Umanistica Digitale - ISSN:2532-8816 - n.5, 2019
G. Rockwell, M. Passarotti – The Index Thomisticus as a Digital Humanities Big Data Project
DOI: https://doi.org/10.6092/issn.2532-8816/8575

The Index Thomisticus as a Digital Humanities Big Data Project

[1]George Rockwell and [2]Marco Passarotti

[1]University of Alberta, Canada

[2]Università Cattolica del Sacro Cuore, Italy

[1]grockwel@ualberta.ca
[2]marco.passarotti@unicatt.it

Abstract. The Digital Humanities (DH), as Rob Kitchin reminds us, have always been interested in the building of infrastructure for research (15., Loc. 222 of 6164). Imagining how emerging technologies could first be applied to Humanities problems and then scaled up to infrastructure for others to use has been one of the defining features of the field, by which we mean the field has evolved through projects that experimented with the application of new computing technologies to the difficult problems of the Humanities. Such experimentation began with Father Busa's Index Thomisticus (IT) project (7.; 26.; 6.) which is why many genetic descriptions of the field returns to the Index. The Index Thomisticus (IT) project was not only the first, but also one of the largest Digital Humanities projects of all time, even though the outcome might, by today's standards be considered "small". The project lasted 34 years and at its peak (1962) involved a staff of as many as 70 persons all housed in a large ex-textile factory in Gallarate. For that time they were dealing with big data, we might even say really big data, and the infrastructure they had to build was unlike any ever built before. If we want to understand what is involved in scaling up to big infrastructure we should look back to the beginnings of the field and the emergence of big projects like the Index. This paper will therefore look at the Busa's project as a way to think through big projects by first discussing the historiography of the IT project and DH projects in general. We will ask how can we study projects as bearers of ideas? What resources do we need/have? Then we will look at specific aspects of the project that shed light on DH projects in general. In particular we will look at how the project was communicated, conceived, and the data processing innovations. Finally, we will reflect on what lessons the IT project has for us at a time when big data has become an end in itself. What can we learn from Busa's attention to data in the face of the temptations of automatically gathered data?

Come Rob Kitchin ricorda, le Digital Humanities (DH) hanno sempre dimostrato interesse nei confronti della costruzione di infrastrutture di ricerca (Kitchin 2014, Loc. 222 of 6164). Immaginare come tecnologie emergenti potessero venire prima applicate a questioni di area umanistica e, quindi, estese ad altri usi a livello infrastrutturale ha rappresentato uno dei tratti distintivi del settore, che si è sviluppato attraverso progetti che hanno messo alla prova dei fatti l'applicazione di nuove tecnologie computazionali a

problemi di area umanistica. Siffatta sperimentazione iniziò con il progetto dell'Index Thomisticus (IT) di padre Busa (Busa 1980; Winter 1999; Busa 1974-1980). Quello dell'Index Thomisticus fu non solo il primo, ma anche uno dei più grandi progetti di sempre nell'area delle DH, benchè esso possa essere considerato "piccolo" alla luce degli standard attuali. Il progetto durò 34 anni e al proprio picco (1962) includeva nello staff 70 persone, tutte ospitate presso una ex fabbrica tessile di Gallarate. Per quei tempi, costoro si trovavano a lavorare su "big data" (di veramente notevole dimensione) e l'infrastruttura che miravano a costruire non aveva precendenti. Se si vuole comprendere a fondo cosa significhi avere a che fare oggi con grosse infrastrutture, è necessario guardare agli inizi del settore e all'emergere di progetti come l'Index Thomisticus. Questo articolo studia il progetto di Busa intendendolo come un mezzo utile a riflettere sui grandi progetti e sulle infrastrutture nelle DH. L'articolo discute brevemente la storia del progetto dell'IT e, più in generale, dei progetti di DH. Ci chiediamo come si possano indagare i progetti di ricerca quali portatori di idee. Quindi, l'articolo tratta aspetti particolari del progetto dell'IT che riguardano i progetti di DH in generale. Nello specifico, ci concentriamo su come il progetto fu comunicato e concepito, oltre che sulle innovazioni che esso portò nel campo del trattamento automatico dei dati. Infine, l'articolo riflette in merito alle lezioni che il progetto dell'IT ancora oggi può insegnare, in un momento in cui i "big data" sono divenuti una sorta di fine in sè stesso. Ci domandiamo cosa possiamo imparare dall'attenzione minuziosa per i dati dimostrata da Busa, alla luce delle tentazioni odierne sollevate dall'acquisizione automatica dei dati.

Historiography of Projects like the Index[1]

If one believes, as we do, that projects are a form of distributed cognition that create meaning, then we should ask how they work at generating and bearing meaning and to what end. Despite a lot of attention being paid to project management, there is little about how digital projects can be read as bearers of meaning and even less on the forms of evidence. As Campbell-Kelly points out in his masterful *From Airline Reservations to Sonic the Hedgehog: a History of the Software Industry* 8. this is particularly true of software projects that leave little behind once the software doesn't work. The exception might be preservation efforts in game studies, but even with these so much of what is written about the history of software is anecdotal as game companies tend to protect their assets until they vanish 18.. We humanists, who since Lorenzo Valla have questioned sources closely and the uses of stories in history, should be part of the solution. We should take the historiography of computing projects seriously. But why bother with projects at all? Are they not the sort of ephemera that clogs the archive? Some reasons to care include,

- First, software is the medium of this age. In so far as the medium is the message we need to understand the projects that led to and maintain our knowledge tools.

1 One of the authors, Marco Passarotti, had a close personal relationship with Father Busa. Many of Busa's opinions discussed in this paper come from Passarotti's many conversations with Busa.

- Second, big data and its uses doesn't spring fully-formed from the head of Zeus. No, data, whether small or big, is input, captured, gathered, managed, aggregated, filtered, enriched, and so on without software. The software is often developed in one off projects that leave little trace other than their data. These projects build out the infrastructure we take for granted that then "gives" us the data as data (given). One can't understand the data without the project that gathered it.

So, back to the Busa *Index* which has the advantage of being extremely well documented. The *Index* can therefore serve as a *case study* right at the threshold between traditional humanities projects and digital humanities projects. More importantly the project initiated a "new era of language engineering" as Paul Tasman, the project's IBM engineer pointed out in a paper in 1957.

The indexing and coding techniques developed by this method offer a comparatively fast method of literature searching, and it appears *that the machine-searching ap*plication may initiate a new era of *language engineering*. It should certainly lead to improved and more sophisticated techniques for use in libraries, chemical documentation, and abstract preparation, as well as in literary analysis. (24., 256)

In other words, the project developed for the first time, methods for dealing with unstructured language, something fundamental to both DH and big data. According to Steve Jones 14., IBM oral history interviews with Tasman confirmed the influence of the project on innovations like Luhn's KWIC (Key Word in Context), which can be thought of as an application of concording.

