DHQ: Digital Humanities Quarterly

2020
Volume 14 Number 3

Tremendous Mechanical Labor: Father Busa's Algorithm

Geoffrey Rockwell <grockwel_at_ualberta_dot_ca>, University of Alberta
Stéfan Sinclair , McGill University

Abstract

This paper looks at innovations in Busa's *Index Thomisticus* project through the tension between mechanical and human labour. We propose that Busa and the IBM engineer Paul Tasman introduced two major innovations that allowed computers to process unstructured text. The first innovation was figuring out how represent unstructured text on punched cards, the way data was encoded and handled at that time. The second innovation was figuring out how to tokenize unstructured phrases on cards into words for further counting, sorting and concording. We think-through these innovations using replication as a form of media archaeology practice that can help us understand the innovations as they were thought through at the time. All this is framed by a letter found in the Busa archives criticizing the project as a "tremendous mechanical labour ... of no great utility." We use this criticism to draw attention to the very different meshing of human and mechanical labour developed at Busa's concording factory.

> 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!) [1]

Early in 1950, Daniel L. McGloin, the Chair of Philosophy at Loyola University of Los Angeles (now Loyola Marymount) wrote a letter critical of the idea of the *Index Thomisticus* project that Father Busa had been promoting in order to get support. We know from the correspondence preserved in the Busa Archive that Busa had sent out letters describing the project and soliciting letters of support. While most correspondents did write supportive letters, at least one did not. McGloin's review of the project was for his Rector who enclosed it in his note back to Busa. The project was, according to McGloin, "a sort of fetish of scholarship gone wild". It was the one negative response we have on record and it is clear that McGloin was appalled. 1

What exactly bothered McGloin so much? What can we learn from the criticism of what is arguably the first digital humanities project? In what could be the first instance of a long tradition of pointing out the dark side of automation in the humanities, McGloin the philosopher was horrified by what he perceived to be the "mechanical verbalism which would tend to deaden rather than revivify the thought of St. Thomas." He could see the value of a concordance of philosophical and theological terms, but not of *all* words. He thought this was "of no great utility", especially considering the "tremendous mechanical labor it would involve."[2] In criticizing Busa's project he played with a contrast between the living (and the labor that revivifies thought) against the mechanical (and the mindless work that deadens).[3] In doing so 2

he alerts us to a tension in how digital methods change the actual work of the humanist: who does the work and how technologies are woven into our practices.

This paper sets out to understand that tension by revisiting the mechanical labor of Busa's project, or more precisely, by trying to replicate exactly how automation was developed for concording Aquinas. Replication here is a praxis of media archaeology that can help us understand the context of a technology as understood in its time [Zielinski 2008] [Parikka 2012] by trying to imitate what was done. Needless to say, to understand the mechanical labor we also have to touch on the human labor and how humans were integrated, though the human labor is not our primary focus here. 3

From the comfortable distance of more than half a century later, living as we do in an epoch defined by language engineering companies like Alphabet (Google), we can say that Busa and his IBM-based collaborator Tasman proved to be prescient, despite McGloin's complaints. No one today would question the usefulness of the tremendous language processing techniques that are currently used, and that is part of what makes it harder to understand the work that went into actually developing ways for processing textual data back then. What were the mechanical processes? What was the data processed and what processing techniques allowed Busa and Tasman to generate a "First example of word indexes automatically compiled and printed by IBM punched card machines" (the subtitle of Busa's *Varia Specima*)? Further, how was what we today would call an algorithm for language processing conceived? For better or worse, these techniques initiated a new era of mechanical and eventually computer-assisted language engineering. It is therefore worth recovering an understanding of what was achieved, and we will do that in four moves. 4

- First, we will briefly discuss the historiography around Busa's project. What resources do we have access to if we want to understand the techniques developed then? Where and how did he discuss this project?
- Second, we will focus on the punched card data format that they developed. How was text chosen and encoded for this project?
- Third, we will look at the key tokenizing process that produced word cards from phrases. We hope to convince readers that this was a significant innovation for its time.
- Fourth and finally, we will return to McGloin's concerns around deadening labor.

Historiography of a Project

If projects are central to digital humanities practice [Burdick et al 2012, 124], then we must recognize that we have largely neglected that aspect in the considerable literature on digital humanities as a discipline, and it is time to ask how we can study projects. Concretely, for Busa's project, a first step would be to do a historiography of the *Index Thomisticus* project and the host of the project, the GIRCSE[4] centre. To understand ourselves in our paradigmatic project practices we need to ask how one studies projects and labs when their scholarly outputs tend to hide (deliberately or otherwise) their labor and infrastructure. 5

In the case of GIRCSE and the *Index Thomisticus* we are unusually fortunate to have four types of resources in addition to the set of publications where Busa and Tasman explicitly discuss their methods. First, we have the extensive Busa Archive at the Library of the Catholic University of the Sacred Heart in Milan (Biblioteca dell'Università Cattolica). Busa's team organized and kept an extraordinary record of unpublished documents related to the project, from sales correspondence to every one of Busa's publications.[5] These are now being catalogued by the Library of the Catholic University. Some of the key documents from the Archive include letters like the one from McGloin and others where Busa described his method in different contexts. There are also sample punched cards; a "Flow Chart"; photographs, including some of the people at work on the project on the ex-factory floor; and a project proposal.[6] 6

A second type of important resource is secondary sources like Steven Jones' book *Roberto Busa, S. J., And the Emergence of Humanities Computing: The Priest and the Punched Cards* (2016). Books like this provide background context including many of the resources, key events, and cultural contexts.[7] It is also the best guide into understanding the origins of the project. Another valuable work is Winter's article "Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance" (1999) which defends the claim that Busa was the first to automate concording. 7

Thirdly, there are oral testimonies by those who worked on the project or knew of it and related materials. Julianne Nyhan and Melissa Terras have been gathering information about the women who worked on the *Index Thomisticus* project.[8] Nyhan traveled to Italy and interviewed many of the women operators; she has a book forthcoming on *Uncovering "hidden" contributions to the history of Digital Humanities: the Index Thomisticus' female keypunch operators.* 8

Finally, there are the publications by the project leaders like Busa and Tasman describing the methods they were developing. Julianne Nyhan and Marco Passarotti are putting together an edited and translated collection of some of Busa's key papers which will make it much easier to study Busa's published thought. The key publications used in this paper are, in chronological order: 9

