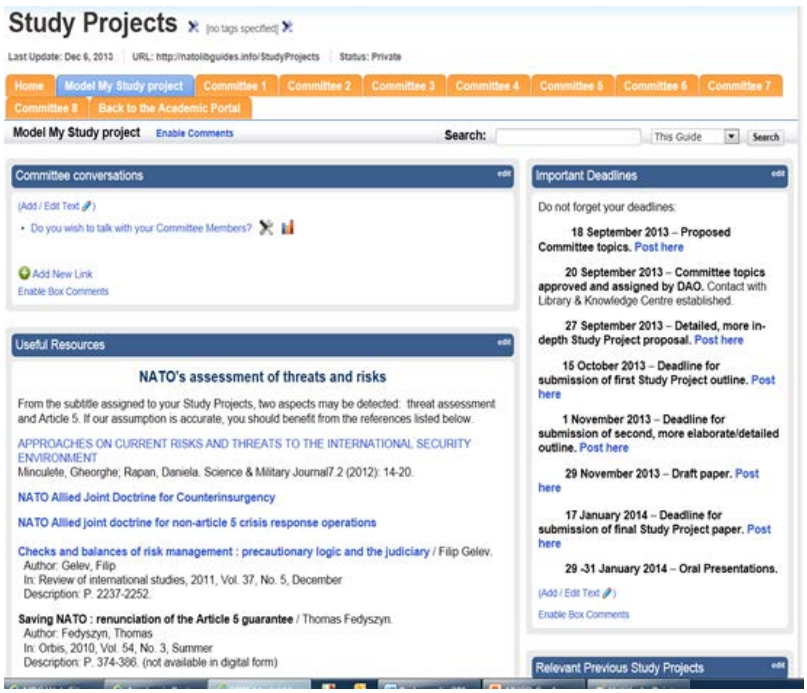


Fig. 2. Study Projects.

Course Members are free to use a blog to consolidate the knowledge they acquire and the conclusions they have reached:

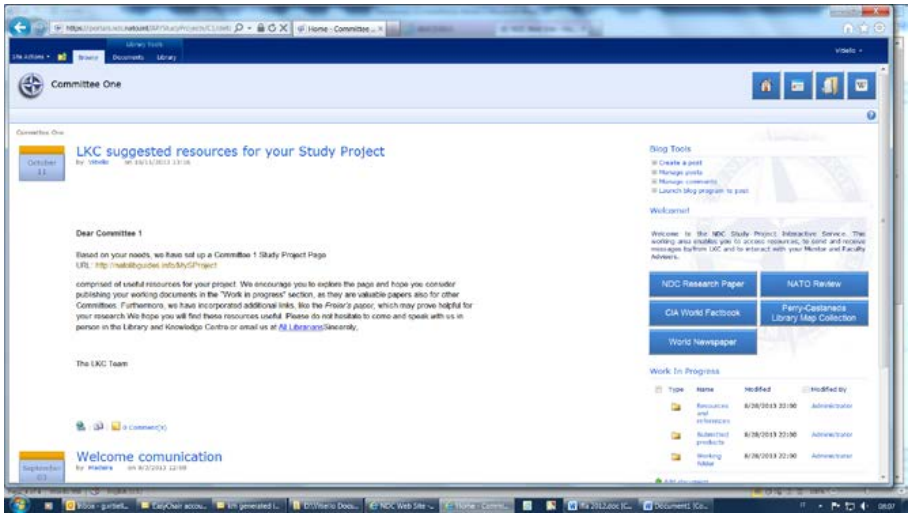


Fig. 3. The blog.

1.2 Organizational changes

I would like now to emphasize the importance of appropriate structuring for the library unit in charge of designing academic portals. Before the creation of the current Knowledge Centre, the operating structure at the NDC was still the classical three-fold service model, in which communication, command and control worked around three poles: acquisitions, processing, and dissemination of the retrieved resources. The library was a repository designed to be generally of interest to all library users.

This organizational asset was redesigned with a view to creating information reservoirs to be accessed by, and shared among, Faculty Advisers, Course Members and Researchers. These communities of practice refer to an activity room, a web page where they can undertake all operations related to their performance. The re-structuring of the library unit is represented in the slide below.

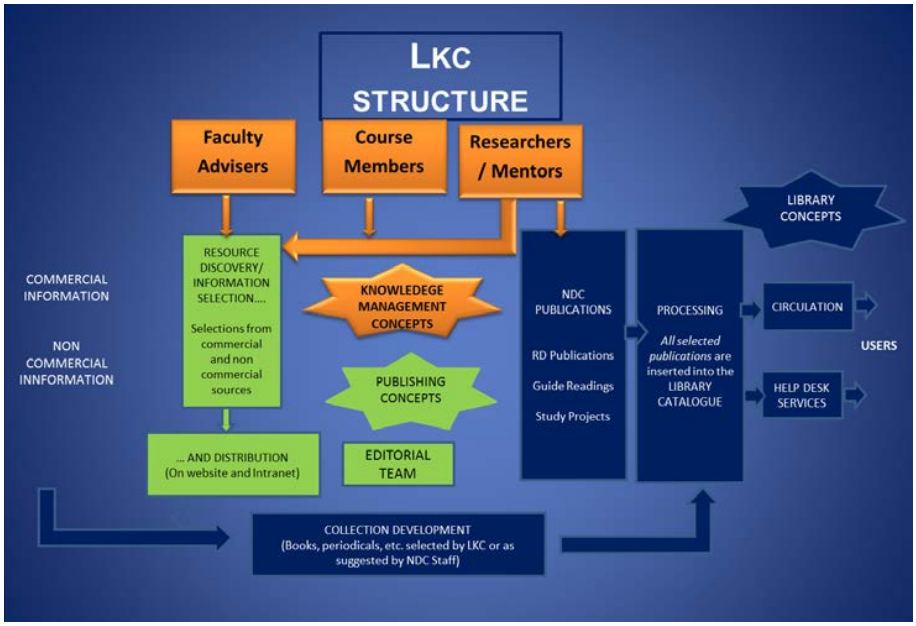


Fig. 4. LKC Structure.

2. Conclusion

The case for the NDC Library becoming a Knowledge Centre is not isolated. When libraries lost their monopoly on documentation, users began systematically bypassing library catalogues in favour of search engines. Cataloguing rules and pre-coordinated indexes became inadequate for search purposes; a simple search on Google was much more effective. Therefore, libraries still using the traditional catalogue system became progressively emarginated from the mainstream workflows of their respective institutions. To maintain a leading position, they have to make the most out of the knowledge / content / document management systems in use in their institution. This has been the only way for them to move back from a marginal role into the centre of information flows.