As for a historiography, it is safe to say that the *Index* is a paradigmatic project due both to its influence and the wealth of materials in the Busa Archives housed at the Università Cattolica del Sacro Cuore in Milan.[2] So what materials do we have available? The Archives cover a time span of around 60 years (from the beginning of the 1950s until 2010) and contain different kinds of materials, which can be summarized as follows:

- Personal materials of Busa, like academic certificates, ordination details, a photocopy of his identity card etc.;
- Documentation about conferences, seminars and workshops attended by Busa, including materials used to prepare his contributions, versions of the text of talks given by Busa, programs of events, handouts distributed by other speakers (with handwritten notes by Busa) and various materials related to practical matters (airline tickets, Visa bills, hotel reservations etc.);
- Press articles in the Italian and international media on Busa and his research;

2 The documents discussed are kindly made available upon request under a Creative Commons CC-BY-NC license by permission of the CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy. The documents and images are contained in the *Busa Archives*, held in the library of the same university. For further information, or to request permission, please contact Marco Passarotti, <marco.passarotti@unicatt.it>, or by post: Largo Gemelli 1, 20123 Milan, Italy. For information on the archives contact the archivist Paolo Senna at <paolo.senna@unicatt.it>.

- Professional correspondence between Busa and his contemporaries (in academia, cultural heritage and libraries, administration, industry, politics, religious organizations etc.) in Italy and abroad;
- Personal correspondence between Busa and close colleagues and friends in Italy and abroad;
- Materials relating to particular phases of the *Index Thomisticus*, like print outs, punch cards, tapes, budgets, page proofs etc.;
- Photographs, each enhanced with the date and the names of the persons pictured;
- *Opera Omnia* of Busa [external to the *Archives*]: one copy of each publication by him.

What follows are three examples that give a sense of the richness of the archive. The first is an assessment of the IT project by Daniel L. McGloin, the Chair of Philosophy at Loyola University of Los Angeles (now Loyola Marymount) to his President critical of the idea of the *Index Thomisticus* project. More on this later.

From the standpoint of philosophy and theology, it is my opinion that the proposed work would have no utility commensurate with the tremendous mechanical labor it would involve. While an Index of technically philosophical and theological terms occurring in the Opera Omnia of St. Thomas would be very useful, the extension of the work to include all words in St. Thomas' works (including, I presume, conjunctions, prepositions, etc.) seems to me : - 1) of no great utility; 2) a sort of fetish of scholarship gone wild, and 3) a drift in the direction of pure mechanical verbalism which would tend to deaden rather than revivify the thought of St. Thomas. (I think he himself would have been horrified at the thought!).[3]

The second is a detail from a flow-chart from 1953 prepared by Paul Tasman and an unknown draftsman at IBM that shows the processes they had developed by that point.[4] It shows in one place the way they were processing information at Gallarate from input to forms of classification. The detail below shows how the process still called for scholars to make decisions about what would be the entry words in the concordance.

3 Letter from Daniel L. McGloin S.J. Chair of the Philosophy department to his Rector at Loyola University of Los Angeles. The date would have been right before February 14th, 1950 when the President, Charles S. Casassa S.J., enclosed this letter with one he wrote back to Busa on that date. These letters are held in the Busa Archive.

4 The Flowchart is dated 1952 and has PT – JEG initials that we assume stand for Paul Tasman and someone else at IBM. This is also at the Busa Archives.

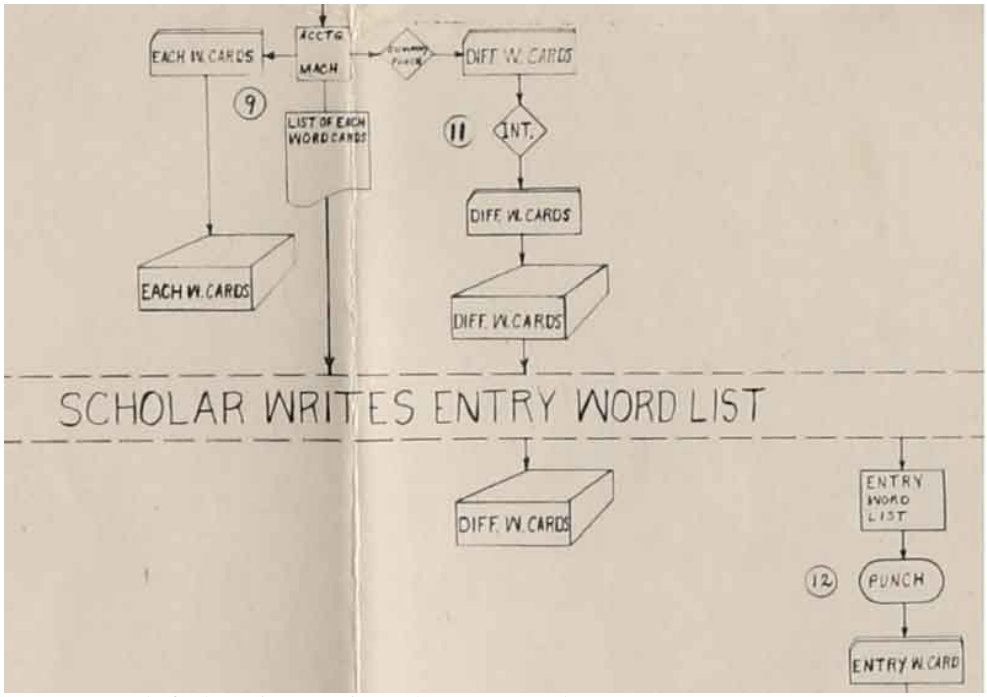


Figure 1: Detail of "Flow Chart: Mechanized Linguistic Analysis Project" 1952

The third example is the "Organigramma"[5] or organizational chart in a proposal "Per Completare Lo Index Thomisticus Per L'Esposizione Mondiale Di New York 1964 – 1965".[6] This was a proposal, as the title makes explicit, "To Complete the Index Thomisticus for the New York World Fair 1964 – 1965." It is, in effect, a grant proposal to IBM for the funding to finish in time so that the project could be featured in the IBM pavilion. This sped up completion of the project for 1964 was not funded, but the chart shows in red what was completed at that point. It also shows the types of work and project organization. Other sections of the proposal provide a financial breakdown for the completion with salaries.

5 This chart from the Busa Archive comes with notes and a legend. The Proposal is not dated but from the text seems to have been created in 1962.

6 Rockwell posted a long blog essay on "The Index Thomisticus as Project" that discusses this. See http://theoreti.ca/?p=6096. The project proposal dates from 1962.