1. The *Varia Specima Concordantium* from 1951 which is a concordance of the poetry of St. Thomas Aquinas that was the "First 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 innovating process.
2. Tasman's 1957 article in an IBM trade journal on "Literary Data Processing."
3. Busa's chapter from 1958 on "The Use of Punched Cards in Linguistic Analysis" in a larger collection on *Punched Cards.*
4. Busa's project proposal from 1962 to IBM for completing the *Index* in time to show it at the New York World's Fair in 1964, "Per Completare Lo Index Thomisticus Per L'Esposizione Mondiale Di New York 1964 – 1965" (For Completing the Index Thomisticus For the New York World's Fair 1964-1965).[9]

In all of these there is a summary of the technical process they developed, a clear indication of their awareness of the importance of the developing process at the time. In the introduction of *Varia Specima* Busa summarizes this process as five stages and this summary is worth quoting as a description of a workflow of the labor involved. 10

> I bring down to five stages the most material part of compiling a concordance:
>
> 1. transcription of the text, broken down into phrases[10], 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)

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 he also talks (p. 26) about the need to have a philologist intervene at stage 3 to disambiguate and lemmatize words. In Paul Tasman's "Literary Data Processing" (1957) the stages outlined are different. Instead of one step for "transcription of text" he has two steps, that of the scholar who marks up the text and that of the keypunch operator who copies it onto cards. This more clearly separates the scholarly from the data entry work. 11

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)

In "The Use of Punched Cards" Busa describes yet another process. Here is the flow chart that Busa provides in "The Use of Punched Cards" (1958, 359). It focuses on who is doing the work: 12

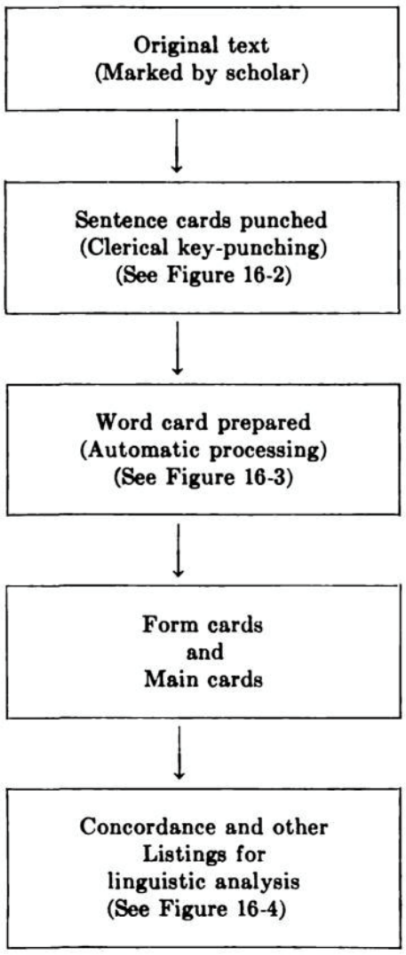


Figure 16–1. Summary of operations.

Figure 1. Flow chart from "The Use of Punched Cards"

In this flow chart from 1958 we again have two operations at the start – that of the scholar marking the original text and that of the entry of the phrases by a key-punch operator. What is new is that the last two stages of 1951 are represented by one operation in 1958, but it is clear that the one box covers many processes for different types of indexes.

One thing that stands out in this process, especially when one looks at other materials like the photographs of the space where the work was done, is how many people were needed for what we today think of as an automated process. The machines may have done some of the tedious work, but it took human labor to prepare texts for data entry, then there was the data entry itself, there was programming, there was the moving of stacks of cards from machine to machine, and then were all the management tasks. The 1962 project proposal gives a break down of the staff needed, admittedly for an accelerated completion. The number of staff proposed, divided by department is listed as: 13

- Machine Department: 33 staff (likely to be mostly women operators)
- Research Department: 21 staff
- Programming Department: 7 staff
- Management and Services: 9 staff

This is a total of 70 staff for a project that was supposed to be showing the advantages of automation! What also stands out is the gendering of the departments. The operators of the Machine Department, which was the largest proposed unit, were mostly women. As Terras points out in her Ada Lovelace day blog post (2013), it was women who actually ran the machines and they are often shown in the photographs demonstrating the computing to men. 14

Returning to the processes that were automated — and looking at the details of the descriptions Busa and Tasman give 15

us — two key innovations stand out that made linguistic data processing possible.

1. First, they developed a way of representing a text fully so that it could be processed by machines, and
2. Second, they developed a process for tokenizing the text — generating words cards with the electro-mechanical machines of the day.

These innovations changed the integration of people and technologies that went into concording, and for that matter, changed how data was processed (previously cards had been used for structured data such as census information, but not for unstructured linguistic data). It seems obvious now, but these innovative uses of machines made possible a different organization of people and technology where data entry was separated from programming which was separated from the scholarly work. Technologies, as philosophers Ihde (1998) and Winner (1980) remind us, are not neutral tools. McGloin may have intuited how Busa's project would still call for tremendous human labor despite the machines and that much of that labor would be mechanical for the women involved.[11] 16

In sum, Busa's team conceived a new way to break down the work of concording by first figuring out how to represent text as data that could be processed by machine and then figuring out how to process that data into tokens (words) that could then be manipulated to generate various types of indexes. In these two connected innovations they essentially developed literary data processing, no small feat at the time. 17

The next two parts of this paper will look at these two stages and discuss the replication of what they might have actually done as a way of probing our understanding. 18

> This is a story from early in the technological revolution, when the application was out searching for the hardware, from a time before the Internet, a time before the PC, before the chip, before the mainframe. From a time even before programming itself. [Winter 1999, 3]

Punched Cards

From today's perspective it is hard to appreciate how different the *data* of "data processing" was in those early years. Busa could not go onto the internet or scan the texts he needed. All input had to be punched by hand — there was no "data", nothing given. There weren't any standard ways of representing texts for data processing and punched cards were limiting. He wasn't even using what would today be considered a digital computer. The *Varia Specima* project (1951) which demonstrated the concept used electro-mechanical machines to sort, replicate, and print on cards. This was a liminal moment between mechanical data processing instruments with pre-defined functions and digital, programmable computers with punched cards in common as a way to enter, store, and record data and results. 19

The one technology that had been at least partially standardized was the punched card as a way of entering and storing data. The punched card, despite later discussion about folding, spindling, or mutilating, was a robust way of entering information so that it could be processed manually *and* by computers.[12] As a data technology it goes back to the Jacquard loom and is still used today, such as in some voting machines like those used in Florida in the 2000 presidential election between Bush and Gore. 20

