SERG – Grupo de Pesquisa em Engenharia Semiótica do Departamento de Informática da PUC-Rio

Série de Notas do EMAPS
Ética e Mediação Algorítmica de Processos Sociais

The Invisible Discipline in Academic Research

EMAPS Notas #04

Clarisse Sieckenius de Souza
www.inf.puc-rio.br/~clarisse

Fevereiro de 2024

Como citar este documento: de Souza, C. S. (2024) **The Invisible Discipline in Academic Research**. *EMAPS-Notas #04*. Rio de Janeiro, RJ - Brasil: SERG, Departamento de Informática, PUC-Rio, 2024. 18 p. URL: www.hcc.inf.puc-rio.br/EMAPS//userfiles/downloads/Notas-InvisibleDisciplineInAcademicResearch2024.pdf

The Invisible Discipline in Academic Research

Clarisse Sieckenius de Souza

Janeiro de 2024

Abstract:
This article presents the ideas that Clarisse Sieckenius de Souza proposed to discuss in the *Digital Technologies and the Knowledge Economy* panel of the II PUC-Rio Colloquium on the Philosophy of Technology. By exploring Erin Glass's concept of *the invisible "technological" discipline*, which has been affecting higher education teaching, the author proposes that this discipline is also affecting research activities in contemporary universities, with the same potentially disturbing consequences as in educational contexts. The article builds its argument around Franco Moretti's notion of "distant reading," helping the user navigate reading and interpretation activities with different kinds and levels of digital technology mediation. Throughout the argument, the reader will find links to watch and experiment *distant reading* with the use of Voyant-Tools, a popular computer-assisted text analysis environment, freely accessible on the Web.

Keywords:
Computer-Assisted Academic Research, Distant Reading Practices in Research, Methodological and Epistemological Challenges for *e*-Science

1 Introduction

This *EMAPS Note* presents my thoughts in preparation to participating in the *Digital Technologies and the Knowledge Economy* panel of the II PUC-Rio Colloquium on the Philosophy of Technology. I am a linguist by training, but my academic career evolved entirely in the area of Computer Science, where I moved from natural language processing (a branch or artificial intelligence) to computer semiotics, human-computer interaction, end-user programming, human-centered software development, and then back to artificial intelligence, now with an emphasis on the philosophy of technology and digital rhetoric. In January 2020, I became *Professor Emerita* of Informatics at PUC-Rio, and have since been part of an interdisciplinary group interested in *Ethics and Algorithmic Mediation of Social Processes* (whose acronym in Portuguese is **EMAPS**).

My life-long interest and concern as an academic has gravitated toward *technology as an artificial language*. In this language, humans who can speak it perform powerful speech acts, which affect their interlocutors (*i.e.*, the "users" of the technology they design and develop), the interlocutors of their users, and potentially more indirect interlocutors along this computer-mediated social communication diffusion process. This is the ethical dimension of semiotic engineering, a semiotic theory of human-computer interaction (de Souza, 2005; de Souza and Leitão, 2009) and, later, a support theory for human-centered software engineering (de Souza et al., 2016), which we have been developing at **SERG** for nearly 30 years, by now. The main idea in the theory is that human-computer interaction (HCI) is a specific form computer-mediated social interaction, where any given system's designers and developers communicate with this system's users through interface protocols that work as their *proxies*, that is, the interface plays the designers' and developers' part (or *speak* for them) in all supported conversations (technically called *interactions* in HCI). The system executes internal programs that contain all, and only, the rules that determine the proxy's capacity to mediate communication between the human parties involved in it. These programs can be interpreted as implementations of human communication models, through verbal and nonverbal signs. One of the hard challenges for HCI design is that computer programs are *automata*, everything they do is strictly determined by rules governing their response in anticipated situations that may or may not occur. This does not happen to (and is therefore not expected from) humans, whose social behavior is hardly ever the same, changing and adapting constantly to even minuscule differences in familiar situations. The fact that humans' computer proxies are governed by specified rules, including when they can learn from interactions, means that their behavior is, at least in theory and unlike ours, predictable (and hence replicable), over time and space. (de Souza, 2017)

"*Knowledge economy*" denotes many different things, most of which I cannot competently discuss. I will, therefore, limit my contribution to some critical considerations based on my own academic studies and practical experience. Thus, I propose to talk about *knowledge production* – more specifically, *academic knowledge production*.

I belong to a generation that *made the transition* from computer-independent research to computer-dependent research. This happened about half way through my career, which means that I have equal experience with *manual* and *digital* research, so to speak. Moreover, all of my research has been directly devoted to enhance the design and development of useful and usable technologies. From this standpoint, as a researcher and savvy user of various kinds of technologies, I will argue that we urgently need to confront the *invisible discipline*, a concept that I have learned from Glass (2018, 2021), and which she defines as follows, with my added emphasis:

> [D]igital technologies are often adopted and evaluated by academics and institutions according to their practical value and professional or community norms. What I want to offer here is an analysis of digital technologies that instead focuses on the broader social and political realities that they reinforce, support, or create. This type of analysis may seem foreign to many academics because of the way higher education has long encouraged (or even taught, if you will) its members to passively accept digital technology within research and learning environments as predominantly natural, neutral, and inevitable. Although the university's tendency to reinforce technological complacency may be accidental, its effects have been politically disastrous [...]. Elsewhere, **I have called this unfortunate tendency in higher education the university's invisible discipline** and have linked it to the mass helplessness we see in response to widely reported ethical infringements carried out by large digital technology companies. (Glass, 2021, p. 24)

Erin Glass, who is a keynote speaker in this colloquium, anchored her analysis in observations made in teaching and education environments. Unlike hers, the arguments I present are not the result of a PhD thesis (*i.e.*, in comparison, they are my *opinion*, not my *claim*), and I am focused specifically on academic research. I suppose that the fact that Glass presented her ideas at the end of her PhD program and that I am inspired by her in presenting mine, at the beginning of my retirement years, within a few years from one another, are telltale. But I will leave it to my readers and the participants of our colloquium to conclude.

2 Knowledge Machines

By the second decade of the 21st century, *knowledge machines* were already extensively used in research. Meyer and Schroeder (2015) characterize such machines as a variety of digital instruments to manipulate a vast range of resources, which researchers and members of the public can access via the Internet to advance, or make use of, research results.