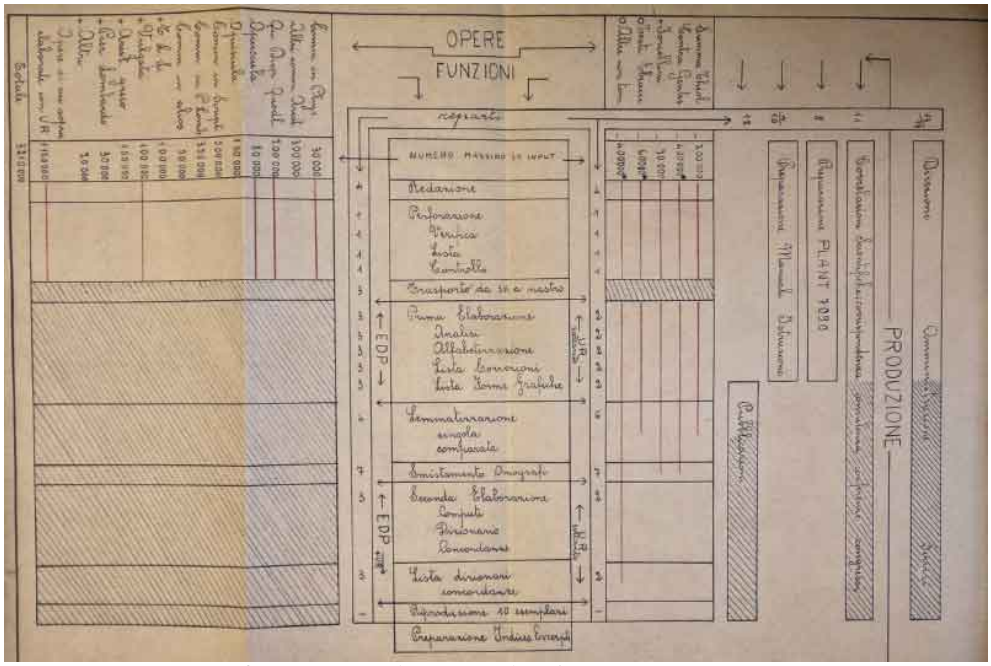


Figure 2: Organigramma from "Per Completare Lo Index Thomisticus…"

The *Archives* provide us with a model as to what might be saved about projects from correspondence to funding proposals. For a large project the materials that could be archived are potentially infinite, but the key is to preserve both documentation of the technologies developed, but also of the human organization and processes of the project. It should be noted that the *Archives* are freely accessible by sending a request to the Library of the Università Cattolica. The irony is that very few scholars seem to avail themselves of the *Busa Archive*, perhaps because the Digital Humanities are young and we aren't yet interested in reflecting critically on our history.

In closing it should be noted that, following a specific request by Busa, the *Archives* still retain their original organization in sections (and related boxes) arranged by Busa so we might say that the very form of the archive is of archival interest. Busa clearly had thought about the form in which his legacy project would be passed down.

The *Index Thomisticus* as Project

Having now argued for the importance of studying projects in general and the *Index* in particular, we turn to what we can learn about Busa's project from the materials archived. Given the scale of the archives and the project, we here will touch on three aspects of the project illustrated by three examples above.

Communications

Let us return to the letter from McGloin critical of the project. This letter is one of many back to Busa discussing the value of the project proposed. We can see in the Archive that Busa, at the beginning of the project, systematically wrote potential supporters and academics asking for expressions of support, and in most cases got positive, if occasionally guarded letters. He seemed to do this in rounds, providing templates for others as to what they might write and to who. In particular he encouraged supporters to write IBM so that IBM would have a sense of the perceived usefulness of the project. In this correspondence one can see a project manager making sure that he gets international support and that is communicated to his sponsor. This is but one example of how the correspondence shows the political management of the project.

Conception

The second aspect of the project is how it was conceived, by which we mean, *how did Busa outline the project to explain its conception*. We focus on this as the "high-concept" of the project was important to communicate in order to get support for the innovative process he proposed which is why some form of summary of stages shows up in a number of the key early documents, sometimes set out as a chart.

For example, in the Introduction of the *Varia Specima Concordantium* (to use a shorter title) from 1951 3., Busa summarizes his project as one of five stages and this summary is worth quoting.

> I bring down to five stages the most material part of compiling a concordance:
>
> 1 - transcription of the text, broken down into phrases,[7] on to separate cards;
>
> 2 - multiplication of the cards (as many as there are words on each);
>
> 3 - indicating on each card the respective entry (lemma);
>
> 4 - the selection and placing in alphabetical order of all the cards according to the lemma and its purely material quality;
>
> 5 - finally, once that formal elaboration of the alphabetical order of the words which only an expert's intelligence can perform, has been done, the typographical composition of the pages to be published. (p. 20)

The *Varia Specima* was a concordance of the poetry of St. Thomas Aquinas that was the "First

7 The word for "phrase" used in the parallel Italian (which Busa probably wrote first) is the more technical word "pericope" which means a "coherent unit of thought" and etymologically comes from "a cutting-out". In Tasman 24. the phrase used is "phrases (meaningful sub-grouping of words…)". (24., 253)

example of word indexes automatically compiled and printed by IBM punched card machines" as the subtitle put it. This publication was the proof-of-concept project for the much larger *Index* and it was the first fruit of the support from IBM for the Index. It has a bilingual Introduction (English and Italian) that describes their innovative concording process that used punched card machines.

In the *Varia Specima* he goes on to say that the IBM system could "carry out all the material part of the work" of steps 2, 3, 4, and 5, though elsewhere (p. 26) he talks about a philologist having to intervene at stage 3 to lemmatize words and so on.

In Paul Tasman's 1957 "Literary Data Processing" the stages are a little bit different 24.. The first stage from 1951 is now broken into two processes. That of the scholar who prepares the text and that of the keypunch operator.

> 1. The scholar analyzes the text, marking it with precise instructions for card punching.
>
> 2. A clerk copies the text using a special typewriter which operates a card punch. This typewriter has a keyboard similar to that of a conventional typewriter and produces the phrase cards.(p. 254)

By now the project is taking shape and the roles are emerging. Busa and Tasman are splitting the work between scholars and clerks.

Seven years later in "The Use of Punched Cards" Busa provides yet another slightly different process. This starts by focusing on who is doing the work and then shifts to the cards and outputs (4., 359). Again, the first stage of 1951 is broken into two operations – that of the scholar marking the original text and that of the entry of the phrases by a key-punch operator. However, the last stages of 1951 are in 1958 represented by a final summative stage for all the listings possible. This emphasizes the value of the initial stages for more than just concording. By then Busa and Tasman have realized that their method can be used for more than concording. They are imagining how it can be used for linguistic analysis and more generally "language engineering".

When we compare these conceptual outlines to the more detailed flow-charts we have we see that the five stages hide a much messier process. In the 1962 Organigramma we see in the central column the actual steps as managed by people that go into digitizing/concording the texts. This has more detail about the technical steps because it was a proposal to management at IBM who knew the project well and because it was a proposal for resources including human resources. It is not a conceptual overview.

Process

When we look at the details of the descriptions Busa and Tasman tell us about, two key innovations stand out. These innovations are now so basic to textual computing that we easily overlook them.

- First, they found a way of representing continuous text on punched cards so that it could be processed at all, and
- Second, they developed a mechanical process that made it possible to generate words cards from the phrase cards with relatively primitive machines – what we today call tokenization.