The Busa project used the IBM card format. This had been developed in 1928 and had become a de facto standard. Each card was 7⅜ inches wide by 3¼ inches tall. They were made of stiff paper with a notch in the upper left for orientation. The arrangement of holes was standardized so that all relevant machines could process them. Given the size, the IBM card could fit 80 columns and 12 rows of punch locations. While the dimensions and punch zones were standardized, projects could overprint with ink whatever they wanted on the cards as what was printed wouldn't affect the data processing; printing was intended for human processing, not machine processing. 21

Figure 2. Detail of *Index Thomisticus* card from the Busa Archive

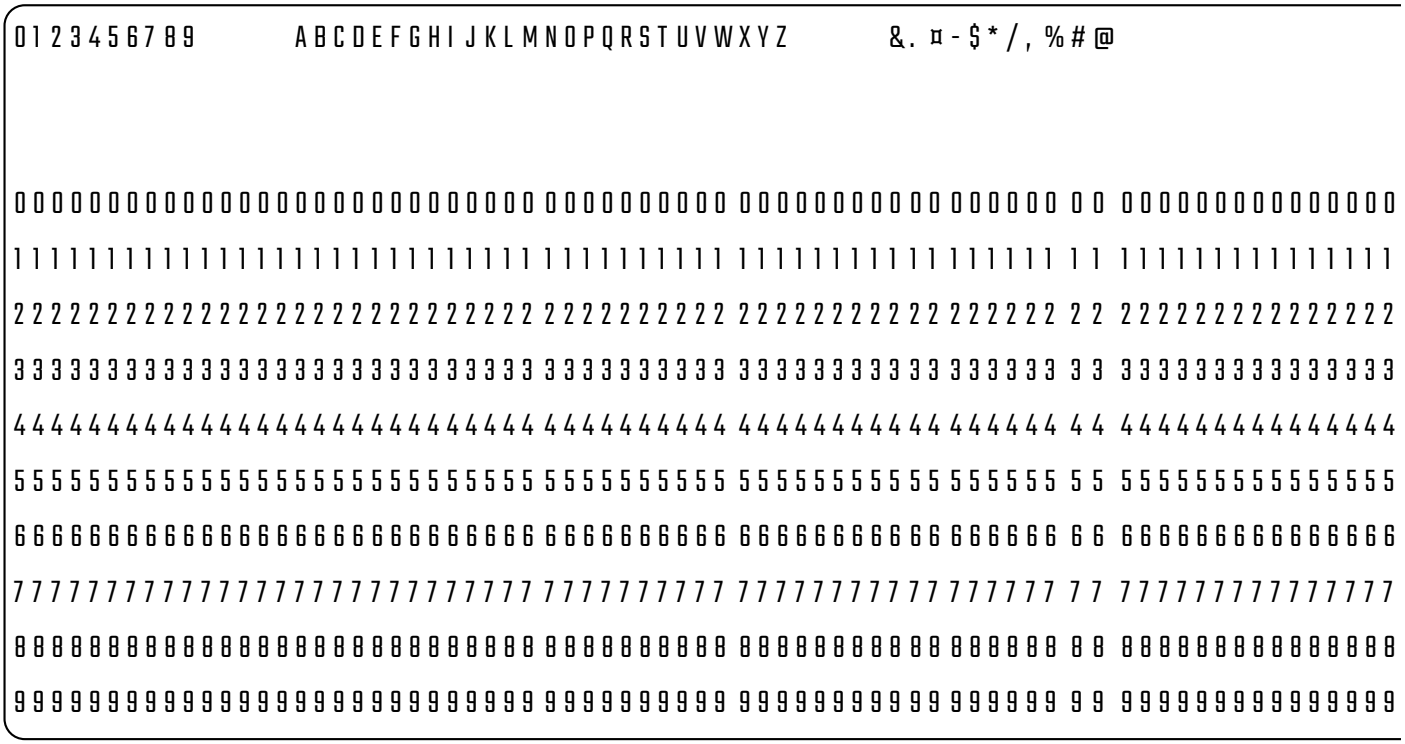
It is also worth noting that Busa's project was large and different enough that they had their own non-standard cards, including some with bubbles for manual pencil marks. Figure 2 shows an example card from the Busa Archive with the areas and labels related to "Philological Analysis." These printed zones could have been used by scholars to manually add annotations as needed for human sorting as part of hybrid machine/human processing. Or they could have been used for easily reading the punched data. Either way, it was common for punched cards in those days to have to be handled by both humans and by machines [Casey and Perry 1951]. The *Index Thomisticus* project was no different; there would have been a tremendous amount of human labor associated with handling the stacks of cards. The human labor would have gone into preparing cards, moving them, checking them, annotating them and even hand sorting them. It would be interesting to try to recreate the flow of Busa's hybrid processes, but that is for another project.[13]

Simple Punch Card Emulator

In the interface below you can type characters into the text box (they will be removed if they're not supported), or, for an interesting challenge, you can try "punching" the characters in yourself by clicking on the boxes.

type word:

```
0123456789      ABCDEFGHIJKLMNOPQRSTUVWXYZ      &.¤-$*/,%#@
```



(CC-BY) Stéfan Sinclair & Geoffrey Rockwell

23 To return to understanding what Busa may have done in the way of digitally representing textual data we have tried to virtually replicate the data dimension of his punch cards with a simple interactive punch card that shows you what holes would be punched in order to represent the data you want the card to bear (see below). This is set up with our best guess of the card data format they used — feel free to enter examples and see if you get what you expected. We have here used replication as a way of thinking through the history of computing. In the replication of a technology we can discover aspects of the technology that wouldn't occur when just documenting it and others can test our hypothesis as to how data was encoded. Among other things the punched card interactive helped us better understand how this encoding system was actually a curious blend of a decimal-based system (0 to 9 with a couple of extra control rows) and the more binary-based digital processing (0 or 1). A typical punched card had 12 rows to represent any one character, which in fact is potentially very high resolution (12-bit = 4096 possibilities), whereas the first character sets for computers were at most 4, 5 or 6-bit (64 possibilities).

24 In this case we were forced to figure out exactly which IBM format was used. Emulating the cards raises the issue of *what* data Busa could encode and *how* it was encoded on cards. Our best guess, looking at example cards in the Archive, and reading descriptions, is that they used something like the 48-character BCDIC (Binary-Coded Decimal Interchange Code). We know that by the early 1930s there was a 40-character BCDIC format that combined more than one hole per column to provide 40 characters (Space, 0-9, A-Z and a couple of special characters.) By the late 1950s

BCDIC had been extended and the IBM chain printers were capable of printing 47 graphics (characters) along with a space to make the 48-character BCDIC with 11 special characters, namely: .'$,#%-&*/ ¤ . Keypunch equipment was also becoming available so users that could input any of these characters with one key stroke, as opposed to having the operator select the punch hole combinations for each character [Mackenzie 1980, 88]. From the Tasman article we know that the project used at least /, #, ‘, and ¤. [Tasman 1957, 254] which is why our guess is that the project was using something close to 48-character BCDIC.