The German philosopher Jürgen Habermas developed the idea that in the 18th century, coffee houses, literary societies and *salons* were instrumental in the exchange and dissemination of different points of

view.[5] The advent of conversation in libraries means that their ultimate goal is to become like the coffee houses and the publishing houses in modern history: an environment with a dense mixture of activities linked with digital work, information dissemination concepts and innovative research discoveries. Through knowledge management tools libraries are set to become incubators of burgeoning ideas, outlets for theoretical innovation and a new arena for public debate.

[5] Habermas J. (1971[5]). *Strukturwandel der Öffentlichkeit: Untersuchhungen zu einer Kategorie der bürgerlichen Gesellschaft*, Neuwied und Berlin, Luchterhand.

Papers

Educational Approaches / Didattica

Moodle as a collaborative platform for digital humanities*

Giuseppe Fiorentino[1], Maria Accarino[2], Alessia Pierfederici[2], Daniela Rotelli[2]
[1]Accademia Navale di Livorno, Livorno, Italy
fiorentino@dm.unipi.it
[2]Laboratorio di Cultura Digitale, Università di Pisa, Pisa, Italy
{maccarino, alessia.pierfederici}@gmail.com, daniela.rotelli@istruzione.it

Abstract. We discuss a Collaborative Project Based Learning (CPBL) experience conducted during the academic year 2012/13 at the Digital Humanities Degree of the University of Pisa. The course ended with a collaborative project to put theory into practice: the creation of a (self-)training Moodle course on Italian grammar. We report on how the students accomplished the task using the Moodle and Cloud Computing resources as a collaborative platform.

Keywords: E-learning, Collaborative Project Based Learning, Virtual Learning Environments, Cloud computing.

1. Introduction

During the academic year 2012/13 the project was focused on the creation of a (self-)training Moodle course on Italian grammar: this task was a real challenge for novices at the tools and methods of e-learning. The project aimed at valorising the interdisciplinary skills promoted by the degree course: their humanistic skills allowed the students to work competently and creatively with the Italian language and grammar, while their ICT skills allowed them translate the contents into suitable online teaching materials. The topic of project was suggested by the finding that students often proceed to higher education with those cultural gaps that the entrance exam aims to spot early. So, we planned to exploit this assessment passage as the starting point for a more ambitious plan, consisting in a process of revision/recovery of the Italian grammar, in which the students could freely use the educational materials, and adapt them to their individual needs.

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.

The course realized during the project was not tailored to teach the whole Italian grammar, a goal that would surely exceed its educational purpose; however, it does not simply provide the knowledge needed to pass the entrance exam either. As a matter of fact, it aims at recalling and deepening notions that should be already known, and it does so with a practical, interactive and friendly approach.

2. Implementation of the project

An e-learning course requires careful planning and implementation; the following paragraphs briefly describe the evolution of the project, the tools used, the main problems, and the solutions adopted.

2.1 Start up and work planning

The interdisciplinary nature of the project suggested involving more teachers in the humanities and in the computer science degree programmes. However, having made this attempt to no avail, the project started with five students and the teacher. Preparatory meetings in presence were followed by intense remote collaboration with Moodle tools (forums, wikis and chat rooms) and cloud computing resources such as Dropbox and Google Drive, followed by audit meetings in order to plan the subsequent activities.

2.2 Analysis of old tests and construction of the syllabus

After having obtained the tests of past editions (not always in digital format), all questions underwent a meticulous classification and control process, to eliminate duplicates and check answers. We also attempted to reconstruct a syllabus trying to induce it from all the questions. However, it soon became clear that, by doing so, just a too small part of the Italian grammar would have been considered. The goal was then achieved using two well-known textbooks (Serianni 2006) (Tavoni 1999) from which to retrieve a syllabus able to guide the preparation of content (lectures, e-books), glossary and questions for all quizzes. This phase of teaching materials collection and coordination was fundamental to build and strengthen the learning and practice community.

2.3 Redaction of the glossary, of the e-books and creation of the interactive lessons

The following decision was fundamental to the project: using the syllabus as the framework of the Italian grammar course. All the items were grouped into six macro-areas (parts of speech, syntactic analysis of the sentence, graphematics, proposition, punctuation, phonetics) and inserted into the "glossary of the Italian grammar" which represents the lowest level of organization of contents, the one with the finest granularity.

All glossary entries share the structure shown in fig. 1: a brief, but exhaustive, description; some examples, to illustrate their use in everyday situations; some links, to deepen the topic and build connections across multiple levels of abstraction and detail (related glossary entries, e-books and interactive lessons).

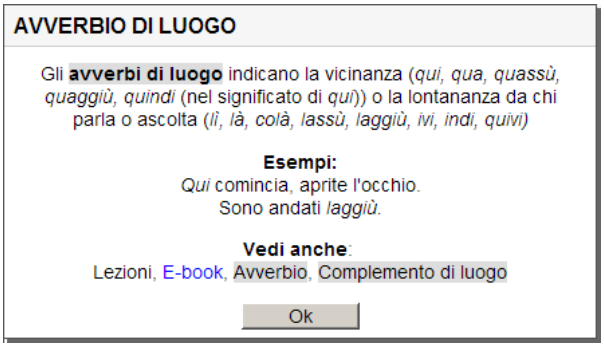


Fig. 1. An entry of the glossary of the Italian grammar.

The glossary provides a concise summary of all terms and makes them always reachable from any part of the course, using Moodle's automatic link feature. Every lemma, wherever used, has also a link that opens a pop-up window like the one shown in fig. 1.

A different approach was adopted for the e-books, one for each macro-area. By going through the syllabus in a systematic way, they represent the most detailed level of treatment of the subject, presenting a comprehensive overview of Italian grammar. Each page summarizes a topic with usage examples, possible variants and exceptions, bibliographic references,

and links for further reading. As for glossary entries, also e-books pages are automatically linked thanks to an extension of Moodle developed by one of the authors (Caldelli and Fiorentino 2011).

In order to avoid the simple linear exposition of the content and allow the development of review and recovery paths, some interactive lessons were set up to adapt the course to individual educational needs.

2.4 Creation of new questions and construction of thematic quiz

The most challenging part of the project has probably been the creation of (at least) one question for each item in the syllabus, providing general and specific feedbacks for each answer. The goal, already not trivial, became more difficult by looking at the tests of past years! The questions therein were repetitive and focused on a number of too restricted topics to serve as a general model. Creativity, network and textbooks made up, not without some difficulty. All questions were first inserted into (Moodle's) database and grouped according to categories and subcategories derived from the syllabus, and then used in quizzes which address the topics covered in the various sections of the course.

We also exploited Moodle's interactive quiz mode which allows trying again after a wrong attempt, after having provided specific feedback which doesn't interrupt assessment. In this way, by suggesting and adjusting the evaluation according to the assistance provided, the quiz retains its assessment purpose, adapts itself to each student's competence and gains an unexpected formative function. In fact, all feedbacks provide useful information: the general one points to the main subject of the question, while the ones for each answer briefly motivate the reply and link learning resources (glossary, e-books and lectures) related to the question.