In short, they figured out how to represent unstructured text so that it could be processed by a computer for the first time and then they figured out how to tokenize or process the data into words such that the words could then be manipulated to generate various types of indexes. In these two coupled innovations they developed literary and linguistic data processing. If you have ever programmed a text processing tool you will know that data representation and the initial analysis into units like words are still fundamental. In this next part of the paper we will try to recover the context of these innovations.

Historical Context

From the distance of half a century it is hard to appreciate how different the *data* of "big data" was in those incunabular years. Busa could not search for and download the Aquinas texts he needed. Everything would need to be key-punched – it was not given without tremendous labor, both human and mechanical. There weren't even standard ways of representing texts for processing. And, he wasn't even initially using what we would consider a computer. The *Varia Specima* proof-of-concept project (1951) used electro-mechanical machines to sort, replicate, and print on cards. The late 1940s and early 1950s was a liminal moment between electro-mechanical and digital computing with punched cards in common as a way to enter, store, and process data.

The one technology that had been standardized at that point was the punchable card itself as a carrier of information. The punched card, despite all the complaints about folding, spindling, or mutilating, was a remarkably robust way of carrying information so that it could be processed manually and with computers.[8] It is a data technology that in its materiality goes back to the Jacquard loom and is still used today in some voting machines.

The Busa project used the by then standard IBM card format that had been designed in 1928. Each card was 7 3/8s by 3 ¼ inches of stiff paper with a notch in the upper left for orientation. The arrangement of holes had to be standardized so that the machines that processed them could all work and the IBM card had 80 columns by 12 rows of punch locations with square holes (in 1964 they introduced oval holes).

While the dimensions and punch zones were standardized, projects could print with ink whatever they wanted on the cards as what was printed wouldn't affect the processing. Busa's project was large enough that they had their own cards with custom areas for writing on.

8 For a nice cultural history of the punched card see Lubar 16.. For examples of manual punched card processing see Casey and Perry 9..

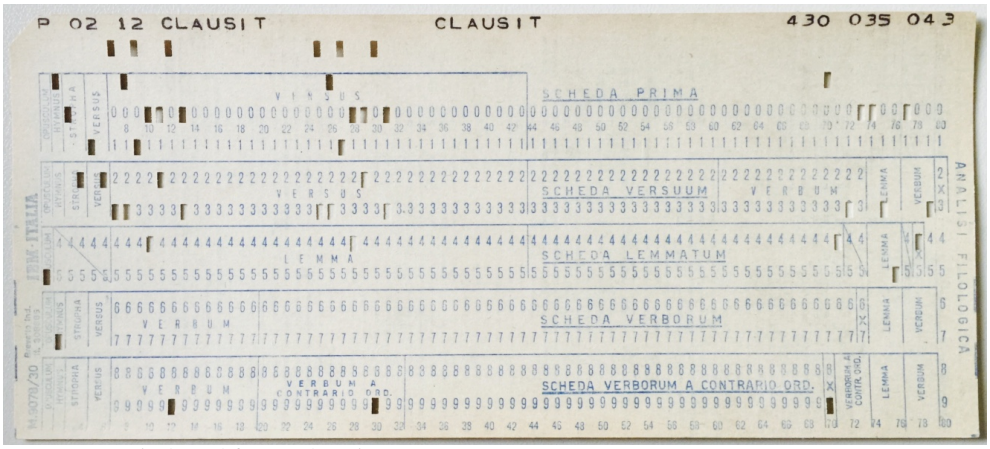


Figure 3: Punched Card from Index Thomisticus project

In 22 you can see the one of the cards in the Busa Archive with, on the right, an area with labels related to "Philological Analysis". As these areas don't have to do with what is punched we can speculate that the printed zones would be used by scholars to manually add annotations if needed. It was common for punched cards in those days to be manipulated both manually and with machines. The medium allowed the data on the card to be annotated by scholars creating two layers of information, that which could be manipulated by computing and that which could be manipulated by people.

To help understand how Busa may have used punched cards, Stéfan Sinclair and Rockwell have tried to virtually replicate his punched cards with a small toy that shows the correspondence between the holes punched and the data carried using the 48-character BCDIC or Binary-Coded Decimal Interchange Code 17.. This is a best guess of the card data format the IT used where you can enter examples and see if you get is what you expected.[9] For more on using replication as a form of media archaeology see Rockwell and Sinclair 22..

One thing that has become clear as one tries to figure out how they punched data onto cards is that there were no standards back then. Data formats were being developed and card processing machines were programmed from scratch for each project. You didn't need to worry about the data format of other people's texts or software and there were no operating systems for electro-mechanical card sorting machines. The issue would have been what characters the IT project needed and what a card could carry. The IT project may well have influenced IBM as they developed 48-character BCDIC 17..

It is also worth noting that the punched cards were not binary. For that matter they weren't really decimal either. Each column had 12 rows that could be punched. Originally 10 were reserved for numbers from 0 to 9, hence the "decimal" and then two extra rows were for accounting purposes (like negative balance). 40-character BCDIC extended this by punching combinations of two holes per column to get to numbers, letters, and a few more special characters. 48-character BCDIC added some three-hole combinations to add yet more special

9 See the Simple Punched Card Emulator at http://stefansinclair.name/punchcard/.

characters needed. These character sets were, of course, represented eventually in a 6-bit binary code that evolved into EBCDIC.

Having looked at the historical context of how the Acquinas texts were represented as data, we will now look at the other associated innovation, and that is the development of a technique comparable to what today we call tokenizing. Busa and Tasman developed a way to use the punched card machines of the day to take the phrases that had been coded on to Sentence Cards (Phrase Cards) and process them to get Each Word Cards (EWC) for each word (token) in the text. This was the great labour saving innovation that gave them two sets of cards, or we might say two database tables. One for each phrase and one for each card. With these two types of linked indexes they could generate just about anything else they needed. They could count words; they could sort on words and they could retrieve the full texts of words in order to build a concordance.

> Each phrase is preceded by the reference to the place where this line is found and provided with a serial number and a special reference sign. (24., 254)

The two types of cards held more than just a phrase or a word. The phrase cards held a location reference to the original text, a serial number and other special marks if the phrase was a quote from someone other than Aquinas. Given the limited space on the punched card, a scholar had to do the initial division of the text into phrases of less than 80 characters.

As for the Each Word Cards, they would have less text and more data. Tasman in his 1957 article, describes EWCs as potentially having,

- The reference to the phrase card (and hence phrase location),
- A special reference mark,
- The word itself as it appears in the text,
- Number (order) of the word in the text,
- First letter of preceding word,
- First letter of the following word,
- Form Card number (alphabetical sequence), and
- Entry Card number.

The Form Cards were what today we would call the word types. The Entry cards were the headings for the different word types after lemmatization and disambiguation. These were created by scholars who had to intervene in the process. This is the intervention in the detail of the Flow Chart in Figure 1 above.