Meyer and Schroeder studied how technology was used in *e*-research, that is, research using knowledge machines at some point in the process of knowledge production. In their book, they present the details of six widely different *e*-research projects. The role of technology is different in every case, but some common aspects prevail. In the next two subsections I will touch on the ones I selected for this panel.

2.1 Styles of Scientific, Technological and Scholarly Research

Because knowledge machines are now used in virtually all areas of scientific, technological and scholarly research, it is important to understand how the *styles* of doing academic work varies from one to the other. Probably, the clearest contrast is between how natural sciences advance, compared to the humanities. While progress in the natural sciences is heavily *cumulative*, in the sense that new knowledge typically *adds* something to previous knowledge, progress in the humanities is *non cumulative*, in the sense that new knowledge is often a *problematization* of previous knowledge. "[S]cience has mechanisms to reach closure (at least temporarily) on major questions, whereas the social sciences often do not, and the humanities largely do not." (Meyer and Schroeder, 2015, p. 203) The idea of *competing interpretations* of phenomena and objects of investigation may well be unknown, sound irrelevant or even absurd, to many natural scientists, although every undergraduate in philosophy, literature, social communication, or anthropology, for example,

is educated to detect and, indeed, encouraged to explore and produce *competing interpretations* of their object of study.

Technology is a *hybrid* territory, though. Considering that *design* is an important part of everything we do in technology, it is clear that competing interpretations of technological solutions and their impact reveal the *non cumulative* aspect of academic studies in this area. However, there is also a large portion *cumulative* knowledge, too, since the advancement of digital technology requires cutting-edge knowledge in physics (for the hardware industry), as much as extended logic types and mathematical models (for the development of new algorithms). This dual nature of technology is what makes it non neutral and prone to permanent ethical, theoretical, epistemological, and methodological questioning, in spite of how many researchers in the *tech* areas carry their projects *obliviously*, without consideration for the consequences and implications of their work *in the wild outdoors* of human psychosocial experience.

2.2 Some Implications of Knowledge Machines in Research

Meyer and Schroeder (2015) offered five challenges at the end of their book. At the time of writing, they believed that each one was a critical issue to be thought through and discussed in the next few years (*i.e.*, by now). First, there are important implications of research knowledge being accessible on the Internet. One of them is that *anyone* can access and use large volumes of research data and reports, regardless of how prepared this person is to understand such findings, let alone discuss them, or use them. Another implication is that students, as well as researchers and scholars, can search and find knowledge produced outside the context of their own practice, their own labs, their own institutions, and so on. If, on the one hand, this has been promoting *interdisciplinarity* at much higher speed than ever before, on the other, it is not clear that the quality and compatibility of interdisciplinary sources for any given research project are always warranted. So, the challenging factor is how much the expert or non-expert users of knowledge machines know about research and researchers, in order to make good judgments about the knowledge sources they access and use.

Second, the pressure for doing *e*-research affects research funding, and *collaboratories* tend to gather research communities with different access to funding and other resources. The challenge is a political and ethical one: what are the standards of *fairness* when different research groups collaborate? Third, some researchers have better conditions (including computer infrastructure) to publicize their work in the form of digital documents, blogs, wikis, videos, free online software tools, than others. The very *product* of their work may be more or less amenable to *e*-publication in different kinds of media. Moreover, whereas some researchers may *test and publish fast*, others – because of their very object of interest and domain – may take long to achieve worthy results. Therefore, given how ranking algorithms of widely-used search engines work, good research in some areas may be much harder to find than bad or biased research. And, since this affects some of the glorified *research productivity indices* automatically calculated for every researcher known in the Internet, the challenge is to make sure that technology is not making it harder than easier to find good researchers in certain areas.

Fourth, the various digital representations of *e*-research process and products are not one and the same in all cases. Technological mediation has a vast range of possibilities, and often (if ever) it is not possible to *desintermediate* (Meyer and Schroeder's word) what we find, in order to have a fair interpretation of what our finding really *means*. The challenge is the obvious threat

to the quality of knowledge being produced, along with multiple opportunities for unnoticed misinterpretations of knowledge across disciplines in interdisciplinary projects. Finally, the fifth challenge is worth quoting in the authors' own words:

> These changes constitute a *scientization* or technological transformation of knowledge [...]. The wider (nonresearch) implications of these changes are that these knowledge machines occupy an ever more central place in society, with the added consequence that researchers are made more aware of how their research is perceived by other researchers (outside of their domain) and by a wider public. (Meyer and Schroeder, 2015, p. 220)

While previous challenges pointed at the absolute necessity to develop philosophical and political awareness to navigate competently in the new territory of *e*-science and *e*-scholarship, the fifth one points at the social responsibility of researchers. The physical, social, and psychological well-being of all of us on planet Earth has been threatened in ways that urgently call for concerted action. This involves all sorts of actors, not the least among them, researchers and scholars with a keen eye for ethics, methodologies, and justifications that are used to promote *valid knowledge*, and discard *false knowledge*, when searching for much-needed solutions to global problems.

3 Close and Distant Reading

Some years ago, Moretti (2007, 2013) sparked an impactful debate in the humanities, when he proposed a *new condition of knowledge*, called "distant reading." He tells us that he had this insight when preparing an essay about the history of European literature. In his own words, [e]volution, geography, and formalism, the three approaches that would define my work for over a decade, first came into systematic contact while writing these pages. (Moretti, 2013) Evolution, from Charles Darwin's *Origins of Species* and Ernst Mayr's *Systematics and the Origin of Species*, called Moretti's attention to the conventional representation of transitions from unity to diversity (*i.e.*, *trees*), and that of the distribution of quantity over some continuous dimension (*i.e.*, *a line*). Geography contributed with the conventional representation of how massive concrete objects, which our natural senses can only (very) partially grasp, are distributed over cognitively meaningful space (*i.e.*, *maps*). Formalism, in literary analysis, aimed at the scientific study of literature, centered on forms, their structure and functions. The influence of this approach on Moretti's work was that he tried to *read* literature by isolating the more traditional socio-cultural and psychological dimensions from the formal properties of text. The interesting question that can begin to be answered with such reading is: What do we learn about text when we look strictly at its *formal properties*? Using *trees*, *line graphs* and *maps*, he could represent what he could see, which became in turn a kind of text representation that deserved its own reading.