Finally, also the choice to submit only a random selection of the available questions and the assessment scheme were adopted with a didactic purpose: the system, by recording the last score, stimulates the student to try the quiz more times. In doing so, he will face a greater number of questions to assess and deepen his knowledge, before or after studying the provided e-books and interactive lessons.

3. Moodle Implementation

Thanks to the great flexibility of its tools, implementing the educational trail with Moodle was simple and not limiting. On the contrary, the abundance of configuration options often suggested educational opportunities difficult to catch otherwise. All Moodle tools and activities (glossary, e-books, interactive lessons, quizzes and forums) were configured to implement both the educational trail and a semantic web able to keep each topic at just one click away from related ones. After a one-semester course one-learning tools and techniques, the students were able to accomplish the task using the platform in a creative and effective way, gradually exploiting its full potential as their experience grew.

3.1 Course structure and schedule

When the organization phase of the course was over (with the completion of the syllabus and the glossary of Italian grammar), the work was divided among the group members into five macro-topics. Each of these resulted into a Moodle section, structured as shown in fig. 2: an e-book, an interactive lesson, a quiz and a thematic forum used as a FAQ repository.

Fig. 2. The structure of the Moodle course on Italian grammar.

Agreed logic and structure of all learning objects, each team member was free to organize his/her work, harmonizing it with other commitments while keeping a constant virtual touch with the group. The teacher guided the entire project coordinating, directing and stimulating the progress of the work, agreeing templates to standardize individual work, evaluating it and supporting students with suggestions and corrections.

After the first meetings in January 2013, at the end of March it was already possible to perform some early tests on the first section of the course, assessing its strengths and weaknesses, and improving some functional aspects. Two more months were necessary to complete more sections (at this stage the work on the project was very irregular, due to the concurrence of other academic duties).

4. Afterthoughts

Almost at the end of the experience, here is the lesson learned so far from this engaging experience for all actors, students and teachers.

The chance to concretely experience theories and methods of e-learning was appreciated by the students. A strong stimulus to their educational growth arose from attending a constructivist course for the first time, a new methodology for (almost) all students, which asked active involvement in the creation of shared meanings. In particular, the opportunity to have a say in the project was much appreciated, it fostered new ideas and the sharing of best practices filing them into a "database of teaching resources".

It is also worth emphasising the quality of the interaction performed using Moodle tools (chat, forums, internal mail, wikis, and polls) and cloud computing facilities such as Google Drive and Dropbox. These resources are still used in the continuation of the project and proved to be valuable socialization tools and effective aid to work constructively together at any time and from anywhere.

An experience like this one shows the benefits of adopting Moodle for university courses, this is especially true for Digital Humanities. In fact, at the end of such degree program a student should be able to apply computer science techniques to the humanities, continuously renewing and deepening his/her knowledge as technology and its potential evolve. This twofold competency produces composite and flexible professionals able to handle humanist contents in digital form and transmit them. One of the main chan-

nels, in our opinion, will be e-learning courses, bound to have an increasing role in teaching and learning.

5. Conclusion and future developments

Conceived as an experiment in CPBL, the project turned out a great gym for the Teaching Technologies course students. The need for a specific education, due to the lack of a well-established syllabus for the entrance test to the Humanistic faculties, contributed in making this task more interesting and stimulating. The task was neither trivial nor obvious, because there was no previous experience but the one realized for scientific faculties (Di Martino *et al.* 2011). Ultimately, there were all the prerequisites to start a good CPBL: a real problem of great interest which could encourage students to create something innovative and useful while valorising their interdisciplinary skills. In our opinion, the CPBL setting is the only one that can allow a quick and long-lasting transition from knowledge to competence. Unfortunately, this approach is almost unknown Italy, despite its wide-spread appreciation abroad. It finds its natural setting in blended courses: in fact, the joint use of Moodle and cloud computing resources provided a valuable organizational and collaborative platform in all stages of the project, from the collection of teaching materials to their deployment in learning paths.

At the end of the work done so far, we believe that the experiment presented in this paper represents a didactic success. The good results already achieved keep us working on the project to complete, deepen and finally evaluate its effectiveness on the field. For this purpose, we are planning to test our project at school with the support of trainees/tutors. This phase will provide valuable feedback to further improve and consolidate the course, before making it available to a broader audience as a MOOC, therefore going beyond the preparation to a university entrance test, towards a lifelong learning perspective.

6. References

Caldelli D., Fiorentino G. (2011). *WikiGlossary e LinkBook, due mod per Moodle*. In Proceedings of MoodleMoot Italia 2011.

Di Martino P., Fiorentino G., Zan R. (2011). *Il progetto ELTP: dai test a scelta multipla ai percorsi individualizzati.* TD Tecnologie Didattiche, vol. 19, pp. 163-169, ISSN: 1970-061X.

Fiorentino G., Rotelli D., Accarino M. (2013). *Il Progetto T.I.F.U. "Test d'Ingresso per le Facoltà Umanistiche" Moodle come Piattaforma per il Project Based Learning Collaborativo.* In Proceedings of MoodleMoot Italia 2013.

Serianni L. (2006). *Grammatica italiana*, UTET Università.

Tavoni M. (1999). *L'italiano di oggi. Educazione linguistica & grammatica italiana*, Lemonnier.

Entry and Assessment Tests to access Science and Technology Faculties, Home page, URL=http://www.testingressoscienze.org/ [last visit on 03.31.2014]

Geostoria del quotidiano. Proposte per un'analisi automatica del testo letterario*

Alessia Scacchi
Dipartimento di scienze documentarie, linguistico-filologiche e geografiche
Università di Roma "Sapienza", Roma, Italia
alessia.scacchi@uniroma1.it

Abstract. L'apprendimento interdisciplinare dei contenuti letterari è un approccio costante specie quando si può giovare della tecnologia. Infatti, è stato avviato un lavoro sui testi letterari che prevede un costante confronto, non soltanto con le metodologie dell'analisi testuale, ma anche con la ricerca di contenuti cui il testo – per sua definizione polifonico – rimanda. un processo in divenire, avviato nelle mie classi, insieme a studenti e colleghi con cui stiamo costruendo uno strumento di condivisione ed analisi testuale che possa essere utile all'apprendimento della letteratura italiana, ma anche al confronto interdisciplinare tra materie diverse (biologia, chimica, cittadinanza e costituzione, diritto, fisica, inglese, matematica, scienze, storia, tecnologia applicata). La finalità principale è quella di far interagire e dialogare settori disciplinari differenti creando un ambiente di apprendimento pensato per studenti delle scuole superiori e, in prospettiva, universitari. Esso consentirebbe la reale interdisciplinarità dello studio a partire dalle opere letterarie, così da rendere agevole la fruizione di contenuti comuni, osservati da diverse prospettive analitiche.

Parole chiave: narratologia, analisi del testo, geostoria, apprendimento interdisciplinare.