It is worth emphasizing that the processes developed by Busa and Tasman were hybrid human and computing processes which is why the project still employed a lot of people, and the types of roles and their respective salaries appear in the Proposal from which we copied Figure 2. These roles were also gendered with women as punched card operators earning less than the

scholars who were likely men. Julianne Nyhan and Melissa Terras have been digging deeper into data entry for the *Index Thomisticus* project and have blogged about their work.[10] It is beyond the scope of this paper to reconstruct the organization of people involved in the project, but suffice it to say that the project at its peak had about 70 people with distinct roles working in a large former factory.[11] The team would have looked very different than a project team today given the number of punched card operators and administrative staff.

One last word about the project and how different it was from how we conceive of projects today. Busa's project would not have been described technically in terms of the data (corpus), the tools (programs) and the computing infrastructure (like web servers) used to deliver the project. Instead the project was described as the punching, sorting, hand annotating, and processing of punched cards. The cards were the surrogate representation of Aquinas' works, at least until they were able to move the data onto tape. From the photographs we have and evidence of the machines, we can imagine that a large part of the project was the careful handling of rows of cards; moving them to tables and then back to machines for simple processes. The computed processes would have been broken into simple tasks that a machine could be wired to do to long decks cards. One can imagine the project as a laborious "cut up" technique whereby the full text was rearranged and duplicated into a concordance.

The Legacy of Father Busa to Reconcile the Two Humanities

In the previous sections, we have shown to what extent the IT can be considered as a relevant historical example of a project concerned both with building and using "big data" in the Humanities. From the example materials from the Busa Archive one can get an idea of how much effort creating the IT required in terms of time, funds and research work, also addressing for the first time a number of fundamental issues in the field, like character encoding and data formats. This section will discuss what the IT as a project and the work of father Busa as an applied method of data handling still have to teach to Digital Humanities (and Digital Humanists) in 2019.

The Origins of the Index Thomisticus

The project of the IT was started because Busa wanted to find an efficient way to handle a very large amount of textual data, namely the 11 million words of the *Opera Omnia* of Thomas Aquinas. This was due to the need of Busa to support a new understanding of Thomas Aquinas' writings with all the information available, that is to say the very contents of his writings.

"Efficient way" here means two things:

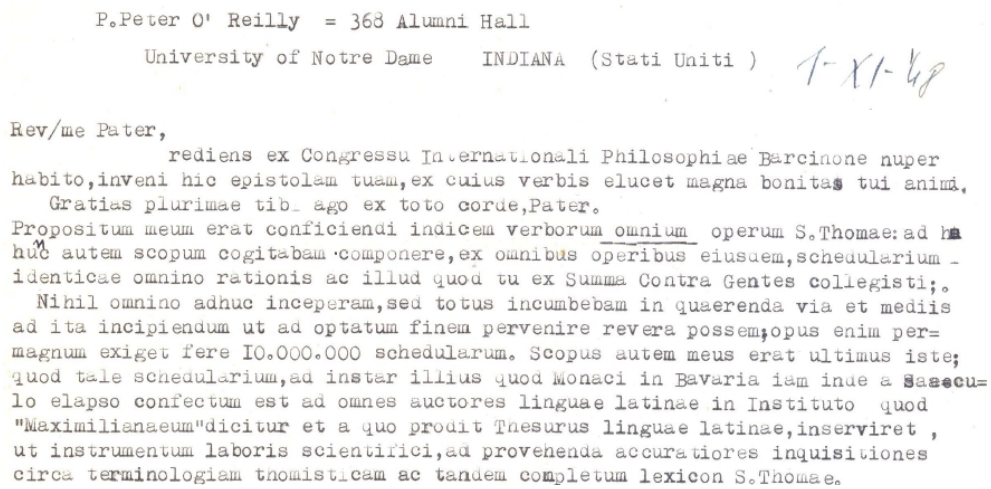
10 See Nyhan & Terras 2017 and also Melissa Terras' blog essay at http://melissaterras.blogspot.co.uk/2013/10/for-ada-lovelace-day-father-busas.html .

11 Steve Jones is developing a 3D walkthrough of what the project space would have been like as part of the Reconstructing the First Humanities Computing Center project that the authors are part of.

- A method to record, identify and retrieve all the occurrences of all the words in all the texts of Thomas Aquinas; and
- A tool for making it real automatically and in a replicable fashion.

This need of Busa was motivated by his extremely rigorous approach to empirical data. Such a careful attention to data comes through clearly in the motto "aut omnia aut nihil" ("all or nothing"), which characterized the entire scientific production of Busa. It was his firm conviction that, in matters of language, reliable conclusions can only be achieved via complete classifications of large amounts of data. Remarkably, in a number of publications of the 50s and the 60s, father Busa reports, as an added value of the IT, that the IT was about to contain the concordances of all the words from every text by Thomas Aquinas, "including *et* (and)". Today, this would not be such a big deal; but in the 50s it was a real innovation and Busa considered it one of the main merits of the IT. This thoroughness is also what McGloin thought was a useless effort that "would tend to deaden rather than revivify" interest in St. Thomas.

Since the very beginning of the IT project, Busa had clearly in mind an overall picture, which was motivated and driven by his approach to data. One can get an idea of this by reading what may be considered the oldest document witnessing the birth of what would later become the *Index Thomisticus*. This is a typewritten letter in Latin from November 1st, 1948 (taken from the Busa Archive) from Father Busa to Father Peter O'Reilly of the University of Notre Dame (Indiana, USA).



P.Peter O' Reilly = 368 Alumni Hall
University of Notre Dame INDIANA (Stati Uniti) 1-XI-48

Rev/me Pater,
rediens ex Congressu Internationali Philosophiae Barcinone nuper habito,inveni hic epistolam tuam,ex cuius verbis elucet magna bonitas tui animi.
Gratias plurimae tibi ago ex toto corde,Pater.
Propositum meum erat conficiendi indicem verborum omnium operum S.Thomae: ad hanc autem scopum cogitabam componere,ex omnibus operibus eiusdem,schedularium - identicae omnino rationis ac illud quod tu ex Summa Contra Gentes collegisti;.
Nihil omnino adhuc inceperam,sed totus incumbebam in quaerenda via et mediis ad ita incipiendum ut ad optatum finem pervenire revera possem;opus enim per= magnum exiget fere IO.000.000 schedularum. Scopus autem meus erat ultimus iste; quod tale schedularium,ad instar illius quod Monaci in Bavaria iam inde a saeculo elapso confectum est ad omnes auctores linguae latinae in Instituto quod "Maximilianaeum"dicitur et a quo prodit Thesurus linguae latinae,inserviret , ut instrumentum laboris scientifici,ad provehenda accuratiores inquisitiones circa terminologiam thomisticam ac tandem completum lexicon S.Thomae.

Figure 4: Part of a letter of father Busa to father O' Reilly

The text of the letter makes clear the practical objective and the methodology of father Busa. As for the former, Busa writes that he wants to build an index of the words of all the works of Thomas Aquinas ("propositum meum erat conficiendi indicem verborum omnium operum S. Thomae").

