Towards a culture of replication in the Digital Humanities: an understanding of transparency and openness practices in published DH papers

Stephen Tiefenbach Keller

Department of Languages, Literature and Cultures

McGill University, Montreal, Quebec

December 2023

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Master of Arts

Contents

Abstract

Valid scientific claims are made based on replicable observations and, in this regard, the replication of published research is an important form of scientific validation. Although the broader scientific community is aware of this and replication has been deemed a cornerstone of the scientific method by the philosophy of science, little incentive exists to promote and facilitate the practice of replicating scientific studies as well as share replication results. This led to what has been called a "Reproducibility Crisis" in science and efforts are now underway to understand and remedy this crisis. The goal of this thesis is to bring current research being conducted in metaresearch - the study of scientific research itself - to the field of the Digital Humanities to understand if it is also affected by issues observed in other fields, and explore topics and issues related to replication in the context of DH. Two studies were conducted to achieve this goal: a survey of papers published in DH and literary criticism journals in 2021 and the replication of published DH research project that analyzed changes in the lexicon and sentiment of popular US songs. The results of the survey indicate that DH exhibits similar transparency indicators observed in other disciplines: half of the 110 papers that relied on empirical data were available as open access; roughly a third shared the code used for data analysis and roughly two thirds shared the data used in the study. Better transparency indicators and journal policies that encourage authors to adhere to a culture of replication would facilitate replication efforts such as the one conducted for this thesis, by reducing time and energy needed to recreate data and code used to validate and extend results.

Résumé

Les affirmations scientifiques valables sont fondées sur des observations reproductibles et, à cet égard, la reproduction des recherches publiées est une forme importante de validation scientifique. Bien que la communauté scientifique dans son ensemble en soit consciente et que la reproduction ait été considérée comme une pierre angulaire de la méthode scientifique par la philosophie des sciences, il existe peu d'incitations à promouvoir et à faciliter la pratique de la reproduction des études scientifiques ainsi qu'à partager les résultats de la reproduction. Cette situation a conduit à ce que l'on a appelé une "crise de reproductibilité" dans le domaine scientifique et des efforts sont actuellement déployés pour comprendre cette crise et y remédier. L'objectif de cette thèse est de transposer les recherches actuelles menées dans le domaine de la métarecherche - l'étude de la recherche scientifique elle-même - au domaine des humanités numériques afin de comprendre si elles sont également affectées par les problèmes observés dans d'autres domaines, et d'explorer les sujets et les problèmes liés à la réplication dans le contexte des humanités numériques. Deux études ont été menées pour atteindre cet objectif : une enquête sur les articles publiés dans les revues de DH et de critique littéraire en 2021 et la réplication d'un projet de recherche publié sur la DH qui a analysé les changements dans le lexique et le sentiment des chansons populaires américaines. Les résultats de l'enquête indiquent que la DH présente des indicateurs de transparence similaires à ceux observés dans d'autres disciplines : la moitié des 110 articles qui s'appuyaient sur des données empiriques étaient disponibles en libre accès ; environ un tiers partageait le code utilisé pour l'analyse des données et environ deux tiers partageaient les données utilisées dans l'étude. De meilleurs indicateurs de transparence et des politiques journalistiques encourageant les auteurs à adhérer à une culture de réplication faciliteraient les efforts de réplication tels que ceux menés dans le cadre de cette thèse, en réduisant le temps et l'énergie nécessaires pour recréer les données et le code utilisés pour valider et étendre les résultats.

Acknowledgments

I would like to thank Prof. Andrew Piper for introducing me to the issues and topics that eventually led to choosing replication and replicability/reproducibility this thesis' main subjects as well as helping me understand how much knowledge and truth we can extract from the methods used in the Digital Humanities. These lessons, I am sure, will extend well beyond the boundaries of this thesis. I would also like to thank him for the guidance in this project and the ideas and suggestions that helped it move forward.

Also, I would like to thank Glenda for the support and patience during the research and writing process of this thesis and cat Anya for sitting beside my computer and keeping me company while I conducted this solitary endeavor.

List of Tables and Figures

1. Introduction

> *"Science is an ongoing, communal conversation and a joint problem-solving enterprise that can include false starts and blind alleys(...). Scientific results should be subject to checking by peers and any scientist competent to perform such checking has the standing to do so."*
>
> **(**National Academies of Sciences, Engineering, and Medicine, 2019, p. 32)

The idea for this thesis came from a critical moment in my Digital Humanities program: the moment I came across a published project that was exactly what I had in mind for my thesis. My immediate reaction was to start thinking of a new subject to study. There is no value in doing what has already been done or discovering what is already known, I thought. This dismissal, it turns out, is institutionalized in how scientific knowledge is produced and affects many disciplines. The second source of inspiration came from a question I had asked myself since the time I started thinking about pursuing a master's degree: what kind of knowledge can we generate with computers and how can we know that such knowledge is valid and correct? This question, especially its second part, relates to how scientific research is performed, and it also connects to the dismissal of my original research idea. Replication of scientific research, from here onwards the main subject of this thesis, is valuable as a tool for validating what is thought to be known. It helps displace trust from a source of authority attached to a scholar onto something external to that individual: data and its associated quantitative methods.