1. La letteratura in(ter)disciplinata

L'approccio utilizzato come docente di Lettere, presso l'Istituto d'Istruzione Superiore "Statista Aldo Moro", deve necessariamente avere presente che gli studenti *digital born* necessitano di metodologie innovative rispetto alla tradizione didattica avviata con l'unità d'Italia.

Il lavoro sui testi letterari viene quindi ad emergere dal costante confronto tra le metodologie informatiche e lo studio dei testi cartacei; le une facenti

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. Cleup, Padova, 2014.

forza sul discretum, l'altro sul continuum macluhaniano. Quindi mutano al contempo le metodologie proposte per l'analisi testuale, per la ricerca e sistematizzazione dei contenuti, per lo studio dei nessi logici. Alle mutate condizioni dell'apprendimento si aggiungono, nel caso specifico, oggettive difficoltà incontrate da studenti di istituti tecnici e professionali, avvezzi alla manipolazione di oggetti, non di materiale testuale, anche in ragione del debole coinvolgimento nel lavoro critico sui testi.

Il Progetto *Geostoria*, quindi, si muove entro questa deissi e, anche per questo motivo, si configura come un processo in divenire che è stato avviato in alcune classi grazie all'apporto di studenti e colleghi interessati alla costruzione di strumenti di condivisione e analisi testuale. Il primo obiettivo, certamente, è che lo strumento sia utile all'apprendimento della letteratura italiana, tuttavia l'intento complessivo è connesso al confronto interdisciplinare tra materie diverse come: biologia, chimica, cittadinanza e costituzione, diritto, fisica, inglese, matematica, scienze, storia, tecnologia applicata per consentire ai discenti di percepire la letteratura come uno strumento di interpretazione del reale utile quanto una disciplina scientifica e/o pratica.

2. Dialoghi materici

Al fine di far interagire e dialogare settori disciplinari differenti è stato necessario creare un ambiente di apprendimento pensato per studenti di scuola superiore che fosse fruibile anche da allievi universitari, ma che consentisse in maniera fluida la massima interdisciplinarità nello studio fatto a partire dalle opere letterarie.

In tal modo sembra utile partire dalle competenze tecniche degli alunni scelti per il progetto affinché non emergano vuoti metodologici nella formazione secondo una serie di competenze di base. Innanzitutto è stato necessario comprendere quale fosse il grado di utilizzo della tecnologia e della rete internet che, apparentemente risultava alquanto ingenua. Infatti il web viene navigato in modo poco cosciente, poco informato e poco strutturato. Infatti i ragazzi conoscono ed usano intuitivamente le tecnologie a loro disposizione – veicolate generalmente tramite uno smartphone rivelando un generale grado basilare di interpretazione del materiale testuale veicolato dal medium elettronico; ovvero, essi mancano di capacità inferenziali, deduttive ed interpretative profonde, perciò la semiosi insita in alcuni enunciati composti da

testi ed immagini resta, ai più, mero ludus entro cui non è necessaria alcuna capacità ermeneutica.

Quindi la questione prima del lavoro sui ragazzi nelle scuole superiori sembra essere un dilemma: si lavora insieme a homini digitali e dunque primitivi? Certo sembra che un certo ritorno collettivo all'esperienza tattile ed intuitiva dei popoli del neolitico sia dilagante, cosa che comporta una scarsa predilezione per l'astrazione. I gruppi di nativi digitali, quindi, comunicano attraverso rapporti di causa-effetto molto lineari, chiaramente rispondenti alla logica dello 0 e dell'1, propria del supporto elettronico di cui si dotano, ma resta oscura tutta una logica ermeneutica di origine storica che essi ignorano e in qualche caso osteggiano.

Quindi, il procedimento orientato al supporto alla logica prende avvio da una corposa destrutturazione delle certezze acquisite tramite l'esperienza acritica del web, ed il conseguente orientamento in direzione di un mutamento delle fonti utilizzate dagli alunni. Tramite l'analisi semantica di social network, wiki e motori di ricerca come Google, è necessario orientare i ragazzi ad un uso cosciente della rete, consapevole di rischi e limiti connessi alla strutturazione dell'informazione in maniera automatica.

3. Techne… e poesia del testo

Il lavoro concreto in classe, con gli alunni divisi per gruppi – secondo le metodologie del peer-tutoring – prevede l'avvio dell'attività con una raccolta di materiali, ovvero i testi e la documentazione scientificamente affidabile presente in rete. La raccolta prevede, quindi, anche una sitografia che consente di elaborare una prima mappatura ideale sul modello del Progetto ancora in via di realizzazione di Ted Nelson[1]. A questa prima fase del processo di apprendimento è possibile far seguire la costruzione di un sistema aperto di implementazione dei contenuti utilizzando, in prima istanza, il blog[2] che è stato già costruito ed utilizzato dagli stessi studenti ai fini dell'apprendimento disciplinare. In costruzione è, invece, un corso sulla piattaforma multimediale dell'Istituto[3], mentre successivamente si ritiene necessario un modello

[1] Si veda la documentazione presente nel sito http://www.xanadu.com/
[2] Si veda il blog all'indirizzo http://letterattiva.wordpress.com/
[3] L'Istituto è dotato di una piattaforma moodle utilizzabile per fini didattici da ogni docente http://www.polocorese.it/moodle/

fondato su una mappa XML, quindi un modello semantico che strutturi le informazioni veicolate dal testo entro una gerarchia di significati codificati e standardizzati.

Le assi interpretative ritenute estremamente utili al fine dell'analisi dei testi e della interazione con contenuti disciplinari scientifici e/o sociologici sono: l'asse spazio-temporale insito nel continuum narrativo; i riferimenti diacronici connessi alla composizione e alla pubblicazione della singola opera presa in esame; i nessi cronotopici utilizzati dall'autore/autrice (Bachtin 2001). Infine, di notevole interesse sono anche i riferimenti testuali a luoghi propri (de Certeau 2001) ed indeterminati.

La selezione delle opere narrative è avvenuta scegliendo uno specifico arco cronologico che si sviluppa nel secondo dopoguerra italiano. Entro tale cornice, sono state scelte – per ragioni connesse alle disponibilità di tempo – opere della tradizione italiana entro cui è stato avviato il lavoro di marcatura dei fenomeni testuali di cui si accennava, di luoghi e date della pubblicazione cartacea, di luoghi propri ed indeterminati, di date citate nel corpo del testo al fine di aprire prospettive didattiche in ambito storico e geografico, con riferimenti guidati a materiali costruiti dai centri di ricerca universitari che nel mondo si occupano di questo periodo storico. Il tutto connesso ad una raccolta di documentazione iconografica, culturale e scientifica in senso lato, comunque connessa al testo letterario preso in esame.