As for the latter, he plans to create a collection of almost ten million cards ("cogitabam componere […] schedularium"; "opus enim permagnum exiget fere 10.000.000 schedularum").

Busa also mentions the overall scientific objective of his work: like the index used in Munich for building the *Thesaurus Linguae Latinae*, the "schedularium" he would make should represent a kind of scientific instrument to support more accurate studies of Thomistic terminology with a complete lexicon of Thomas ("tale schedularium […] inserviret, ut instrumentum laboris scientifici, ad provehenda accuratiores inquisitions circa terminologiam thomisticam ac tandem completum lexicon S. Thomae"). From the clause "ut instrumentum laboris scientifici", one can see how the need to bring the scientific method into the Humanities is already present in Busa's thought since the very origins of the *Index Thomisticus*.

At the time when the letter was written, Busa was still looking for the way and the means to make it all real ("totus incubebam in querenda via et mediis"). Just one year later, he had found both way and means after his famous meeting in New York with Thomas Watson Sr., the founder of IBM, which funded the IT project for more than 30 years 20..

Towards a Methodological Turn in the Humanities

For Father Busa, collecting textual data in machine-readable format and making them searchable automatically was not only a valuable service for the research community, but it represented the main way towards a methodological turn in the Humanities. He was convinced that striving to formalize language for computing purposes represented an extraordinary method to get to a detailed knowledge of it. He argued that preparing textual data for computer analysis required the scholar to dedicate more time (and effort) than that required for non-computer-aided research. This is clear if we look at the detailed Flow Chart that Tasman prepared for the building of word concordances for the *Index Thomisticus* (see 2, dated 1952; see also 24.).

So, doing things faster is just one (and, surely, not the main) reason for using computers for processing linguistic data. Even before being practical, this is a methodological innovation, which becomes crystal clear if we look at the following excerpt from a paper of Busa published in 1962:

> [I]l 'libro magnetico' rappresenta un vero e proprio cambiamento di dimensione. Ma non è solo quantitativo né solo di velocità. È anche qualitativo. […] L'interpretazione induttiva del fenomeno linguistico […] promette di far ricominciare il ciclo della consapevolezza linguistica e grammaticale con maggiori profondità, sistematicità e documentazione. (5.: 117)[12]

"Induction" is a key word in Busa's thought and, thus, method. Any "interpretation" of linguistic data must be of the inductive type, i.e. based on as much as possible complete empirical evidence working as available documentation in support of reliable conclusions. Such a central role played by induction makes the concern for the quality of source data the essential

12 Translation (by Philip Barras): "the 'magnetic book' […] represents a real change of dimension. But it has not merely to do with quantity and speed; it is also a matter of quality. […] the inductive interpretation of the phenomenon of language […] promises […] to restart the cycle of linguistic and grammatical awareness with greater depth, methodicalness and documentation".

question in Busa's work. Not by chance, he used to quote the famous saying: "garbage in, garbage out".

Reflecting on the future of the discipline (being it Computational Linguistics or Digital Humanities), Busa was worried about the production and use of data. He used to say that the discipline would experience a "big boom" thanks to increasingly powerful computers, the widespread diffusion of digital technology, and the ease of transferring information across the Internet. This boom would lead to excellent results, but it would also be necessary to deal with the risk of upsetting the identity of the discipline, which he considered to be closely linked to the data.

He foresaw that the wide availability of large collections of digitized textual data and of tools for processing them automatically would run the risk of being incorrectly exploited. Busa believed the greatest danger lay in considering Computational Linguistics (and Digital Humanities, too) not as a discipline aimed at doing things better, but rather as a tool to do things faster, both in the phase of collecting data and in that of exploiting data. He feared that the computational linguists and the digital humanists of the third millennium would cease caring for the quality of data and lose the humility to check them carefully, preferring instead to process huge masses of texts quickly and approximately, without even reading a line.

Today, we see this fear of Father Busa coming true in opportunistic research works. There are projects in Digital Humanities dealing with enormous amounts of textual data; but unfortunately it turns out that often the data is not carefully checked, if checked at all. The availability of large sets of data tends to replace the careful gathering, enrichment and curation of appropriate data. The result is projects that do not move anymore from information to knowledge, which should be the real added value of any computer-based research work in the Humanities.

More generally speaking, this issue goes beyond the narrow borders of research in the Humanities, as it is strictly connected to one of the main risks of the so-called "information age", as presaged by Thomas S. Eliot in his poem "The Rock" (1934), where he asks:

> *Where is the Life we have lost in living?*
>
> *Where is the wisdom we have lost in knowledge?*
>
> *Where is the knowledge we have lost in the information?*

Sure, gathering data remains essential. Busa spent his life building what he used to call "documents", namely collections of carefully curated textual evidence (this is, essentially, what the IT is), which were given, like a gift, to the research community as the necessary empirical bedrock upon which any scientific knowledge should be based. Here, the stress is on the words "carefully curated", which does not always apply to contemporary projects in Computational Linguistics or Digital Humanities. This is not unproblematic, because such laziness in checking the quality of data is what today mostly alienates "the two Humanities", usually referred to as "Digital Humanities" and "Traditional" Humanities".

So many times, we have heard questions about "reconciling the two Humanities", or "strengthening" the dialogue between them. Paradoxically, it looks like this distinction today risks becoming even stronger. Digital Humanists tend to look at "traditionalists" as out-of-time and old fashioned scholars. And traditional humanists tend to consider their "digital" colleagues as impromptu geeks and, basically, unsuccessful humanists who recycled themselves as odd technicians.

This is even more of a paradox, if we think that such a distinction should just not exist at all and it is basically out of time. In 2019, any Latinist should be able to use a PoS (Part of Speech) tagger and a treebank just like he/she uses the *Du Cange* dictionary or *L'Année Philologique*. These are all tools that should be on his/her desk: some are papery, some are digital, but today being digital shouldn't be what matters, which shows why the very name "Digital Humanities" sounds old-fashioned.

In 2019, any humanist who wants to develop computational/digital tools and/or resources supporting research in the Humanities (a highly valuable task in itself) must be reminded that the very core of any empirical research work is data and its careful curation. Otherwise, we run the risk of falling into another paradox: the same discipline that wanted to be the most rigorous in dealing with data has lost its very core. Today we can quickly process large amounts of linguistic data automatically, but too often we do not know them. Equally, building huge masses of linguistic data is a pretty easy task, but too often we do not even take a glance at what we built and we provide the community with.

It seems like we take much greater care of the machine, but too little of the input data processed by the machine and of the output data produced by the machine. The focus of that area that is called Digital Humanities has been for years on digital tools in humanities contexts rather than on humanities tools in digital contexts. Now, we must consider what kinds of models, tools and approaches from contemporary Computer Science and Computational Linguistics Humanists might find useful, rather than starting with the tasks and approaches that Computer Science and Computational Linguistics researchers are most familiar with and asking how they can be applied to data in the Humanities.