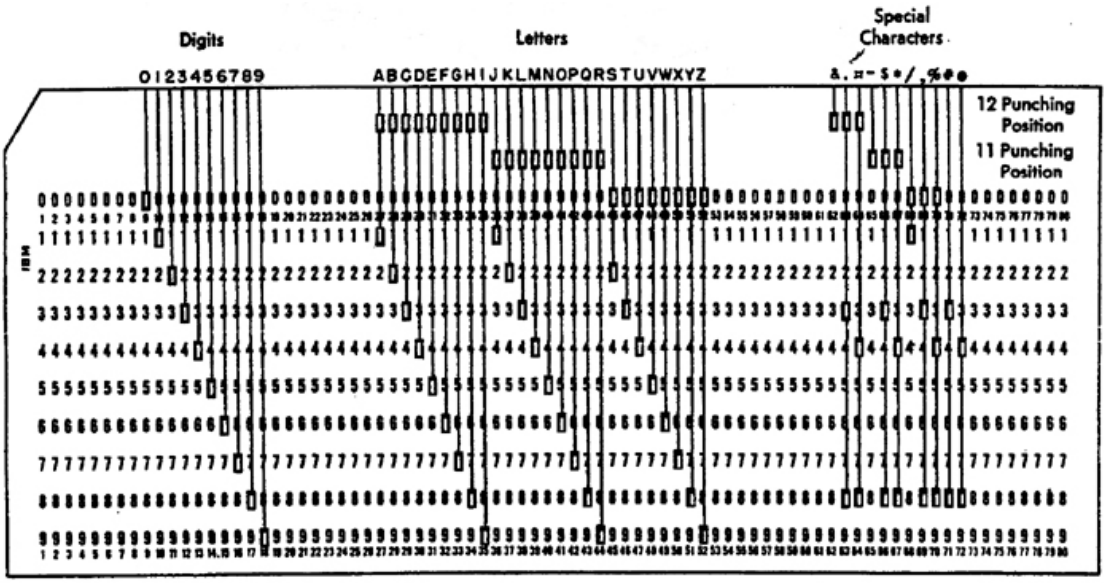


Figure 2. Punching Positions in Card

Figure 3. Punching Positions for 48-character BCDIC[14]

It is worth noting however, that they did not need to adhere to any particular standard the way we need to today. There was no operating system or off-the-shelf software enforcing standards. There were no other projects that might use their data. The only issue was what the keypunch could enter, what could be read, and what the printers could print, if printing was needed. At that time you could extend things and program the machines to operate on your own hole combinations. It could very well be that the *Index Thomisticus* project provided some of the impetus behind the extension to 48 character BCDIC as it would have been a major project defining IBM’s move into linguistic engineering. 25

Having explored some of the *how* of Busa’s work with punched cards, we will now turn to the *what* or purpose of the work, because that is where Tasman and Busa figured out a practical way to adapt the technology at hand to text processing. The key was realizing that they only needed two sets of cards for most literary/linguistic operations: a set of Sentence (or Phrase) Cards making up the whole text, with associated information, and a set of Each Word Cards (EWC) with all the words (tokens) and associated information. In today’s terms they needed only two database tables to generate the various forms of indexes they wanted; one was the full text, the second was a word index. In the processing they showed that you can go from one to the other using the one best suited for the analytic process. This is the precursor of analytical tools that cache a word index in addition to the text for efficiency purposes. 26

> 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. [Tasman 1957, 254]

What was actually held on the two types of cards? The sentence or phrase cards held the location/reference in the original text, the text of the phrase, a number for the phrase (serial number) and a special reference mark to indicate if the phrase was thought to be by St. Aquinas himself or a reference to words by another. The amount of text for each card, given the limitations of 80 columns (not counting the columns reserved for reference data), would be decided by the scholar who divided the text up into logical thoughts and then into meaningful phrases that would fit on a card. It is important to note that human intervention was needed to fit the ongoing unstructured text onto 80-column cards. 27

The EWCs had less text and more associated information. By the time of Tasman's article, each word card would have encoded in the punched holes the following (some fields are further explained below): 28

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

The Form Cards were the headings for each different word. The Entry cards were the entry headings that presumably would appear in the final printed concordance for the different word types after lemmatization and disambiguation (*dance* as a verb or as a noun would have different entries, for instance). These were created by a second intervention by scholars and would be the entries in any published concordance. In Figure 4 you can see where in the flow chart the Different Word Cards (or Form Cards) are generated along with the summaries and lists that allow the Scholar to write a list of Entry Words. 29