L'esempio scelto per questa relazione prevede la ricostruzione delle tappe che precedono la pubblicazione dei racconti di Anna Maria Ortese confluiti nel 1955 nella raccolta *Il mare non bagna Napoli* da cui trarremo le esemplificazioni principali. Questo perché le storie, che descrivono la situazione postbellica partenopea, narrano il dopoguerra dal basso, ovvero focalizzando l'interesse e la riflessione sul popolo minuto e sulle difficoltà dell'esistenza nel quotidiano, cristallizzando panoramiche sociologiche dalle tinte forti e dai richiami letterari autorevoli[4].

4. Geostoria del quotidiano[5] in Anna Maria Ortese

La fascinazione che si prova nel leggere l'attacco del pezzo giornalistico intitolato *La città involontaria*, firmato Anna Maria Ortese e pubblicato

[4] Il progetto di scrittura sembra voler eternare personaggi ignoti oltre il silenzio cui li costringe la storia, in tal senso si avvertono echi foscoliani.

[5] Per la definizione come ricchezza e complessità delle interazioni ordinarie si veda Goffman 1956.

nel settimanale di politica e cultura «Il Mondo» nel 1952, corrisponde alla sensazione di nudità e di difficoltà difensiva che comunica il lucido e radiografico sguardo di Eugenia nel racconto *Un paio di occhiali*, il primo della raccolta della scrittrice edita nel 1953, con il titolo *Il mare non bagna Napoli*.

Il lead dell'articolo, che risponde alla rubrica *Napoli a zero*, squarcia le certezze del turista medio e dell'intellettuale italiano come uno dei boati dei bombardamenti, i quali costringevano la popolazione partenopea nella Napoli sotterranea. L'apparente sguardo asettico, scientificamente distante, si fa invito alla condivisione dell'esperienza che genera conoscenza:

> Una delle poche cose da vedere a Napoli, dopo le visite regolamentari agli Scavi, alla Zolfatara, e, ove ne rimanga tempo, al Cratere, è il III e IV Granili, nella zona costiera che lega il porto ai primi sobborghi vesuviani. È un edificio della lunghezza di circa trecento metri, largo da quindici a venti, alto molto di più. L'aspetto, per chi lo scorga improvvisamente, scendendo da uno dei piccoli tram adibiti soprattutto alle corse operaie, è quello di una collina o una calva montagna, invasa dalle termiti, che la percorrono senza alcun rumore né segno che denunci uno scopo particolare[6].

Napoli non ha molto da offrire al di fuori degli itinerari "regolamentari", sostiene la giornalista, eppure improvvisamente si scorge una collina, mentre la zona costiera fa da collante tra il porto ed i sobborghi attraverso una costruzione monumentale – in senso quotidiano – trecento metri di struttura, in cui un formicaio, in perenne movimento, sfida l'inutilità della storia di foscoliana memoria. Il brulichio prende la forma di una calva montagna di storie, prive di rumore per gli storici professionisti ma fragorosamente centrali nel racconto documentario che ne fa la scrittrice.

È un'osservazione scientifica del quotidiano che rappresenta, per Ortese, la via maestra per comprenderne i singoli atomi di significato; essa si colloca nel solco dell'anelito meridionalista proprio della rivista che pubblica l'articolo, animata da un eminente studioso del fenomeno[7] e testimoniata dai numerosi articoli usciti nell'immediato dopoguerra intorno alle questioni meridionali.

Eppure, una donna scrive da anni di sud dal sud, accrescendo il suo interesse per la materia narrativa offerta dalla quotidianità. Il male di vivere

[6] Ora in Ortese 1957, 28.
[7] Mario Pannunzio fonda e dirige la rivista "Il Mondo" dal 1948 al 1966.

emerge dalle rovine della Napoli postbellica, squarciata e ricostruita innumerevoli volte, abitata da una "plebe regina". La città decaduta è dipinta programmaticamente con un'ottica intradiegetica, cinica e debitrice nei confronti delle massime prove di letteratura verista di fine Ottocento, e si conforma come fruttuosa erede dell'impronta cromatica dell'espressionismo da secolo breve.

Buonanno afferma che una delle peculiarità della professione giornalistica al femminile coincida con la predilezione per uno «sguardo sensibile, dell'ascolto attento, del rapporto diretto», in tal senso la visione dovrebbe rivelarsi strumento di conoscenza peculiare per le professioniste della carta stampata. Perciò, sarebbe possibile istituire la priorità del «vedere rispetto al fare» ed Ortese stessa, quindi, sarebbe ascrivibile al solco delle diversità che si mutano in risorsa, proprie delle giornaliste che iniziano ad operare tra gli anni Trenta e Quaranta.

La guerra che ha coinvolto il mondo occidentale sembra essere svanita, relegata al racconto della memoria – vedasi la notevole produzione memorialistica del periodo – sbriciolata nella nube di detriti che segue al bombardamento. Affiorano le macerie di una realtà necessariamente diversa che, per essere narrata, ha bisogno di strumenti nuovi. Appaiono deboli i primi aneliti di ricostruzione che poggiano, tuttavia, sulle sconnesse assi sociali dei palazzi del potere, in gran parte divelti dalle deflagrazioni.

La guerra ha aperto numerose questioni, ha moltiplicato i detrattori del fascismo, occultato ignobilmente i conniventi, mentre Napoli rimane immobile con la sua umanità da termitaio, con le sue vene aperte che insozzano le strade, sotto il «cielo nero del soprannaturale»[8], Napoli resta «senza mare»[9]. La città appare come assopita, simile ad un «tappeto di carne»[10] assiepato da stanchi «mendicanti, minorati o semplicemente professionisti»[11], sdraiati a terra nell'inutile richiesta di un obolo.

Il dopoguerra guardato da sud è come osservato in salita, quasi intravisto nelle pieghe dei volti che Ortese si sofferma a descrivere. Tuttavia, la narrazione è schietta, semplice: la parola è riconquistata – come sottolinea Ghilardi – ed il silenzio è squarciato dalla "cattiveria" con cui l'autrice sente la necessità di radiografare il presente. Per questo aspetto, quindi, appare necessario

[8] Ora in Ortese 1957, 25.
[9] Come recita anche il titolo del volume.
[10] Ora in Ortese 1957, 25.
[11] Ortese 1957, 24.

dissentire da Faustini, quando afferma che la tendenza del new journalism diffuso in America dal XX secolo non ha epigoni in Italia. Infatti, già negli articoli di Ortese è possibile rintracciare, all'indomani della guerra mondiale, da una prospettiva periferica come quella del sud Italia, «tecniche di scrittura, di costruzione, di taglio del pezzo [...] proprie del racconto e del romanzo».

A fare da sostrato iconografico alla ricostruzione che segue l'istituzione della Repubblica Italiana è, certamente, un insieme di opere di generi differenti – critico-letterarie, filosofiche, fotografiche – che si adoperano per la proiezione di un meridione che dimentichi la guerra; tra le tante pubblicazioni, tuttavia, spiccano la raccolta di immagini di Pane e l'iniziativa del giornale di cultura di Prunas. Infatti, esse aprono squarci inattesi alla accomodante visione di un paese liberato a nord, ma deturpato nelle viscere a sud.