Moretti was paving the way for data-driven literary studies, where mostly quantitative (but also, although less frequently, qualitative) *metadata* about written texts are used to represent *patterns* in typically big samples of textual objects. His work, and that of his followers, set out to demonstrate what can be known with the *distant reading* of texts, and what this knowledge is worth. Close and distant readings, he insists, are *different conditions of knowing* textual objects. Their role, we conclude from several of his examples, is to pose a permanent intellectual challenge for each other, because it is obvious that some of the knowledge in literature can only be the result of *close reading*, while some other knowledge in the field can only be captured with *distant*

reading. These ideas stirred the digital humanities (DH) community; the concepts expanded the *vocabulary* of researchers in the area, inasmuch as *close reading* is a COMPUTER-INDEPENDENT activity, whereas *distant reading* is a COMPUTER-DEPENDENT one. Yet, computers can be used to enrich and support *close reading* in various ways. One example is the *social experience* of sharing highlighted passages from digital books; another example is the *embedding* of copied passages in new contexts of text production. Other examples come from research *close reading*, where qualitative analysis software like Atlas.ti[1] and MAXQDA[2], allow text analysts to *code* and *categorize* topics and themes, as well as to add indefinitely many and varied annotations (including content from other media), organize annotations, link text and annotations to web content, share this material, and so on.

With respect to *distant reading*, in addition to providing networked infrastructure, computers provide information search, retrieval and management, in addition to all kinds of numerical and symbolic calculations, pattern discovery and visualization, and – since OpenAI inaugurated large-scale use of AI in all sorts of computing – machine learning. The range of *distant reading* experience with texts can start with the extremely useful but simple, in comparison, citation management systems, like Zotero[3], Mendeley[4], EndNote[5], and JabRef,[6] all of which can search and retrieve text and document metadata, generate bibliography reports, support recursive text classifications, annotations, filtering, and more. They are essential tools for research at all levels and in all areas of knowledge. The *distant reading*, in this case, is what we do with bibliography we collect in this way. We do not read all the papers, and books, and technical reports we retrieve. We browse or *mine* the abstracts for relevant content, search for common references, create clusters, create links between items or clusters, and *read* these structures to decide which publications we are going to examine with a *close reading*.

At the time of my writing this article, Texas A&M University Libraries Research Guides on the Web provided a selection of AI-based literature review tools. Semantic Scholar[7] is probably the most popular one listed on their page. Likewise, MAXQDA and Atlas.ti now incorporate AI. A relevant point for this panel's discussion is expressed in the advertising paragraph on Atlas.ti homepage. This is what it says (https://atlasti.com/, visited on February, 2024):

> ATLAS.ti bridges human expertise with AI efficiency to provide fast and accurate insights. Communicate directly with your documents and have them automatically coded based on your intent for customized results. Leverage the most advanced AI tools that make suggestions while you have the final say.

Automatic coding and *having the final say* seem to be at odds with each other, at least in some intuitive interpretations of the paragraph above. If AI is going to do the coding automatically, I presume that the analyst has *asked AI to do the reading*, from which coding emerges. I am therefore confused about what kind of *final saying* this technology leave to qualitative researchers who adopt this strategy.

In the next section, I will briefly present a small study with comparative *close and distant*

[1] https://atlasti.com/
[2] https://www.maxqda.com
[3] https://www.zotero.org/
[4] https://www.mendeley.com/
[5] https://endnote.com/
[6] https://www.jabref.org/
[7] https://www.semanticscholar.org/

readings of a short text. It is deliberately a *pre-AI* scenario, where I wanted to explore the contributions of natural language processing to the formal reading of texts, as is done in research using automated literature search and retrieval in large bibliographic databases, without having a chance to *delegate* the reading to an artificial intelligence. I was constantly the 1st-person reader of the text, looking at it closely, or at a distance. This was my strategy for keeping the mutual *intellectual challenge* between the two conditions of (my own) knowing (Moretti, 2013) healthily alive.

4 A Small Exercise Comparing Close and Distant Reading

The short-term goal of the simple exercise described in this section is to have a first-hand perception of *what* I learn, and *how* I learn it, when reading text with and without computer text analytics. The long-term goal is to take the first step in search of understanding the epistemic revolution that technologies being used for distant and "delegated" reading (like ChatGPT) have set in motion. At this point, this is a personal learning effort, not an academic research project. I chose to use **Voyant Tools**[8] for *distant reading* for important reasons. *Voyant* is freely

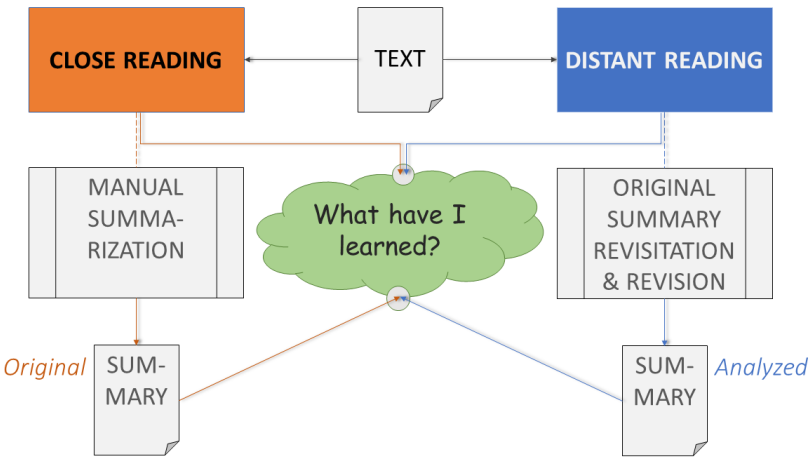


Figure 1: Main steps in the exercise comparing *close* and *distant reading*

accessible on the Web. It does not require a login or client software installation, and there is no paywall. *Voyant* offers very powerful *hermeneutic* tools, not only in terms of analytic processing, but also in terms of information visualization, publicly accessible communication of results, and end-user customizations through programming (*on wheels*, with Spyral Notebooks pre-programmed resources, and *the brave way*, with full programming in Javascript).[9] Moreover, *Voyant* is the result of combined theoretical and practical work in digital hermeneutics. Its creators, Geoffrey Rockwell and Stéfan Sinclair, are both experienced researchers in the field of computer-assisted interpretation in the humanities, and competent technology design leaders. In their book called *Hermeneutica* (Rockwell and Sinclair, 2016), they provide the theoretical

[8] https://voyant-tools.org/
[9] See *Voyant* online documentation.

foundations behind *Voyant*, as well as the *rationale* for many of their design decisions. This contributes to especially well crafted technological transparency, not commonly seen among similar tools.