In other words, referring to the name "Digital Humanities", we focus too much on "Digital" and too little on "Humanities", forgetting that the head of the phrase is "Humanities", while "Digital" is (only) a modifier that may have almost had its time as a distinctive feature of the field.[13] Indeed times look now mature enough to change the very name of the discipline from Digital Humanities to Computational Humanities, or to revive the old name Humanities Computing, as this would underline more the actions performed on data ("computing") than just their format ("digital"), thus giving back to data that central role that they had at the beginning of the discipline and, more generally, in every rigorous scholarship in the

13 There are several definitions of what "Digital Humanities" means; and this is a further sign that the name is not the best choice. A brief discussion on the topic, with a special focus on "Digital", is provided by Ciula 2017. See also Cecire 2011, whose definition of "Digital + Humanities" is close to our vision of what the discipline should be about. However, our discussion here concentrates more on the very use of the word "Digital" in "Digital Humanities" than on the different interpretations and definitions of what this label should denote.

Humanities.

This does not mean that we must become undigital. Rather, it is exactly the opposite. Busa claimed that "a machine made us realize" ("ci ha resi consapevoli") how limited is the knowledge that we have even of our own language. He says that "a machine has revealed that there is still too little humanism of the serious and systematic type" ("La macchina […] ha documentato che di umanesimo, di quello serio e sistematico, ce n'è ancora troppo poco") (5.: 105). Keeping the legacy of father Busa means pursuing such "humanism of the serious and systematic type".

Without such a methodological turn, there will not be a common language between Digital Humanists and Traditional Humanists (who have always been dealing with data) and we won't be able to achieve the natural objective of these years: i.e. finally to get rid of Traditional and Digital Humanities and have just one and only Humanities, based on rigorous and high-quality empirical evidence thanks to the application of computational methods, techniques, resources and tools and by complying with a bunch of best practices, which basically consist in being aware of the properties and the quality of the data to process and analyze before making any inductive inference based on them.[14] This is not in contrast but in line with Traditional or Historical Humanities, considering the highly empirical-based origins of some Humanities disciplines, and more profoundly of the shaping of the scientific method, well before computers were available, like for instance via the Italian humanism philological paradigm shift and the Enlightenment natural philosophy, to mention the most known examples belonging to the Western tradition.

Such "humanism of the serious and systematic type" requires one of the main (if not, the main) tenets of Busa's legacy: the need of replicability of results also in the Humanities, to move the field out of a partial and impressionistic approach to a more "scientific" one.

The FAIRness of Keeping the Legacy of father Busa

The concern for data curation and replicability of results is today regular practice in the scientific world and affects all disciplines, like for instance computational biomedical research, which makes use of open, web-based platforms for accessible, reproducible and transparent research, like for instance Galaxy (https://galaxyproject.org/).

Although this is still less the case in the Humanities, thanks to the current wide availability of easily accessible and usable data in several areas of the Humanities, the question of the replicability of results of empirical experiments is now at the heart of the validation of scientific knowledge and of the scientific endeavor even in the Humanities. A number of events were held recently on the topic, especially in the area of Computational Linguistics,[15] and there is a

14 Others have argued before for an empirically based Digital Humanities (for instance, see what Smithies calls 'Software Intensive' Humanities research in 23.), as well as for a strong alliance of Digital Humanities advocating for rather than in contrast with the Humanities (for instance, http://4humanities.org/).

15 See, for instance, the Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language (4REAL http://4real2018.di.fc.ul.pt) and the Workshop on Data

growing tendency by scientific journals to support sustainable research, asking submissions that involve data and/or code not already publically available to submit these along with the papers.[16]

The motivation for the increased interest in replicability is to be found in a number of factors, including the realization that for some published results, their replication is not being obtained (e.g. 21.; 1.); that there may be problems with the commonly accepted reviewing procedures, where deliberately falsified submissions, with fabricated errors and fake authors, get accepted even in respectable journals 2.; and that the expectation of researchers vis a vis misconduct, as revealed in inquiries to scientists on questionable practices, scores higher than one might expect or would be ready to accept 13..

Despite such interest in the issue, a general consensus about the very meaning of the main terms connected to the question of replicability is still missing. These terms are "repetition", "replication", "reproduction" and "reuse". Moving from repetition to reuse means going from confirming the reliability of the output results of a specific experiment to exploiting the scientific results of various research works.[17]

Repetition involves running the exact same solution or approach under the same conditions in order to arrive at the same output result. The conditions to keep the same are the environment of the experiment (e.g. the same lab), the workflow and the execution settings. One of the aims of repetition is providing reviewers with a proof of the reliability of the results reported in a paper describing a research work.

Replication entails arriving at the same overall conclusions. The aim is not to achieve the same output results, but to appropriately validate a set of results, by replicating the same answer to a given research question by different means (i.e. with a certain degree of variation), e.g. by reimplementing an algorithm or evaluating it on a new dataset. The aim of replication is to evaluate the robustness of a given research process.

Reproduction means running the same experiment with a different set up. The aim is not to get to the same output results but to get the same scientific results. It differs from replication in that it does not only entail a variation in workflow, execution settings and environment, but it makes use of different workflows, execution settings and environment. The aim is to evaluate if a different output is in accordance with the scientific results of an experiment with a different set up.

Provenance and Annotation in Computational Linguistics 2018 (WPACL https://typo.uni-konstanz.de/dataprovenance/).

16 One example is the forthcoming special issue of *Computational Linguistics* on "Computational Approaches in Historical Linguistics after the Quantitative Turn" (guest-edited by Taraka Rama, Simon J. Greenhill, Harald Hammarström and Gerhard Jäger).

17 The meaning of these four terms reported here echoes that presented by Sarah Cohen Boulakia in her speech at WPACL (*Scientific Workflows for Computational Reproducibility: Experiences from the Bioinformatics domain, Status, Challenges, and Opportunities*). See Cohen-Boulakia et al. 12..

Reuse gets not only to different output results but also to different scientific results. The experiment is different, as just some parts of it remain the same. These may be tools, scripts, data, workflows, results of an experiment that are reused in another one. The aim is to exploit the results of one or more previous experiments to move one step further in science.

Such growing interest in the different layers of replicability of results in research is reflected also in the fact that Horizon 2020[18] considers Data Management Plans (DMPs) a key element of good data management, describing the data management life cycle for the data to be collected, processed and/or generated by a H2020 project. A particular focus is put on making research data findable, accessible, interoperable and re-usable (FAIR).[19] Following Wilkinson et al. (25., 4), making (meta)data FAIR means that,

- They are assigned a globally unique and persistent identifier (Findable);
- They are retrievable using an open, free and standardized communications protocol (Accessible);
- They use a formal, accessible, shared, and broadly applicable language for knowledge representation (Interoperable); and
- They are released with a clear and accessible usage license and are associated with detailed provenance (Reusable).

According to the *Guidelines on FAIR Data Management in Horizon 2020* (page 4), to make data FAIR, a DMP should include information on,

- The handling of research data during and after the end of the project;
- What data will be collected, processed and/ or generated;
- Which methodology and standards will be applied;
- Whether data will be shared/made open access; and
- How data will be curated and preserved (including after the end of the project).