La scrittrice, dunque, realizza dei racconti che sembrano simili a documentari, in cui la fatica del quotidiano è narrata in forma critica; lo stilo dell'autrice affonda nelle vene aperte di una città assopita da secoli per pungolare intellettuali e popolo affinché siano protagonisti di un risveglio della ragione. Quindi, seguire l'ordine sincronico e diacronico dell'uscita in rivista dei racconti ortesiani si rivela molto proficuo in quanto è possibile ricostruire la temperie culturale che rappresentava il sostrato contestuale delle prove narrative per costruire una vera e propria geografia delle pubblicazioni su quotidiani e riviste per poi connettere le citazioni intratestuali ed extratestuali che contengono rimandi storici.

In tal modo appare evidente che il dopoguerra di Ortese è osservato dall'interno di un basso e contiene le immagini di una Napoli svuotata dalle bombe – come documentato nella raccolta d'immagini pubblicata da Roberto Pane[12], coeva all'uscita degli articoli – la città si presenta alla scrittrice con un volto inconsueto; è possibile osservare solamente l'ombra che traspare dal negativo della famosa cartolina che immortala la città per pura propaganda turistica.

Per il rifacimento del percorso geostorico che conduce alla redazione dell'opera ortesiana è necessaria, inoltre, una rassegna degli intenti e delle collaborazioni della rivista diretta da Pasquale Prunas – cui lavorò fattivamente Ortese fin dal suo rocambolesco avvio – con particolare riguardo allo spazio che la rivista dedica alla voce di donne e quello riservato alla rappresentazione realistico-verista della quotidianità. Una Napoli personificata, cittadina del mondo, è la protagonista degli articoli pubblicati su il Corriere

[12] Pane 1949.

di Napoli, il Nuovo Corriere di Firenze, Omnibus e Milano-sera. Interessanti sono anche i riferimenti all'analisi semantica delle pagine che ospitano i racconti/documentari, compresa la titolazione, che consente di rilevare ed interpretare le differenze programmatiche di ciascun quotidiano.

È possibile rilevare quali siano gli intenti della scrittrice nell'allestimento dei testi per le testate giornalistiche. Nel Corriere di Napoli la città è vista dal suo interno, così che la narrazione procede fluida agli occhi di chi conosce le quotidiane rinunce in: *Napoli senza mare* e *Pietà per i poveri*.

Nel Nuovo Corriere di Firenze, il discorso si orienta ammiccando alle dolcezze e ricchezze della capitale dell'arte rinascimentale in: *Oro a Forcella* e *Il giardino dei poveri*. Mentre per Milano – la pubblicazione avviene in Omnibus e Milano-sera – la narrazione e la titolazione necessitano di spiegazioni ampie e riferimenti concreti, ovvero ciò che Ortese realizza in: *Nelle strade della miseria è dorato anche il piattello del mendicante* e *Una farfalla al monte di pietà*.

Procedendo, quindi, all'analisi semantica delle pagine che ospitano i racconti, o meglio i documentari, l'attenzione va posta sulla titolazione, al fine di rilevare ed interpretare le differenze. Tuttavia l'analisi procede secondo uno schema geografico, ovvero si analizzano i testi a seconda della città in cui sono pubblicati, fino a ritenere possibile che il messaggio veicolato dai testi volesse oltrepassare le frontiere campaniliste per rivolgersi a «Il Mondo», rivista su cui appaiono: *Un paio di occhiali*, *La plebe regina*, *La città involontaria* e *L'orrore di vivere*.

5. Ontologie o non-ontologie? Una provvisoria conclusione?

Le ontologie presenti in rete per definire il letterario servono ad ottimizzare, standardizzare il sistema testo secondo i parametri di diacronia e sincronia; questo processo in ottica accademica è utile soprattutto per approfondire il materiale bibliografico sull'autore, sul testo e sul concetto che si intende evidenziare.

Al contrario, centrare il discorso sulla didattica multidisciplinare consente di evidenziare che il Progetto Geostoria rappresenta un esempio delle possibilità di modellizzazione; esse certamente si avvalgono delle potenzialità della rete senza naufragare in una dilagante e omogenea medietà.

Gli studenti hanno diritto ad un'istruzione superiore che consenta loro di acquisire un metodo critico.

6. Bibliografia

Asor Rosa A. (2009). *Storia europea della letteratura italiana*, Einaudi.

Bachtin M. (1979). *Estetica e romanzo. Un contributo fondamentale alla scienza della letteratura*, trad. it. Clara Strada Janovič, Einaudi.

Benjamin W. (1991). *L'opera d'arte nell'epoca della sua riproducibilità tecnica* (1955), nota di Paolo Pullega, Einaudi.

Bolter J.D., Grusin R. (2003). *Remediation. Competizione e integrazione tra media vecchi e nuovi*, Guerini.

Braudel F. (1982). *Le strutture del quotidiano. Civiltà materiale, economia e capitalismo* (secoli XV-XVIII), trad. it. Einaudi.

Carbonaro A. (1992). *Riproduzione sociale, vita quotidiana e soggetti collettivi nella sociologia italiana degli anni ottanta*, in L. Gallino (a cura), *Percorsi della sociologia italiana*, Angeli.

Ciotti F., Crupi G., a c. di (2012). *Dall'Informatica umanistica alle culture digitali. Atti del Convegno di studi in memoria di Giuseppe Gigliozzi (Roma, 27-28 ottobre 2011)*, Roma, Università La Sapienza.

Ciotti F. (2007). *Il testo e l'automa. Saggi di teoria e critica computazionale dei testi letterari*, Aracne.

de Certeau M. (2001). *L'invenzione del quotidiano*, trad. M. Baccianini, Edizioni Lavoro.

Eco U. (1979). *Lector in fabula. La cooperazione interpretativa nei testi narrativi*, Bompiani.

Fiorentino F. (2011). *Al di là del testo. Critica letteraria e studio della cultura*, Quodlibet.

Foucault M. (1977). *Microfisica del potere*, trad. it. Einaudi.

Ghisleni M. (2000). *Vita quotidiana*. In A. Melucci (a c. di), *Parole chiave. Per un nuovo lessico delle scienze sociali*, Carocci.

Gigliozzi G., a c. di (1987). *Studi di codifica e trattamento automatico di testi*, Bulzoni.

Goffman E. (1956). *La vita quotidiana come rappresentazione*.

Greimas A. (1970). *Del senso*, Bompiani.

Holister G. (1985). *Pensare per modelli*, Adelphi.

Luperini R. (1999). *Il dialogo e il conflitto. Per un'ermeneutica materialistica*, Laterza.

McGann J. (2003). *La letteratura dopo il World Wide Web. Il testo letterario nell'era digitale*, Bononia University Press.