The exercise consisted of three major steps (see **Figure 1**), after the selection of a short text freely available on the Internet. The first step was reading it *closely*, in the traditional way, and registering my understanding of the text in the form of a manuscripted summary. The next stage was to submit the selected text to *Voyant*, freely explore and use the available computer-assisted interpretation tools to analyze it, and then revisit the previously produced summary, making annotations and revisions when appropriate. In the final step, I asked what I learned with each kind of reading, and – most importantly – with the entire process of reading a text both ways.

The text I chose was the **Message of Pope Francis for the 2024 World Day of Peace**. The message is short and has been officially translated into several languages. I selected the **Portuguese** version because I am a native speaker and did not want second-language deficiencies to interfere in my reading. Moreover, the message is about *Artificial Intelligence and Peace*, a theme not too distant from what we are discussing in PUC-Rio's colloquium this year. This was an additional reason for working with this particular text. For those readers who understand Portuguese, the summary I produced after close reading and interpretation is available **here**.[10]

The details of my distant reading are described, in Portuguese, elsewhere.[11] In the next few paragraphs I provide the highlights of process and findings, with direct links to *Voyant*'s website, so that interested readers can visit it and *play* with the material I have produced. The uploaded corpus and the *default* dashboard of the analysis can be accessed at: voyant-tools.orgcorpusc̄26037787d03fae83fffc44803844456&viewc̄orpusset.

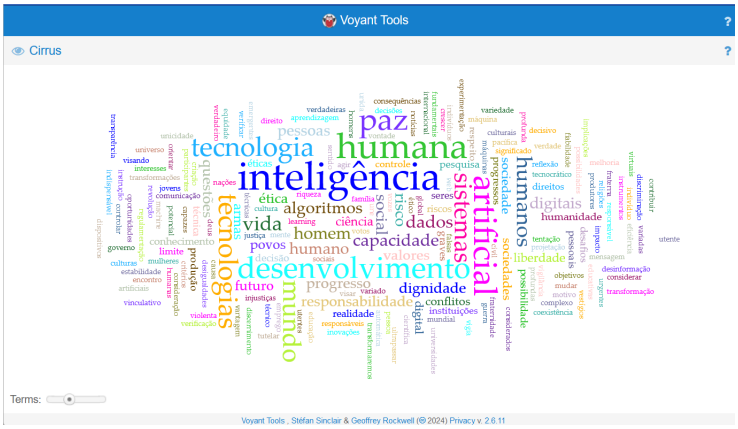


Figure 2: A word cloud for Pope Francis's message, showing 195 terms

With quantitative information provided by the **Summary Tool**, especially the list of "Most frequent words in the corpus,"[12] along with the **Terms Tool** and the **Phrases Tool**, I was able

[10] English speakers who wish to have a flavor of this summary can submit the text to their favorite machine translation system. I haven't done the translation, myself, because I wouldn't able to escape the temptation of writing a new version of the summary in English.

[11] de Souza, EMAPS-Note #05, to appear.

[12] In this online Summary, the parameter for how many words are shown as the "most frequent" is set to 600. This

to build a list of terms to be excluded (**here**) and two other lists with significant terms to be included (a **long** and a **short** list). With these resources in hand, I was able to use the **Cirrus Tool** and produce a popular visualization of document content called *a word cloud* (**Figure 2**). If you decide to go online and *play* with my configuration for Cirrus, try using the "Terms" slider to change the number of words in the cloud. You can also create your own inclusion and exclusion lists to see how this kind of representation of the Pope's message content can be improved.

In a more qualitative kind of exploration, I used the concordance list (**Contexts Tool**) to examine the contexts where words from my inclusion lists (long and short) occurred. I also used the **Collocates Tool** to find out which words appeared next to each other, and how frequently. The result was a revised and condensed version of my inclusion lists (**here**), which allowed me to move on to an important step in the analysis, the categorization of key content-related terms. A nice side-effect of revising the inclusion lists was to improve the *word cloud* representing the message content (see the new Cirrus **here**).

The categorization of included key terms (*i.e.*, words and expressions) involved iterated readings of the message text using *Voyant*'s **Reader Tool**, as well as the *consultation reading* (Rockwell and Sinclair, 2016, p. 87) with the **Contexts Tool**. The resulting list of categories included the following: THREATS AND RISKS; PROGRESS; TECHNOLOGY; HUMANITY; VALUES, ETHICS, MORAL; RELIGION; and REGULATION. [a]

The best visualization for the results is a line graph, produced with the **Trends Tool**. Here, some technical details are worth mentioning. The **Trend Tool** can be customized using several parameters. **Figure 3** shows how the line graph shown **here** has been configured. If you place the pointer on the "Segments" slider control (or examine the URL carefully), you will see that its value is set to **8** (`bins=8`). Segments are *slices* of texts. Therefore, using 8 segments means that the text has been *sliced* in 8 pieces. I suggest that, if you decide to go online and interact with this tool, you change the number of segments and watch what happens on the line graph.

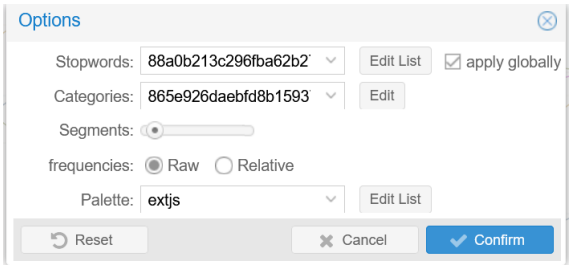
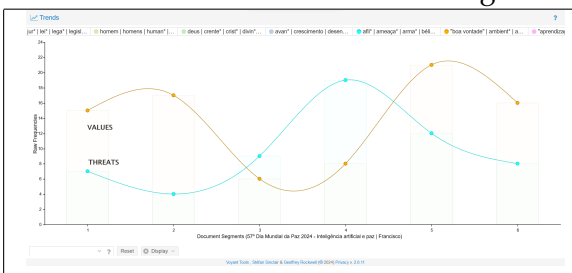


Figure 3: Parameters used with the **Trend Tool**

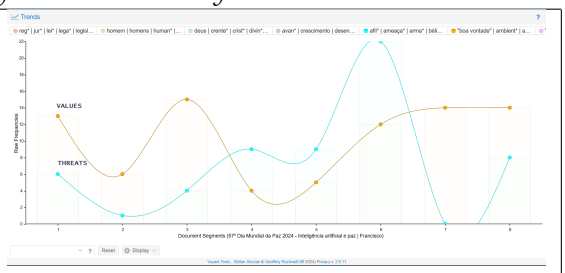
Table 1 shows that a small change in the number of "Segments" reveals the *rhythm and harmony* of the Pope's message, to use a musical analogy. On the left-hand column, the number of segments is set to **6**; on the right-hand column, to **8**. Although some rythms and harmonies

is a high limit parameter value, more than 10 times the default one, 5. It requires considerably more computation for conclusion, which is why the page takes longer to load. Only with hundreds of words has this tool been useful to me. The last word in the 600-word list is `obsessão (1)`. If you don't see it at the end of the list, the page has not been completely loaded.