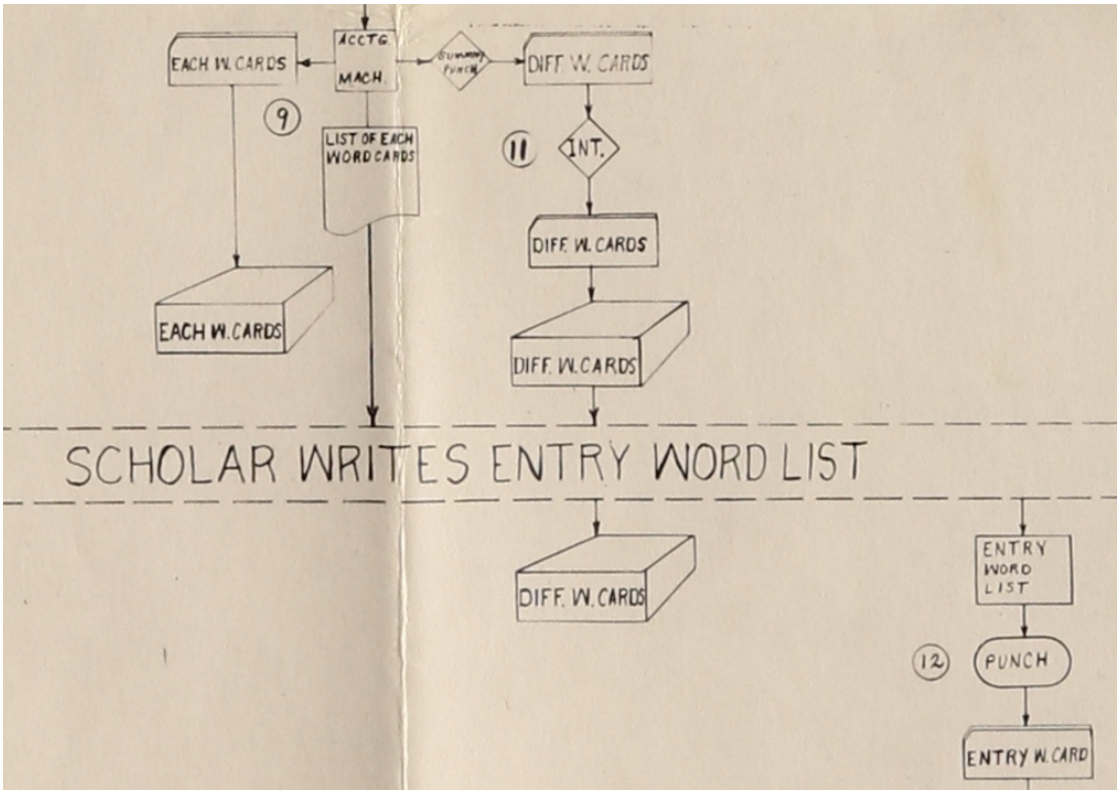


Figure 4. Detail from the 1953 "Flow Chart" from the Busa Archive

There are a few noteworthy aspects about the project's use of punched cards. First, data on cards was material in the sense that users could physically touch the carrier of the data — the punched card. Many of the processes from keypunching to sorting involved physically manipulating the cards themselves, not data in memory. The human was also part of the sequence of operations as someone would have to load cards, move stacks from machine to machine, and cards would actually be consulted by humans at certain points. Busa would have been aware that he was setting up a system that integrated people and machines in new configurations, at least for the humanities. 30

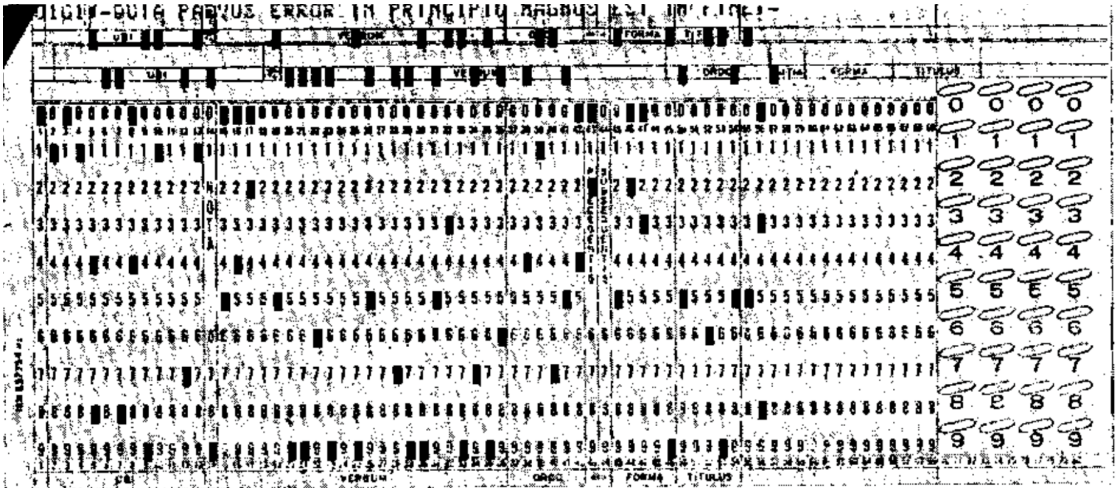


Figure 5. Image of EWC in Tasman 1957 with a “Mark sensing area for use of scholar in selecting, collating, printing, or revising information” (p. 250-1)

Busa and Tasman made full use of the materiality of the punched card. They actually had at least three orders of information on the cards. There was what we are going to call the data proper that was punched on the card and could be operated on mechanically. Then they would print (not punch) extra information on the front or back of the cards for human consultation. For example, they would print the data punched on the top of the fronts of the cards so that they could be checked by eye, or they could print up to 9 lines (phrases) of context on the backs of EWCs so a scholar could easily check the context when lemmatizing without having to call up the phrase card. Finally, they had an area on the left of cards that would take lead pencil marks so scholars could mark cards for additional processing.[15] In Tasman’s (1957) article we see that both the Sentence Cards and EWCs had such areas. 31

One way to appreciate the difference between data then and now is to reflect on how the project was not about using computers or software as we understand them now. The actual work involved punching information onto cards, proofing of cards, using electromechanical machines to process the cards, creating new cards, printing things on cards, moving cards from machine to machine, duplicating them, and marking them for further processing. The “programming” in this context was the design of the cards so that various sorting, replicating and counting processes produce what was needed. (It was also the wiring/configuring of the machines.) The cards enabled and constrained what could be done. The work of the project as described by Tasman was really the work of managing punched cards as tokens to be interpreted. The cards were a surrogate for the text that could be processed to produce a new concorded text. 32

Processing

Now to the second innovation of Busa and Tasman and that is the algorithmic one. It is fascinating to reconstruct the mechanical processes with which all this information could be generated and added to the cards in successive passes with a limited number of standard machines. For example, to add the first letter of each following card the stack of EWCs would be sorted in reverse order of appearance and then run through starting with the last word. The first character of each word would then be carried forward (or backward in this case) and written out on the next card, which illustrates another property of the material cards, the fact that data can be incrementally added to the blank areas with proper planning. We do not have space here to go through all the succession of processes. Instead, we will focus on the one process that made literary data processing possible, and that is the tokenizing of the Sentence Cards to generate Each Word Cards (EWCs). This is the crucial innovation that showed how systems designed for accounting could be used for unstructured text. 33

> This is equivalent to state that each line was multiplied as many times as words it contained. I must confess that in actual practice this was not so simple as I endeavoured to make it in the description; the second and the successive words did not actually commence in the same column on all cards.

> In fact, it was this lack of determined fields which constituted the greatest hindrance in transposing the system from the commercial and statistical uses to the sorting of words from a literary text. The result was attained by exploring the cards, column by column, in order to identify by the non-punched columns the end of the previous word and the commencement of the following one; thus, operating with the sorter and reproducer together, were produced only those words commencing and finishing in the same columns. [Busa 1951, 24/26]

This would have been considered difficult at the time. As Busa writes in 1951, the existing commercial accounting and statistical uses of punched cards assumed that there would be zones on the card that constituted predictable fields. These fields could hold alphabetic information like names in a personnel dataset, but the start of each name – the zone of the card where it appeared - would be known in advance, allowing processing operations to be easily set up.