Knowledge production in the humanities relies on close readings and case studies. Researchers move from the single to the whole by generalizing. In fact, the previous two sentences are examples of generalizations: I did not back up these claims with any actual data to support them. Piper (2020a) shows that roughly half of all sentences in the introductions and conclusions of literary studies, history and sociology papers contain generalizations (this figure drops to roughly 40% in more conservative predictions) (p. 37). Piper proposes “transparency, openness, and self-assessment" (p. 54) as a way of dealing with generalizations. By being transparent and open about the evidence used to build arguments and theories, it is possible to put boundaries on generalizations. And as additional evidence is gathered, these boundaries can gradually and incrementally expand. This thesis will focus on the mechanisms that permit transparency and openness to be achieved and on replication as validation of evidence that has already been gathered. This topic is of special interest to the Digital Humanities and its reliance on quantitative and computational methods to study the humanities. DH’s dependence on data and code as a way of obtaining its results brings the topics of reproducibility and replicability (the ability to obtain consistent results and the availability of data and code) to the forefront.

My interest in the topic of reproducibility and the questions and issues associated with it emerged when I first got in contact with Brian Nosek and the Open Science Foundation (OSF) work, especially the Reproducibility Project: Psychology (Open Science Collaboration, 2015). It quickly became evident to me that a lack of focus on the replication of published research and relying only on peer-review as the main source of evaluating the validity of the claims made by authors has become one of the Achilles Heels of science. Evidence gathered by the OSF and others has shown that a significant amount of research is not easily available and/or accessible for researchers willing to conduct replications (low reproducibility) and, when replications are

conducted, the same results are not obtained (lack of replicability). This has led to what has now been called the Reproducibility Crisis, an issue that has the potential to affect the credibility of science. This does not mean that the scientific method is to blame, but that the tools available for science to self-correct - replication being one of them - are not being used as they should. Not only is replication not being used with the frequency it should, but there are also no incentives for researchers to engage in replication efforts.

To begin connecting the topic of replication with the humanities and, more specifically, the Digital Humanities, I will refer to a 2019 study by Nan Da that generated several academic responses to it and fostered a prolific discussion with scholars arguing for and against computational methods. I will not focus on any of her methods of inquiry as that has been done already (see, for example, Hermann et al., 2020; Piper, 2020b; and the long list of responses on "Computational Literary Studies: A Critical Inquiry Online Forum") and is also outside of the scope of this thesis but will, instead, dwell on two points made by the author. Da focuses on a subfield of the Digital Humanities called Computational Literary Studies (CLS), a method of inquiry in literary studies that, according to Da, runs “computer programs on large (or usually not so large) corpora of literary texts to yield quantitative results” which are subsequently evaluated using statistical methods for measuring their significance, and then visualized and used as the basis for making literary claims (p. 601-602).

The first point made by Da and that I will review is when she states that CLS has “adopted an approach to critical contribution characterized by modesty, supplementarity, or incrementality, reframing setbacks as a need to modify methodology and generate more testing” (p. 602) a scholarly posture that leads to a ”‘strategic incrementalism,’ a bad-faith pragmatism

that merely justifies the lack of a method and whose epistemological false modesty and positivistic aspirations suggest that it is at base a pseudoscience" (Mulligan, 2021). Contrary to this position, I would argue that it is precisely the possibility of CLS to rely on incremental knowledge to modify and improve its own methodologies that makes CLS (and other DH subfields that uses quantitative data and methods) not a pseudoscience but a scientific field that allows claims made by its researchers refutable (Popper, 2002) by allowing the collection of more data and evidence to corroborate (or not) the theories derived from those claims (Ioannidis, 2005; Goodman et al., 2016).

This thesis aims to promote replication as a tool for verifying claims (potentially refuting them) and methods, a tool that helps increase our confidence in what has been published and accepted as truth, and not as a tool that should be used to dismiss claims and, more broadly, entire fields of research and inquiry. Da questions, when discussing topic modelling, the efficacy of that method because her results differed from the results of the original study she attempted to replicate (pp. 628-629). Instead of using that opportunity to understand what may have caused the difference in results and propose ways to improve the method, it is simply stated that it did not pass the reproducibility test, thus rendering topic modelling ineffective. As it was said on the quote at the very beginning of this introduction, scientific inquiry is a communal problem-solving enterprise, it requires researchers to work together and not against each other, should it wish to succeed. One of the main takeaways of this thesis is that this communal problem-solving spirit should be a guiding principle for those willing to engage in replication efforts.