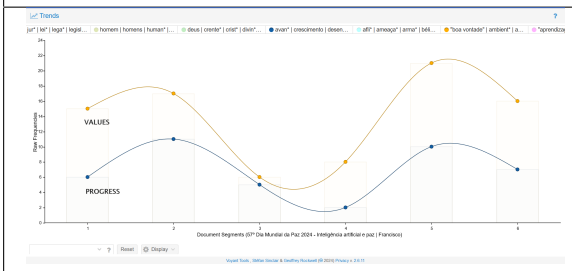
Table 1: Visualizing the *Rhythm and Harmony* of Text



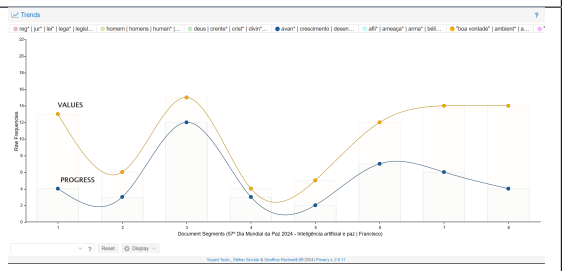
Threats and Values 6 segments



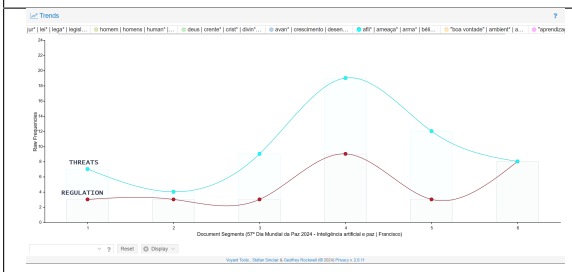
Threats and Values 8 segments



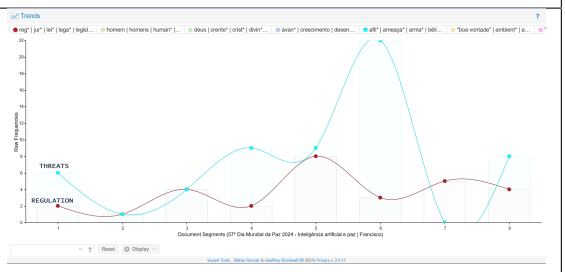
Values and Progress 6 segments



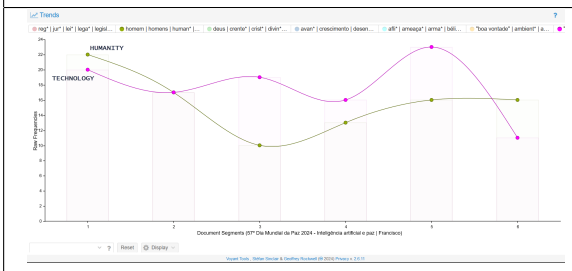
Values and Progress 8 segments



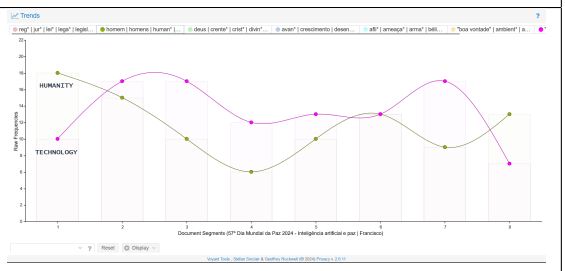
Threats and Regulation 6 segments



Threats and Regulation 8 segments



Technology and Humanity 6 segments



Technology and Humanity 8 segments

look the same on both sides, others are only visible with 6 segments. This means that the *number 6* semiotically *indicates* a relevant pattern of text organization. Unfortunately, it is not clear – to a novice Voyant user like me – where segment boudaries are placed in the text. This information would be fantastic for a *formal rhetorical analysis* of the message. In my small study, I haven't followed this thread any further, but any interested reader can do it online, from where I stopped.

The concluding steps in the exercise consisted of a *re-reading and revision* of the summary manually produced, in view of my learning from *distant reading*. Some observations are worth sharing in this article.

1. I was surprised to see how the first summary produced after a *close reading* clearly reflects the *pathos* of the Pope's message, its rhetorical effect on me. The vocabulary and phrases in my text mirror the Pope's eloquent choices. In comparison, my *distant reading* was thoroughly semantic, centered on the message's *informational content*, rather than rhetoric.

2. Possibly as a consequence of the above, REFLECTION and EXHORTATION are two themes that completely escaped my analysis with Voyant, although it obviously stands out as the key point of the message in the manuscripted summary.

3. The reexamination of the original summary also showed that I had picked up some of my favorite cherries in the Pope's message, like the fact that technology is not neutral, that human intelligence is not fragmented, that we may not be able to tell truth from falsehood, and so on. Although I don't believe that this is *a problem* – quite contrarily, I see it as my individual contribution to the vast universe of possible readings that all texts have to offer – it can be misleading in summaries. For example, my favorite themes haven't all been developed to the same extent by Pope Francis. This was clear at *distance reading*, which helped me revise certain passages (see the revised Portuguese version **here**).

4. The *distant reading* also suggests that this message was not about religion. Even if it was written by the supreme authority of the Catholic Church and references to religious documents are abundant, the perspective of this document is *ethical* rather than *religious*. This is visible with the **Trends Tools**. The *ethos* of the text, the rhetorical effect of it having been written by *the Pope*, has influenced my *close* and *distant* readings alike. The original summary, in particular, is full of references to the Pope, as the author of the message. The text itself, of course, only mentions it in its heading and the signature at the end. It is written in *first person*.