By contrast, with a phrase of unstructured text only the first word on each card is in a predictable place and even then there is no information about where the field ends. Thus the problem they had to solve was how to explore the Sentence Cards over and over, identifying each successive word … and in 1951 they had to do this using only sorting machines with replicators to create the new EWCs. This was the key mechanical process that allowed one type of data — phrase cards – to be processed to generate another — word cards. Everything else could in theory then be generated using these two sets of cards (along with a bit of human work lemmatizing/disambiguating.) 34

To figure out how this could be done we have again replicated the process, though in code rather than with an emulated machine.[16] We imagined that one would start by exploring all the cards, sorting out those with a space in the second column. Those would be the cards with a one letter word at the start. These would be run through the replicator which would now know what zone the word to find was in and thus could create an EWC for the one letter word. Then one would sort the remaining cards to find those with a space in the third position and so on until one had explored every possible location of a word between two spaces. We suspect that this actually would have involved a lot of physical moving of stacks and also keeping track of the subsets of cards – not an entirely mechanical process, but one less prone to error and much faster than writing things out by hand. Tremendous semi-mechanical labor indeed. 35

Later they found other ways to do this. By 1958 Busa had actually developed three ways to do this crucial process. In “The Use of Punch Cards” (1958) Busa describes the 3 options (on page 361) as: 36

1. The keypunch operator could just punch individual word cards (whether in addition or instead of phrase cards.) Punching just the EWCs was the most efficient way if all you need was a word index.[17]
2. The keypunch operator could punch the Sentence Cards and then, using sorting and replicating machines, the EWCs could be machine generated. This is the approach described in the *Varia Specima* quoted above.
3. Or a last process “using the Cardatype, recently developed by IBM” [Busa 1958, 361]. This is the process that is diagramed in the 1952 “Flow Chart” prepared by PT and JEG, probably Paul Tasman and an unidentified draftsman. Figure 5 shows the relevant detail of the flow chart where Sentence Cards are handled by the Cardatype:

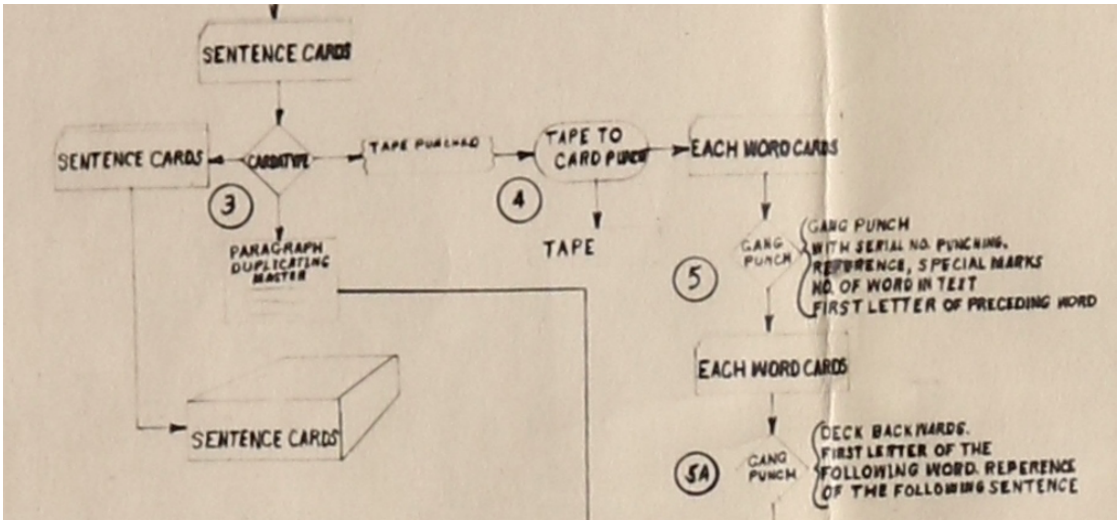


Figure 6. Detail from 1952 “Flow Chart” from the Busa Archive

The Cardatype was a modular and multi-function machine which makes it hard to figure out exactly what they did with it. It had one or more IBM electric typewriters, a verifier, and could have a tape-punching unit attached that would output a paper tape which could then control a card punch that would punch the EWCs [Stanford 1951, 403] There seems to have been enough intelligence in the Cardatype to output to paper tape the data needed to punch the EWCs. 37

Conclusions

To conclude we begin by noting that Busa provided answers, of sorts, to McGloin’s criticism. Father Busa in his 1980 reflection on the *Index Thomisticus* defended the concording of function words by arguing that “all functional or grammatical worlds … manifest the deepest logic of being which generates the basic structures of human discourse. It is this basic logic that allows the transfer from what the words mean today to what they meant to the writer” [Busa 1980, 83]. The paradigmatic example for Busa was the one he used to open his 1980 article: understanding *presence* through prepositions like *in* as he tried to do in his doctoral thesis. Busa the philologist valued all the language of Aquinas as a way of understanding the historical context. McGloin wanted to focus on the philosophy and theology which he felt needed to be brought to life and made relevant to the modern world. This is an old tension in the humanities [Turner 2014]. McGloin may not have objected to the concording of function words *per se*; rather it was amount of work that he expected to be tremendous and useless. One wonders if Busa could have convinced him of the philosophical value of concording prepositions like *in*? That would not, however answer McGloin’s concern about the amount work required. As mentioned, McGloin probably (and rightly) inferred that the tremendous mechanical work would also mean tremendous human labor to set it all up and then deal with the results. And it should be noted that the *Index Thomisticus* project did involve an extensive human crew working in a converted factory space for decades [Jones 2016]. McGloin probably suspected that Busa was proposing a concording assembly line that would not be philosophically rewarding work but mechanical work that could have been better dedicated to something else. 38

A different sort of answer can be found in an article by Paul Tasman, Busa’s IBM engineering collaborator. He believed that innovations like the machine-searching application developed by the project “may initiate a new era of language engineering.” 39

> The indexing and coding techniques developed by this method offer a comparatively fast method of literature searching, and it appears that the machine-searching application 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. [Tasman 1957, 256]

Tasman saw the business potential for IBM in the innovations of the *Index Thomisticus*. While Thomas J. Watson Sr. 40

(the CEO of IBM) may have been skeptical of the project when he first met Father Busa [Busa 1980], he ended up supporting it and it paid off in terms of innovation. In this Tasman was right. The language engineering innovations of the project underlie the digital turn of our age where not just language, but multiple media are either born digital or digitized for engineering on the computer. The computer has become the universal machine for processing information in any form. The project was the first step away from using the computer solely for numbers and towards using it for other forms of information. It should also be noted that IBM took advantage of the publicity that the Busa project offered, including it as part of their display at the 1964 New York World's Fair (see [Jones 2016]).