Mordenti R. (2007). *L'altra critica. La nuova critica della letteratura fra studi culturali, didattica e informatica*, Meltemi.

Orlandi T. (2010). *Informatica testuale. Teoria e prassi*, Laterza.

Ortese A.M. (1957). *Il mare non bagna Napoli*, Einaudi.

Pane R. (1949). *Napoli imprevista*, Einaudi.

Riva M. (2011). *Il futuro della letteratura. L'opera letteraria nell'epoca della sua (ri) producibilità digitale*, Scriptaweb.
Scardigli V. (1983). *La consommation. Culture du quotidien*, P.U.F.
Segre C. (1999). *Avviamento all'analisi del testo letterario,* Einaudi.
Szondi P. (1992). *Introduzione all'ermeneutica letteraria* (1975), trad. di Bianca Cetti Marinoni, introd. di Giorgio Cusatelli, Einaudi.
Jedlowski P. (1994). *Il sapere dell'esperienza*, Il Saggiatore.
Jedlowski P. (2002). *Memoria, esperienza e modernità. Memorie e società nel XX secolo*, Angeli.
Wittgenstein L. (1967). *Ricerche filosofiche*, trad. it. Einaudi.
Zancan M. (1998). *Il doppio itinerario della scrittura, La donna nella tradizione letteraria italiana*, Einaudi.

Managing Educational Information on University Websites: a proposal for Unibo.it*

Federico Nanni
Department of Philosophy and Communication Studies, University of Bologna, Italy
federico.nanni8@unibo.it

Abstract: This article is focused on the complexity of finding and analyzing the totality of educational information shared by the University of Bologna on its website during the last twenty years. It specifically emphasizes some issues related to the use of the Wayback Machine, the most important international web archive, and the need for a different research tool which would guarantee more solid analyses of the corpus. This tool could initially be characterized by the use of standard Natural Language Processing techniques (such as tokenization, stop-words removing, parsing, etc.) but we also have to take into consideration more complex solutions, such as text mining analyses, WordNet integration and an ontological representation of knowledge. Thanks to approaches like the one here presented, future historians will be able to efficiently study the evolution of a website.

Keywords: web historiography, digital history, web archives, natural language processing, latent semantic analysis.

1. Introduction

It is difficult to deny that the World Wide Web is the most important technological innovation of the last Century. As we all know, it has not only exponentially improved our capacity of communicating with other people, but thanks to constant improvements, it has been offering year after year new ways of sharing information (websites, forums, YouTube, Twitter). As IBM remarked[1], in 2012 2.5 quintillion bytes of data were created each day, so much that 90% of the data in the World Wide Web today has been produced in the last two years alone. Therefore, it is self evident that being able

* M. Agosti, F. Tomasi (Eds). *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. 2nd Aiucd Annual Conference, Aiucd 2013. CLEUP, Padova, 2014.
[1] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

to guarantee a precise and rapid access to all kinds of information available online is one of the most important tasks of these decades, and a robust interdisciplinary approach is needed.

However, the exponential increase of digital information is not the only problem which researchers interested in studying the Web have to face in this era. Since 30th April 1993[2], people have been creating websites and sharing information online. Facing the volatility of documents digitally born, web preservation has become another important task, which now involves National libraries, National archives and other no-profit privates organizations.

Taking all these reasons into account, my primary research objective is to emphasize both the difficulties and the potentialities that emerge when web historians[3] have to deal with documents that are digitally created. Specifically, I intend to remark the importance of Natural Language Processing and Text Mining techniques in studying these new sources.

In this article I will focus my attention on one specific case of study: how the University of Bologna has used its website as a platform to communicate educational information to its students, during the last twenty years. My interest will not be uniquely limited to a historical reconstructions of these materials, but I will also aim at performing a diachronic analysis of the changes in didactic at this university during the last two decades.

Even if my analysis is focused on a particular medium (the website) and a circumstantial kind of information (educational ones) related to a specific academic institution (the University of Bologna), it is important to notice that the same approach could be effective for other similar researches.

Given all the above points, the rest of the article is focused on three major tasks:

- First of all the "hypothetical corpus" I intend to recreate is described, and the importance of web archives for this specific task is remarked.
- Secondly a discussion on what kind of digital tools could be effective with a synchronic analysis of these materials is presented.
- Finally the importance, for historians, of developing new interdisciplinary competences and becoming able to deal with new born digital sources is emphasized.

[2] On this date CERN announced that the World Wide Web would be free for everyone, with no fees due.

[3] In my works I use the term "web historian" as Niels Brügger does in *Web History* (2010).

2. Twenty years of educational information on www.unibo.it

Web historiography is a recent and challenging field of study, different from but closely related both to Internet Studies and Digital History, as Niels Brugger (2012) remarked. It involves a solid interdisciplinary approach in the creation of web archives and at the same time it requires specific tools in order to analyze the materials preserved. An active participation of historians is also crucial in dealing with more theoretical issues, such as the difference between web materials and archived web materials or the reliability of these preserved documents.

Therefore, I will take into consideration all these factors and I also intend to comment both on the digital preservation of materials and their automatic analysis.

2.1 The hypothetical corpus

My research objective is the study of how the University of Bologna has shared educational information with its students since 1993 through its website. Thus, the corpus I intend to rediscover includes all course programs offered online by the University during the last twenty years. To do so, I intend to use both the materials still present on Unibo.it and, if they are available, snapshots of its website preserved in the Internet Archive.

2.2 Piecing together the past

It is well known that Italy doesn't have a National Web Archive and, even if important preservation projects have been conducted since 2006[4], currently the only way of consulting Italian digital past is through international web archives.

The most important and extended one is the Wayback Marchine[5]: launched by the Internet Archive in 1996 it has preserved more than 240 billions URLs. This platform offers an enormous collection of snapshots of

[4] http://www.rinascimento-digitale.com/magazzinidigitali.phtml
[5] https://archive.org/web/

websites, in chronological order. Nowadays it is the only web archive which guarantees the preservations of almost two decades of Italian websites.

Unluckily, the University of Bologna's website is not preserved there (fig. 1); supposedly it does not allow Internet Archive crawlers. Therefore, the only viable way of analyzing University of Bologna's digital past is by consulting materials still present on its website. Focusing on educational information, it has been offered on departments pages until academic year 2004/2005; however, because of an important website update in 2005, these information are often not available anymore[6].

Fig. 1. University of Bologna is not preserved in the Internet Archive.

In conclusion, it is presently possible to have access only to educational information from academic year 2004/2005[7], but without a web archive we are not able to guarantee that the layout offered today is the same as ten years ago and we cannot be certain of the preservation of educational materials linked from those pages.

[6] For instance, this is the old version of the homepage of the Department of Classic and Medieval Philology http://www2.classics.unibo.it/

[7] It is important to notice that this academic year is present both in the old version of the department websites and in the new ones.