The H2020 requirement of repetition, replication, reproduction and reuse of results also in projects in the Humanities is good news for the research area, as it stems from a need of the community emerging in a bottom-up fashion. Indeed, the above mentioned lack of careful curation of the quality of both input and output data of processes of automatic natural language processing is mitigated by the growing spread of this interest for replicability of results also in the Humanities. In some way, this is a kind of natural turn that the Humanities must take, if they want to exploit at best the amazing benefits that can come from the application of computational methods and tools and, at the same time, if they do not want to separate into

18 Horizon 2020 is the biggest EU Research and Innovation program ever, with nearly €80 billion of funding available over 7 years, from 2014 to 2020. https://ec.europa.eu/programmes/horizon2020/.

19 See the Guidelines on FAIR Data Management in Horizon 2020 here: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

two parties that do not talk to each other, thus becoming weaker.

Today the *Index Thomisticus* (and, more generally, Latin) is on the front line in keeping the FAIR principles, thanks to its forthcoming inclusion in the LiLa Knowledge Base of linguistic resources and Natural Language Processing tools for Latin (https://lila-erc.eu/), which are connected via an explicitly-declared vocabulary for knowledge description fitting the Linked Data paradigm. LiLa is currently under construction; its building is funded by a two million euros ERC-Consolidator Grant (2018-2023), which shows how much the development of such infrastructures is considered essential and supported at international level.

For them to be Findable, the (meta)data collected by LiLa are assigned a globally-unique and persistent identifier (Uniform Resource Identifier = URI). In LiLa, identifiers are meant to capture the different degrees of granularity of the (meta)data available in the resources, ranging from the most generic (e.g., the type of resource) to the most specific, such as the single occurrence (token) of a word (type) in a text, for instance taken from the *Index Thomisticus*.

To make (meta)data Accessible by their identifier, LiLa uses the widely adopted SPARQL Protocol (https://www.w3.org/TR/rdf-sparql-query/), an RDF query language able to retrieve and manipulate data stored in the RDF format (https://www.w3.org/RDF/). In LiLa, (meta)data are made Interoperable through authoritative references to other (meta)data. These are provided in terms of explicit relations between the objects of the Knowledge Base, with the aim of helping users to identify and discover the (meta)data.

Finally, to make LiLa's (meta)data Reusable, these are described with a plurality of accurate and relevant attributes about the data usage licence and their provenance.

The FAIR guiding principles for data management should be welcomed in the Humanities and LiLa is one project applying the principles. As the Digital Humanities experiment with big data approaches they should learn from the *Index Thomisticus* and follow the careful approach to data pioneered by Busa. His legacy is not only in the innovative uses of technology, but also in the humanistic attention to data even when overwhelmed with it.

Acknowledgments

Marco Passarotti gratefully acknowledges the support of the project LiLa (Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin). This project has received funding from the European Research Council (ERC) European Union's Horizon 2020 research and innovation programme under grant agreement No 769994.

References

1. Begley, C. Glenn, and Lee M. Ellis. 2012. "Drug development: Raise standards for preclinical cancer research." *Nature* 483:531-33.

2. Bohannon, John. 2013. “Who’s Afraid of Peer Review?” *Science* 342 (6154):60-65.

3. Busa, Roberto. 1951. *S. Thomae Aquinatis Hymnorum Ritualium Varia Specima Concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*. Milano: Fratelli Bocca.

4. Busa, Roberto. 1958. “The Use of Punched Cards in Linguistic Analysis.” In *Punched Cards: Their Applications to Science and Industry,* edited by R. S. Casey, J. W. Perry, M. M. Berry and A. Kent. 357-73. New York: Reinhold Publishing.

5. Busa, Roberto. 1962. “L'Analisi linguistica nell'evoluzione mondiale dei mezzi d'informazione.” In *Almanacco Letterario Bompiani 1962*, edited by S. Morando. 103-8. Milano: Bompiani.

6. Busa, Roberto. 1974-1980. *Index Thomisticus*. Stuttgart-Bad Cannstatt: Frommann-Holzboog.

7. Busa, Roberto. 1980. "The Annals of Humanities Computing: The Index Thomisticus." *Computers and the Humanities* 14 (2): 83-90.

8. Campbell-Kelly, Martin. 2003. *From Airline Reservations to Sonic the Hedgehog: a History of the Software Industry*. Cambridge, MA: MIT Press.

9. Casey, Robert S. and James W. Perry (1951). *Punched Cards: Their Application to Science and Industry*. New York: Rheinhold Publishing.

10. Cecire, Natalia. 2011. "When Digital Humanities Was in Vogue." *Journal of Digital Humanities* 1(1): 54-9.

11. Ciula, Arianna. 2017. "Digital palaeography: What is digital about it?" *Digital Scholarship in the Humanities* 32 Suppl. 2:89–105.

12. Cohen-Boulakia, Sarah, et al. 2017. "Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities." *Future Generation Computer Systems* 75: 284-98.

13. Fanelli, Daniele. 2009. "How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data." *PloS ONE* 4 (5):e5738.

14. Jones, Steven E. 2016. *Roberto Busa, S. J., And the Emergence of Humanities Computing: The Priest and the Punched Cards*. New York: Routledge.

15. Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Los Angeles: SAGE.

16. Lubar, Steven. 1992. ""Do Not Fold, Spindle or Mutilate": A Cultural History of the Punch Card." *Journal of American Culture* 15(4):43-55.

17. Mackenzie, Charles E. 1980. *Coded Character Sets, History and Development*. Reading, MA: Addison-Wesley.

18. Newman, James A. (2012). *Best Before: Videogames, Supersession and Obsolescence*. New

York: Routledge.

19. Nyhan, Julianne and Terras, Melissa. 2017. “Uncovering ‘hidden’ contributions to the history of Digital Humanities: the Index Thomisticus’ female keypunch operators.” Paper presented at Digital Humanities 2017, Montréal, Canada.
20. Passarotti Marco Carlo. 2003. “One Hundred Years Ago. In Memory of Father Roberto Busa SJ.” In *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3),* edited by Mambrini, F., Passarotti M., Sporleder C. 15-24. Sofia: Bulgarian Academy of Sciences.
21. Prinz, Florian et al. 2011. “Believe it or not: how much can we rely on published data on potential drug targets?” *Nature Reviews Drug Discovery* 10 (9): 712.
22. Rockwell, Geoffrey and Stéfan Sinclair 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, MA: MIT Press.
23. Smithies, James. 2017. *The Digital Humanities and the Digital Modern*. Berlin: Springer.
24. Tasman, Paul. 1957. "Literary Data Processing." *IBM Journal of Research and Development* 1 (3):249-56.
25. Wilkinson, Mark D., et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data*. 3. https://doi.org/10.1038/sdata.2016.18
26. Winter, Thomas Nelson. 1999. "Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance." *The Classical Bulletin* 75 (1):3-20.

Last URLs access: 08/05/2019