41 Having a sense of some of the tremendous mechanical labor involved in Busa's project, something McGloin didn't really have when commenting, we can conclude by asking again if such labor was worth it. Obviously, from an efficiency point of view, it was worth it if you grant that Busa was right about the value of a full concordance. Tasman [Tasman 1957, 255] estimated that mechanizing concordancing would reduce the time needed for the *Index Thomisticus* to a tenth of what a manual process would take, with a fifth of the people. One could also argue, as Tasman does, that Busa's project not only produced a concordance but also developed innovative ways of ways of using the technologies at hand for "language engineering." These developments we now know led to much more than concordances. They led to a whole alphabet of services. But, to be fair to McGloin, he was only reviewing the usefulness of the *Index Thomisticus* for studying Aquinas, not reviewing the potential of the technology. McGloin argued that focusing on the technology would become a fetish of its own, distracting us from the interpretation of Aquinas, and he was right. But, would such tremendous labor not make it easier for others to interpret Aquinas? To claim that some words are more important and therefore more worthy of indexing for research seems to betray a limited imagination of what research could be. McGloin's view, in retrospect, is conservative when it comes to imagining research usefulness but prescient if he meant to also alert us to reconfigurations of labor and machines.

> Big Data has emerged a system of knowledge that is already changing the objects of knowledge, while also having the power to inform how we understand human networks and community [Boyd and Crawford 2012, 665].

42 McGloin was right to focus on labor and the changes in attention, what he calls fetish, though we disagree with his unquestioned association of mechanism with a loss of life. Changes in the labor of research change the object of research and the methods with which we think-through (and vice versa). Busa and Tasman were aware of this in that they discussed some of the implications of the new methods they were developing and could see some of the possibilities for different types of research and answers. McGloin is right to ask about these changes, but wrong to assume there is something necessarily dead about mechanised methods while the existing objects and methods are sacred. McGloin poses the critical question about what and how we are interpreting, but he is a romantic who does not question his own practices. As boyd and Crawford argue for big data (2012) "it is time to start critically interrogating this phenomenon, its assumptions, and its biases" (665). We would add that to critically interrogate a new phenomenon one needs to also ask about that with which it was contrasted in its time and also ask about the metaphors of contrast. One of the best ways to interrogate such a phenomenon of literary and linguistic computing is to do the archaeology of the technologies that actually changed the labor, objectives and methods, and that is what we have tried to do here. Much can be brought to life in the replication of liminal moments like Busa's *Index Thomisticus* project.

43 Following boyd and Crawford, an aspect of the labor that McGloin might have objected to if he had had more time would have been the factory-like management systems developed for this big data project and the ways they might change what is considered knowledge in the humanities. Busa and Tasman didn't just develop new technologies, they had to develop a physical center where the work could take place and an organization capable of carrying out so much human and mechanical labor. This wasn't the first large scale humanities project, there had been large projects for some time, including concording projects, dictionary projects, archaeological projects, and editorial projects. It was, however, one of the first projects to integrate information technology and human labor so closely. The project needed deliberate planning, fund raising, management, public relations, and training in the use of computers. Jones (2016) has gone a long way towards documenting how this was the first humanities computing project. Nyhan and Terras have documented the gendered division of labor between male scholars and female punch card operators.[18] Jones leads a web project "Reconstructing the First Humanities Computing Center" to which many have contributed that is

reconstructing the physical space in Gallarate of the Centro per L'Automazione dell'Analisi Letteraria (Center for the Automation of Literary Analysis) or CAAL.[19] Rockwell and Passarotti (2019) have looked at the project as a project.

From today's perspective McGloin might have commented on how the challenge of dealing with so much text led to the development of a new organizational model for humanities projects where the integration of technology forced a new division of labor on the humans. It forced humanists to develop practices which integrated scholars with punch card operators, technical staff and engineers. We are still struggling with issues of credit and project management around these new configurations. These new configurations change the communities of knowledge and may have also changed the way knowledge is conceived. Information, often in the form of large amounts of data, has become synonymous with knowledge. The datafication of the cultural record has changed our thinking about knowledge. Scholarship has given way to information management. 44

Replications

- Punched Card Emulator Blog: http://stefansinclair.name/punchcard/
- Simple Punched Card Emulator: https://cdn.rawgit.com/sgsinclair/epistemologica/master/punchcard.html
- Busa Operations: http://nbviewer.jupyter.org/github/sgsinclair/epistemologica/blob/master/PunchcardOperations.ipynb

Image Credits

All images from the Busa Archive are property of the Biblioteca dell'Università Cattolica. They are shared in this presentation with permission. For further information, or to request permission for reuse, please contact Paolo Senna at paolo.senna@unicatt.it. 45

Acknowledgements

This paper was originally presented at Loyola, Chicago in 2016. Rockwell and Passarotti also presented some of these ideas as part of a different paper in Rome, Italy in 2017 at the 6th AIUCD Conference. That paper was published in 2019 under the title "The Index Thomisticus as a Big Data Project." 46

Notes

[1] Letter from Daniel L. McGloin S.J., Chairman of the Philosophy department to the President/Rector at Loyola University of Los Angeles. The date would have been before February 14th, 1950 when the President/Rector, 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.

[2] Jones (2016) discusses this letter in Chapter 3 of his book. He points out that McGloin believed the project was "insufficiently interpretative — a complaint that would continue to be brought against computing in the humanities for decades and, indeed, continues today to figure in critiques of the digital humanities." (Location 2336)

[3] McGloin's judgement of Busa's project can be understood as an extension of an essay he published while in graduate school in 1945 about "Revitalizing Liberal Education." (Note the "vital" language in the title.) In this he complained about the technological materialism guiding life at the end of the war. "Unconsciously, we have come to view life after the analogy of an assembly-line. We construct an educational system as we blueprint an efficient factory, which is an aggregation of machines and operators. Bring your material in, run it through the machines, and out comes a tank." [McGloin 1945, 135]

[4] GIRCSE or the Gruppo Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (Interdiscplinary Group for Research into the Computerization of Expressive Signs) was set up by Busa in the 1970s. GIRCSE evolved into CIRCSE (Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione — Interdisciplinary Centre for Research into the Computerization of Expressive Signs). See http://centridiricerca.unicatt.it/circse-centro-interdisciplinare-di-ricerche-per-la-computerizzazione-dei-segni-dell-il-centro-di-ricerca.