2.3 From the "hypothetical corpus" to the real one

By analyzing this corpus it is evident that every year the University of Bologna offers online a massive amount of information related to its educational organization. As an example, let us focus on academic year 2012/2013: we have digital access to more than 6.300 different courses descriptions, characterized by a brief summary of the lessons, a bibliography, sometimes links to other educational materials, and the curriculum, research interests and a list of publications of the professor. If we multiply this corpus by ten years we obtain an enormous amount of non structured information which, were we able to analyze them, could give us a more complete understanding of the topic.

3. A possible approach for a synchronic analysis

Given the enormous size of the corpus previously mentioned, it is evident that in order to analyze this case of study an advanced search tool is needed: this could, for example, guarantee us to find specific courses by searching their main topic.

Nowadays this is not possible either using the generic search tool (http://search.unibo.it/) or the one dedicated to educational information (http://www.unibo.it/it/didattica/insegnamenti). Therefore, if we are interested in discovering which course is focused on "the historical importance of Charles Darwin", this is a piece of information difficult to retrieve with the system currently adopted.

3.1 Advantages given by NLP techniques

Given all these reasons, I believe that a viable solution could be a new research tool based on a Natural Language Processing approach. It will consider all the information listed in courses descriptions, on professors pages, and in publications and books abstracts as a single corpus. For instance, to improve the retrieval of educational information all courses programs will be tokenized; then all the stop-words and the other "useless words" (as "exam", "lessons", "program", etc.) will be eliminated. Next, a matrix will be created with the courses on the rows and the words selected on the col-

umns and each word will be then represented as a vector <word, weight>, where the weight is determined by its recurrence and its position in the text.

After a training period exploiting users feedbacks and machine learning algorithms to improve the process of selecting the exact "word-weight" relationship, this tool will allow easier asking of structured query to the system while also receiving more precise information.

Another advantage of this solution could be the possibility of emphasizing similarities between courses, especially when they are from different departments. For instance with this approach we could easily discover that professor Fabio Vitali (Web Technologies) and professor Francesca Tomasi (Digital Humanities) both teach, from different points of view, "semantic web technologies", as the topic is present in both of their courses programs and in their publications.

3.2 Other problems, future solutions

It is clear that a straightforward comparison between two vectors could be an initial improvement for the retrieval of documents but this approach will not be able to extract the main topics present in each program, as the simple recurrence of a word is not enough[8]. Becoming capable of automatically extracting and representing knowledge from a textual document is one of the most difficult challenges that computational linguists have been faced with for more than fifty years. Certainly some progress has been made, as Google Knowedge Graph or IBM's Watson showed us[9], but we are still far from the complete comprehension of meaning from an unstructured text.

However, we could improve the approach previously described on three major aspects, following the advancements in the field. First of all it will be important to guarantee the integration of the tool with structured bases of knowledge, such as WordNet[10] and DbPedia[11], in order to help the system

[8] As Dan Cohen remarked here http://www.dancohen.org/blog/posts/its_about_russia

[9] Here IBM's Watson project http://www.ibm.com/smarterplanet/us/en/ibmwatson/ and here Google Knowledge Graph http://www.google.com/insidesearch/features/search/knowledge.html

[10] WordNet is a gigantic lexical database for the English language, but several versions for other languages (including Italian) have been realized http://wordnet.princeton.edu/

[11] DbPedia is a project that aims at extracting structured information from Wikipedia and making them available on the Web http://dbpedia.org/

notice relationships between tokens which are synonyms and identify named entities.

Secondly it is evident that, with a bigger and bigger structured database, it will become vital to consider a more complex semantic structure of knowledge. Thus, creating an ontology representation of the educational activities offered at the University of Bologna would definitely help the planning of more complex analyses and information retrieval activities. A solution like this could be projected and initially experimented considering a limited number of courses, as the ones offered by the Department of Philosophy; here, another interesting option would be to automatically generate this ontology "from the data", with a bottom-up approach (Gangemi *et al.* 2013). Both methods will be considered.

To conclude, the last fundamental aspect that has to be taken into account is the use of more complex techniques for document clustering in order to discover relationships between different courses. Therefore topic models such as Latent Dirichlet Allocation (LDA) or the Pachinko Allocation Model (PAM) could definitely be appropriate for this purpose (Mendes, Antunes 2009 and Templeton 2011). For these reasons the software MALLET[12] will be a fundamental tool during this part of the analysis.

4. Conclusion: from synchronic analyses to diachronic ones

In this paper I underlined the complexities of reconstructing the digital corpus and developing a tool that could be able to automatically analyze the contents, offer more precise access to the information and also evaluate the proximity between two programs by their main topics.

However it is evident that the approach previously described could also be an efficient solution for scholars who wants to study temporal changing of this new kind of sources.

4.1 Studying the web of the past

As early mentioned, ten years of courses programs are preserved on the website of the University of Bologna, starting from the academic year 2004/2005. Therefore, by implementing the NLP approach here presented

[12] http://mallet.cs.umass.edu/

to documents created in different years, it would become possible to emphasize correlations between past programs, even when they are taught by different professors, to describe the changes of theme in the same course during the last decade and to discover the increase or decrease of reference to a specific topic. Being able to extend the corpus by digitizing the course programs from previous years will definitely offer materials for way more complex analyses.

However, it is also important to note the fact that the documents here analyzed are not always a satisfying testimony of what the real course was (or is) about: they tend to be vague and, sometimes, an exact copy of the ones written the previous year, with no updated information. Therefore it would also be important to sensitize professors to the importance of these contents, which often represent the only reference for prospect students interested in the course.

4.2 New sources, new approaches but also new historians

Studying sources which are digitally born is about to drastically change historiography. New problems are arising, such as web preservation issues and the unstoppable increase of documents, but at the same time scholars are developing different solutions every day. As the web historian will be one of the researchers who will benefit the most from these innovations, I believe it is important that he will become more and more able to cooperate in this intensively interdisciplinary field of study.

5. Bibliography

Brügger N. (2010). *Web History*, Peter Lang.

Brügger N. (2012). *When the Present Web is Later the Past : Web Historiography, Digital History, and Internet Studies.* «Historical Social Research», vol. 37, no 4, pp. 102-117.

Gangemi A. *et al.* (2013). *A Machine Reader for the Semantic Web*. International Semantic Web Conference (Posters & Demos), pp. 149-152.

Mendes A. C., Antunes, C. (2009). *Pattern mining with natural language processing: An exploratory approach.* «Machine Learning and Data Mining in Pattern Recognition», Springer, pp. 266-279.

Templeton C. (2011). *Topic Modeling in the Humanities: An Overview*. URL=http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/ [Last visited 26.03.2014].

Author index

Stampato nel mese di settembre 2014
presso la C.L.E.U.P. "Coop. Libraria Editrice Università di Padova"
via G. Belzoni 118/3 - Padova (t. 049 8753496)
www.cleup.it - www.facebook.com/cleup