5. The third rhetorical aspect of the message – which may well has affected me, but I did not notice it – was its underlying rules and structure (the *logos*), nicely depicted in the *waves* shown in **Table 1**. Here, the value of a formal *distant* approach was evident. The way how the themes were distributed and combined throughout the text plays an important part in making messages clear and convincing. The **Trends Tool** is, thus, a window open to the *logos* of texts.

My findings confirm Moretti's point that *close and distant readings* are **two ways of knowing** that intellectually (and very positively) challenge each other. Whichever one is used in isolation means that the wealth of knowledge to be gained with the other is lost. My study had the advantages and disadvantages of using a tiny volume of data, which is not what *distant reading*

is (or, according to some, should be) used for. *Distant reading* is good for *Big Data*. Yet, the exercise described in this section shows that it can be good for *Small Data* as well. In the context of current data-driven practices in scientific research, I suppose that we can correctly assume that **both kinds of readings** make equally legitimate contributions to knowledge production and are equally necessary. The problem is, as will be discussed in the next section, how to convince the academic community that such is the case, not only because of old "science wars" fought in new guises, with technology being used to claim and gain political power, but also because of the engrossment we can easily feel in computer-assisted research steps. There is even the risk of mistaking assistance for delegation, as has been lately seen in cases of ChatGPT misuse (Sallam, 2023). At this point, we meet the *invisible discipline*, which I will discuss in the next and final section of this article.

5 Reasons for "Reprogramming the Invisible Discipline" in Research

In the introductory section, I quoted a passage from Glass (2021), where she concisely expressed her concern with *the invisible discipline* in higher education. She defines it as the a kind of silent ongoing operation whose effect is that university faculty, students, researchers and administration are *"encouraged (if not taught)"* to accept digital technologies passively, as if such technologies were *"natural, neutral, and inevitable."* Decisions about which types of technologies are used for teaching, learning, and administration tend to be selected for technical, financial, and productivity reasons, typically without discussing their potential for alienating previous work practices and people, whose contributions *don't quite fit* in the new technological environment. Non-fitting contributions are frequently lost and the people who used to make them must accept to be re-educated.

The first question I should ask in my concluding remarks is: *Is there an invisible discipline in academic research?* Assuming that the answer is "yes," the second question is: *Are there reasons to reprogram it?* In this section, I will *begin* to answer these questions and add some thoughts about what we can do next.

5.1 Is there an invisible discipline in academic research?

I definitely believe that *there is* an invisible discipline in academic research. I will mention just three indications that this is the case, coming from very different directions. The first indication is how *systematic literature review* (SLR) has been swiftly and increasingly adopted as *the standard* method to survey the "the state of the art" in virtually all areas of science, technology, and scholarship. According to (Shaffril et al., 2020, p. 1320):

> SLR has several advantages compared to traditional review such as its numerous unique procedures. SLR encourages researchers to look for studies outside their own subject areas and networks through the introduction of extensive searching methods, predefined search strings, and standard inclusion and exclusion criteria (Robinson and Lowe 2015[13]). This kind of review stress on transparency, all terms in inclusion criteria for example, must be defined and justified while

[13] Robinson, P., Lowe, J.: *Literature reviews vs systematic reviews*. Aust. N. Z. J. Public Health 39(2), 103 (2015)

> exclusion of articles must be reasoned (Greyson et al. 2019[14]). Furthermore, SLR heavily focuses on evidence, impact, validity and causality, it urges researchers to examine information on research design, analytical methods and causal chains, and by practising this, SLR is controlling the quality of review by ensuring the robustness of evidence (Lockwood et al., 2015;[15] Mallet et al. 2012[16]).

Since the Internet provides a vast collection of academic publications and citations that we, humans, cannot even begin to search and browse without the use of digital technologies, the SLR method requires the use of search engines and text mining technologies. One of the symptoms that an *invisible discipline* is driving the movement is the ill-informed ways in which a disturbing portion of researchers are using it. Authors like Puljak and Lund (2023), Delaney and Tamás (2017), Okoli (2015), Boell and Cecez-Kecmanovic (2015) and Boell and Cecez-Kecmanovic (2014), for example, have expressed their concerns about such widespread misunderstanding of the method and waste of its merits. All of them mention technology in their analyses, although it is my personal hypothesis that technology may be playing an important role in the drama. The *engrossment* with digitality or the obstacles that typically come along with the use of powerful computer tools by more or less experienced researchers is something I have very frequently observed. We shouldn't underestimate the effort: a smooth, productive use of research technologies is *not easy*.

The second indication is is that, because all researchers must use technology to keep up-to-date in their domain of interest, the visibility, or *findability*, as it has been coined, of online documents (including multimedia items, such as academic lecture videos, software demonstrations, etc.) is a critical factor for researchers and publishers. Researchers depend on the *findability of their work* to achieve their mission and to climb the highly competitive steps of academic careers. Publishers, in turn, depend on the *findability* of their products because this is their business. The mutual interests in optimizing the *format* of academic publications to improve their visibility is strongly tied to the mechanics of existing search engines and text-mining systems (Marks and Le, 2017; Schilhan et al., 2021). The Association for Computing Machinery (ACM), a leading publisher in computer science, for example, has turned to full-fledged *publication workflows* as substitutes for the traditional *submission templates* (Forget everything you know about "templates."[17]), a clear indication of how technology has turned research publication into a bureaucratic, machine-regulated, process.

The third indication of the *invisible discipline* is sharply depicted in a famous article written by Anderson (2008), then editor-in-chief of *Wired* magazine. The last 3-sentence paragraph of the article is: "There is no reason to cling to our old ways. It's time to ask: *What can science learn from Google?*" I could rest my case here, but some additional comments are worthwhile. The article is short and non-academic, but written by someone who has been trained in the sciences (physics, quantum mechanics) and has worked as a science journalist for the prestigious *Nature* and *Science* magazines. Anderson strongly argues that the *petabytes era* inaugurated by Google

[14] Greyson, D., Rafferty, E., Slater, L., MacDonald, N., Bettinger, J.A., Dubé, È., MacDonald, S.E.: *Systematic review searches must be systematic, comprehensive, and transparent: a critique of Perman et al.* BMC Public Health 19(1), 1–6 (2019).

[15] Lockwood, C., Munn, Z., Porritt, K.: *Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation*. Int. J. Evid. Based Healthc. 13(3), 179–187 (2015).