[5] In addition, Busa's family friend, student, and colleague, Dr. Marco Passarotti is at the Catholic University and he has been of great help promoting the value of the Archive and interpreting the gems found. One of the authors (Rockwell) heard Passarotti talk at a conference in

Darmstat and as a result went to visit. Once at the archive Rockwell met the archivist Paolo Senna, who is working hard to preserve and digitize the extraordinary collection. He kindly scanned and allowed us to show the “Flow Chart”.

[6] Rockwell posted a long blog essay on “The Index Thomisticus as Project” that describes many of the documents we found useful in the Busa Archive. See http://theoreti.ca/?p=6096. This also discusses the need for a historiography. The project proposal dates from 1962.

[7] Steven Jones is now developing a web project “Reconstructing the First Humanities Computing Center”, with a contribution from us, that documents the Busa project site in Gallarate. The site includes a 3D reconstruction of the factory and a selection of images from the Busa Archive. See https://avc.web.usf.edu/images/RECAAL/.

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

[9] See the blog entry by Rockwell mentioned above at “The Index Thomisticus as Project,” http://theoreti.ca/?p=6096.

[10] 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”. (See the Wikipedia entry on Pericope, https://en.wikipedia.org/wiki/Pericope.) In Tasman (1957) the word used is “*phrases* (meaningful sub-grouping of words… .” (p. 253)

[11] It should be noted that there was also repetitive labor, often gendered, in traditional hand techniques of concording. What the technology changed was the division of labor or the arrangement of humans and technologies.

[12] The one technology that had been at least partially been standardized was the punched card as a way of entering and storing data. The punched card, despite later discussion about folding, spindling, or mutilating, was a robust way of entering information so that it could be processed manually and by computers. As a data technology it goes back to the Jacquard loom and is still used today, such as in some voting machines like those used in Florida in the 2000 presidential election between Bush and Gore.

[13] One can get some sense of the flow from the images and recreation of the project space in the Steve Jones’ “Reconstructing the First Humanities Computing Center” project. See http://www.recaal.org/.

[14] This comes from a web page on the IBM 026 Key punch which we think Busa used as he listed it in the 1962 project proposal “Per Completare Lo Index Thomisticus…”. See http://www.columbia.edu/cu/computinghistory/026.html.

[15] It isn’t clear if Busa used optical mark readers on the cards or if cards thus marked would be manually processed.

[16] See http://nbviewer.jupyter.org/github/sgsinclair/epistemologica/blob/master/PunchcardOperations.ipynb.

[17] In a letter dated March 16, 1957 Busa tells a Rev. William Le Saint S.J. who had enquired about creating an index of Tertullian that “If you intend only an index verborum, then work will be extremely easy and economical for then you could start by punching the word cards directly.” (p. 1 of 2)

[18] They have a forthcoming book with translated interviews and research on exactly what the processes were.

[19] See https://avc.web.usf.edu/images/RECAAL/. Accessed June 11, 2019.

Works Cited

Boyd and Crawford 2012 Boyd, D. and K. Crawford (2012). “Critical Questions for Big Data.” *Information, Communication & Society.* 15:5. 662-679.

Burdick et al 2012 Burdick, A., et al. (2012). *Digital_Humanities.* Cambridge, MA, MIT Press.

Busa 1951 Busa, R., S. J. (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.

Busa 1958 Busa, R. (1958). “The Use of Punched Cards in Linguistic Analysis.” *Punched Cards: Their Applications to Science and Industry.* Eds. R. S. Casey, J. W. Perry, M. M. Berry and A. Kent. New York, Reinhold Publishing: 357 - 373.

Busa 1980 Busa, R. (1980). “The Annals of Humanities Computing: The Index Thomisticus.” *Computers and the*

Humanities. 14:2: 83-90.

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

Ihde 1998 Ihde, D. (1998). *Expanding Hermeneutics: Visualism in Science.* Evanston, Illinois, Northwestern University Press.

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

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

Mackenzie 1980 Mackenzie, C. E. (1980). *Coded Character Sets, History and Development.* Reading, Massachusetts, Addison-Wesley.

McGloin 1945 McGloin, D. (1945). "Revitalizing Liberal Education." *Jesuit Educational Quarterly.* VII:3. 133 - 138.

Nyhan and Terras 2017 Nyhan, J. and Terras, M. (2017). "Uncovering 'hidden' contributions to the history of Digital Humanities: the Index Thomisticus' female keypunch operators." *Conference Proceedings of Digital Humanities 2017.* Montréal, Canada: 313-4.

Parikka 2012 Parikka, J. (2012). *What is Media Archaeology?* Cambridge, UK, Polity.

Rockwell and Passarotti 2019 Rockwell, G. and Passarotti, M. (2019) "The Index Thomisticus as a Big Data Project." *Umanistica Digitale.* Vol. 5. DOI: 10.6092/issn.2532-8816/8575.

Stanford 1951 Stanford, S. C. and C. D. Gull (1951). "Transcription Problems in Preparing and Using Punched-Card Files". *Punched Cards: Their Application to Science and Industry.* Eds. R. S. Casey and J. W. Perry. New York, Rheinhold Publishing: 395-404.

Tasman 1957 Tasman, P. (1957). "Literary Data Processing." *IBM Journal of Research and Development.* 1:3. 249-256.

Terras 2013 Terras, M. (2013). "For Ada Lovelace Day — Father Busa's Female Punch Card Operatives" *Melissa Terras' Blog.* http://melissaterras.blogspot.com/2013/10/for-ada-lovelace-day-father-busas.html.

Turner 2014 Turner, J. (2014). *Philology: The Forgotten Origins of the Modern Humanities.* Princeton, New Jersey, Princeton University Press.

Winer 1980 Winner, L. (1980). "Do Artifacts Have Politics?" *Daedalus.* 109:1. 121-136.

Winter 1999 Winter, T. N. (1999). "Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance." *The Classical Bulletin.* 75:1. 3-20.

Zielinski 2008 Zielinski, S. (2008). *Deep Time of the Media: Toward an Archaeology of Hearing and Seeing by Technical Means.* Cambridge, Massachusetts, MIT Press.