[16] Mallet, R., Hagen-Zanker, J., Slater, R., Duvendack, M.: *The benefits and challenges of using systematic reviews in international development research*. J. Dev. Eff. 4, 445–455 (2012)

[17] https://www.acm.org/articles/pubs-newsletter/2019/blue-diamond-mar-2019#3

puts an end to an *old problem* for scientists, namely, that *correlation is not causation*. In Anderson's view, we no longer have to use theoretical models to validate the meaningfulness of data, hence the provoking title of his essay, *"The End of Theory: The Data Deluge Makes the Scientific Method Obsolete"*:[18]

> "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

Indeed, the "*let the data speak*" mantra is catchy and has been used and misused by many, who mistake the representational essence of *big data* for the naturalist conception of *observed data*. If interpretation was a topic for fierce dispute in pre-petabyte philosophy of science, today it is indisputable that *the data* stored in petabyte databases is a humongous set of **binary representations of** whatever they are claimed to represent by those who created the representation. So, of course there is interpretation; there must be. The dispute is over in data science. The fact that we delegate interpretation to algorithms does not make it vanish from the picture. And yet, this is seldom discussed in data science publications, which typically herald the superior validity of their findings. We already see some discussion about how *Big Data* seems to be reviving old positivism in a machine-justified version of an old vision of science that we thought was already behind us (see Fuchs (2017), Jones (2018) and Skees (2020)).

5.2 Why should we reprogram the invisible discipline?

I think that the short answer for why we should reprogram the invisible discipline is intellectual honesty. But this answer is clearly insufficient because its metonymical rhetoric obscures the extent and depth of causes and consequences. Intellectual honesty *depends* on education that is founded on firm ethical values and is competently guided by duly qualified professionals. Among other factors, intellectual honesty must be able to detect and avoid (internal and external) political manipulations of research, as well as threats to the quality of knowledge (with its own consequences for the living conditions of all the world population, for other forms of life on the planet, and beyond).

I would like to elaborate very briefly on two aspects of the infinite digression where a discussion of *intellectual honesty* might lead us. One is related to the often neglected need for teaching philosophy to young scientists and technologists. The truth is that many of us, experienced researchers and scholars without training in philosophy, eventually begin to ask philosophical questions about what we are doing (or should be doing). We realize the perils that we have or have not avoided in our long careers, for sheer lack of an adequate philosophical perspective on how to work with knowledge, how to relate to truth, and how to think critically about various forms of knowledge validation that we may have interpreted in a more instrumental way than we should. We wish we could have been better educated in the philosophy of science, technology, and ideas. But we haven't, although many have had successful research careers without any kind of *educational rewiring*. So, what evidence do they have that they would have been better scientists and scholars if they had been educated differently? They don't have any. They *just know it*. The paradox is that, because of currently dominant views on science, we may need *data* to prove it before the research community is convinced that the matter deserves attention.

[18] https://www.wired.com/2008/06/pb-theory/

The second aspect I would like to mention is *interdisciplinarity*. There has been a certain pressure for interdisciplinary work in the last two or three decades. The complexity and the magnitude of current societal and environmental challenges can only be solved by a collective effort of researchers coming from all academic domains. There are many ways of defining and practicing interdisciplinary. Although all of them are typically gratifying and illuminating from a personal development perspective, not all are equally productive from a knowledge perspective. For example, if the members of an interdisciplinary research team are using data-driven methodologies, say, but not all of them have the same understanding of the *implications* of such methodologies for the validity of those parts of the overall results that they are expected to contribute, the entire project may be compromised. The pressure for using computer technology means that we have to *operationalize* knowledge, that is, describe explicitly *what* constitutes our objects of interest in some given domain and *translate* this description into a computationally operable representation. As Bonino and Tripodi (2021) remark, verifiability criteria in computer-assisted research requires that we accept to decouple our *natural* (or intuitive) conception of objects of interest from their *operationalized* version. These two forms are not the same (just think of how many exceptions even the simplest attempt to describe, formally, a natural entity like "fish" can generate). So, there is always some loss-and-gain evaluation to be made when operationalizing a concept. A formal, operationalized description may surprise us because it includes some intuitively unexpected entities (*e.g.*, whales are mammals), and frustrate us with the exclusion of others (*e.g.*, whales do not belong to the same species as we intuitively think of as *fish*). Problems of operationalization in interdisciplinary groups require particularly competent epistemological examination, which I believe can only emerge in groups that are not only *inter**disciplinary***, but also *inter**epistemic***, by which I mean that they have a solid practical and theoretical experience with different modes of knowing, as well as sharp sensitivity to epistemological and methodological shifts.

5.3 Where could we start to reprogram the invisible discipline?

There must be dozens, if not hundreds, of starting points to address the problem we have discussed in this article. Moreover, since we are immersed in the problem, it is not too clear which ones are likely to lead us to dead ends and vicious circles. So, what follows is *a wishful guess*.

I believe we can start by *promoting education in the philosophy of science and technology* for all the community involved in research: faculty, graduate students, academic lab researchers, industrial researchers, R&D administrators inside and outside the university, and so on. To promote education does not necessarily mean to *teach graduate courses*, but to provide rich learning environments in the form of workshops, permanent discussion forums, and cross-disciplinary epistemologically-sensitive group projects, among other possibilities. It is clearly a challenge for those who will design and execute such educational activities, but to dodge the challenge can lead us into the *disastrous consequences* that Glass (2018, 2021) warns us to escape.

In a thought-provoking article about "the big challenges of big data biology" (BDB), Callebaut (2012) addressed the fact that BDB practitioners might be acting more like *makers* than *scientists*: BDB practitioners don't care too much to know, or tell us, what they are doing – they are doing it. (p. 72) His analysis was influenced by microbiologist Carl Woese's earlier evaluation that there was a serious "lack of vision" in recent developments in biology: (Woese, 2004, p. 2)

> A society that permits biology to become an engineering discipline, that allows that science to slip into the role of changing the living world without trying to understand it, is a danger to itself. Modern society knows that it desperately needs to learn how to live in harmony with the biosphere. Today more than ever we are in need of a science of biology that helps us to do this, shows the way. An engineering biology might still show us how to get there; it just doesn't know where "there" is.

The influence of technology in promoting the peril that Woesel dramatically expresses is clearly stated by (Callebaut, 2012) in a single sentence: Characterizations of bioinformatics often equivocate between the biological information 'contained' in the genetic code, and the information that is the daily bread of information scientists (who may never bother about the differences between, say, syntactic and semantic information). (Callebaut, 2012, p. 72) I am not sure that information scientists "may never bother about the differences between syntactic and semantic information," but I agree that, because concepts must be *operationalized* to be processible as information in computing systems (Bonino and Tripodi, 2021), semantics must be *"syntacticized"*, or formalized.

An education in the philosophy of science and technology holds the promise of making us all, researchers, more *aware and sensitive* to what we are doing, as well as to how, why, under which conditions and limitations, and for what reasons and ends. This is what Callebaut and Woesel very eloquently call for in their papers: more *reflective* processes of knowledge production. The study with Voyant has shown me how much *reflection* is stirred when close and distant readings of *discourse* are compared and composed. I thus have reasons to believe that the reprogramming of the invisible discipline could begin in universities, with the offer of different kinds of reflective practices using these two modes of reading. In this way, we would gain more familiarity – if not proficiency – with the two "conditions of knowing" (Moretti, 2013). And, who knows, if reflective practices like these can contribute to reprogramming the invisible discipline, they can also have a role in *saving us*, as a society, from being the reprogrammed ones.

Acknowledgments

I would like to thank Maria das Graças Volpe Nunes and Simone Diniz Junqueira Barbosa for their thoughtful comments and questions regarding my study with Voyant and several drafted parts of this article. I still don't have the answer to most of the questions they have asked. I hope that we will have a chance to look for them together some time in the future.

— o —

Notes

[a] The Portuguese expressions for the seven categories are the following:

THREATS AND RISKS
afli* | ameaça* | arma* | bélic* | conflito* | desigual* | econ* | ego* | empreg* | espada* | guerra* | injustiça* | leta* | militar* | necessitad* | perigo* | pobre* | preconceit* | risco* | terrorista* | trabalhador* | trabalho*

PROGRESS
avan* | crescimento | desenvolvi* | dinâmico* | efici* | expan* | facilit* | fáceis | progre*

TECHNOLOGY
"aprendizagem de máquina" | "inteligência artificial" | "inteligências artificiais" | "machine learning" | "nova tecnologia" | "novas tecnologias" | algor* | comput* | dados | digita* | internet | sistemas | tecno*

HUMANITY
homem | homens | human* | jove* | mulher* | mundo | pessoa* | povo* | sociedade*

VALUES, ETHICS, MORAL
"boa vontade" | ambient* | amor* | bem | convivência | dign* | educ* | equitat* | etic* | fratern* | justiça* | justo* | liberdade | moral* | pacif* | pacífic* | paz* | solid* | valor* | étic*

RELIGION
deus | crente* | crist* | divin* | dádiv* | graça | relig*

REGULATION
reg* | jur* | lei* | lega* | legisla* | respons* | limit* | chefe* | govern* | aten*

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, June 23rd.

Boell, S. K. and Cecez-Kecmanovic, D. (2014). A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for Information Systems*, 34.

Boell, S. K. and Cecez-Kecmanovic, D. (2015). On being 'systematic' in literature reviews in is. *Journal of Information Technology*, 30(2):161–173.

Bonino, G. and Tripodi, P. (2021). Distant Reading and the Problem of Operationalization. Goldilockean Considerations. *COSMO*, 18:187–196.

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1):69–80.

de Souza, C. S. (2005). *The Semiotic Engineering of Human-Computer Interaction*. Acting with Technology. The MIT Press, Cambridge, MA.

de Souza, C. S. (2017). *Semiotics and Human-Computer Interaction*, chapter 2, pages 33–49. John Wiley & Sons, Ltd.

de Souza, C. S., Cerqueira, R. F. G., Afonso, L. M., Brandão, R. R. M., and Ferreira, J. S. J. (2016). *Software Developers as Users. Semiotic Investigations in Human-Centered Software Development*. Springer International Publishing, London, 1 edition.

de Souza, C. S. and Leitão, C. F. (2009). *Semiotic Engineering Methods for Scientific Research in HCI*, volume 2 of *Synthesis lectures on human-centered informatics*. Morgan & Claypool, San Rafael, CA.

Delaney, A. and Tamás, P. A. (2017). Searching for evidence or approval? a commentary on database search in systematic reviews and alternative information retrieval methodologies. *Research Synthesis Methods*, 9(1):124–131.

Fuchs, C. (2017). From digital positivism and administrative big data analytics towards critical digital and social media research! *European Journal of Communication*, 32(1):37–49.

Glass, E. R. (2018). *Software of the Oppressed: Reprogramming the Invisible Discipline*. Phd dissertation, New York, NY.

Glass, E. R. (2021). Reprogramming the invisible discipline. In McGrail, A. B., Nieves, A. D., and Senier, S., editors, *People, Practice, Power Digital Humanities Outside the Center*, Debates in the digital humanities, chapter 2, pages 24–42. University of Minnesota Press, Minneapolis.

Jones, M. L. (2018). How We Became Instrumentalists (Again): Data Positivism since World War II. *Historical Studies in the Natural Sciences*, 48(5):673–684.

Marks, T. and Le, A. (2017). Increasing article findability online: The four cs of search engine optimization. *SSRN Electronic Journal*.

Meyer, E. T. and Schroeder, R. (2015). *Knowledge machines : digital transformations of the sciences and humanities*. Infrastructures series. The MIT Press, Cambridge, Massachusetts.

Moretti, F. (2007). *Graphs, maps, trees*. Verso, London, paperback edition edition.

Moretti, F. (2013). *Distant Reading*. Verso, New York.

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, 37.

Puljak, L. and Lund, H. (2023). Definition, harms, and prevention of redundant systematic reviews. *Systematic Reviews*, 12(1).

Rockwell, G. and Sinclair, S. (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. The MIT Press, Cambridge, MA.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6):887.

Schilhan, L., Kaier, C., and Lackner, K. (2021). Increasing visibility and discoverability of scholarly publications with academic search engine optimization. *Insights the UKSG journal*, 34.

Shaffril, H. A. M., Samsuddin, S. F., and Samah, A. A. (2020). The abc of systematic literature review: the basic methodological guidance for beginners. *Quality & Quantity*, 55(4):1319–1346.

Skees, M. (2020). A new traditional theory: Fetishizing big data analytics. *Constellations*, 29(2):146–160.

Woese, C. R. (2004). A new biology for a new century. *Microbiology and Molecular Biology Reviews*, 68(2):173–186.