To contrast with the use of replication as a tool for dismissing claims and fields of inquiry, what follows are three examples of replication conducted by Digital Humanities researchers that augments or attempts to correct previous research. These were conducted in

what I consider to be the proper ethos that should guide such research enterprises. The first example is Piper's (2022) replication of a study conducted by Langer et al. (2021), using data sets and methods different to those used by the original authors. In the original study, Langer et al. analyzed a corpus of literary works ranging from the years 1705 to 1969 and taken from Project Gutenberg and observed – via the use of a biodiversity word dictionary - a decrease over time in the use of words associated with biodiversity. They linked this decrease to the rise of urbanization and industrialization, two historical processes that have created a separation between humans and nature and alienated the former from the latter. Piper replicated this study and applied the original dictionary method as well as a machine learning approach to two new datasets and observed results that contradicted the original findings. What I want to focus on here is that, instead of dismissing the original methodology and/or claims, Piper attributes the difference in results to corpus construction and points out the need to further investigate the issues relevant to these studies. This serves to illustrate and reinforce the incremental nature of research and is also a case for replication as a method of scientific verification and validation conducted in a manner that is in line with the communal ethos that science needs for it to succeed as a problem-solving and knowledge-generation device.

In the second example, Rizvi (2021) contests the results of a study conducted previously by Craig (2017). Rizvi starts the paper by immediately positioning it as a challenge of Craig's study, one that is, itself, a challenge of another study on Shakespeare and authorship. Rizvi follows this initial remark by stating that the piece's goal is not to rebut the claims made by Craig or to argue for or against any of the points made by either Craig or the challenged author, instead it is to warn researchers against the improper use of a statistical tool called t-test. For the purposes of this thesis, I won't delve into t-tests and the broader scholarly discussion which

Rizvi's piece revolves around (that of Shakespeare and authorship). Instead, I want to focus on the fact that it is possible to critique the use of a method without dismissing it or dismissing the field of the Digital Humanities and/or its subfields. Rizvi argues for conducting experiments with rigorous use of statistics so claims derived from statistical results are valid. He also commends Craig for advising their readers not to generalize too much from the results shown in his paper, to which I will take the opportunity to emphasize the need to place boundaries on the knowledge we can derive from our evidence. It is through the gathering of more evidence that we can expand these boundaries and move towards more generalizing statements.

The third and last example is a rebuttal of a replication conducted by Pervez Rizvi (the same author from the previous example). Egan et al. (2023), creators of a method for text authorship attribution called Word Adjacency Netword (WAN), a method that compares the proximities of high-frequency words across texts to establish their authorship, (Brown et al., 2022), wrote their piece with the goal of defending their method against Rizvi's claims (Rizvi, 2023a and Rizvi, 2023b) that WAN does not produce the knowledge that it is supposed to produce (that of attributing the authorship of texts) and that it is simply a word counting tool hidden behind a layer of superfluous mathematics. Egan et al. argue that Rizvi's claims are the result of a lack of understanding of the mathematics behind their methods and also that he conducted a replication that omitted key aspects of the WAN method (p. 2). But even in the case of this counter-argumentative piece that was built on the case of one's alleged misunderstanding or lack of understanding of underlying mathematical concepts of a method there are silver linings. The authors conclude their paper by looking at the positive outcomes of Rizvi's replication: they had the opportunity to clarify some of the key aspects of their methodology and improve the way they explained them. And, more importantly, they state that this process of

critique, argument and counterargument is important for the advance of their field (authorship attribution), it "builds upon each new advance, and abandons approaches that turn out to be fruitless" (p. 13). Their statement is in accordance with the idea defended in this thesis, that of the incremental nature of science and of replication as a tool that contributes for such.

Moving forward towards the second point that was made by Da, one that is simply a footnote on her paper but relates directly to the study that will be conducted in this thesis. She points out to the importance of having access to computational code and data as a way of checking the validity of CLS work (p. 602) and comments that it took her nearly two years to request and retrieve data and code for the projects she attempted to replicate. According to her, authors contacted by her either did not reply to her requests, could not or did not want to share data and code or only provided the necessary material after repeated requests. Although this is an isolated, anecdotal case, there is evidence from the field of metaresearch showing that this is indeed a common issue faced by researchers (Alsheikh-Ali et al., 2011; Collberg and Proebsting, 2016; Hardwicke et al., 2020; Iqbal et al., 2016; Raghupathi et al., 2022; Wallach et al., 2018). In an ideal scenario, data and code would be easily available and accessible in public databases, but at the very least, researchers willing to replicate a study should be able access such materials by contacting the original authors. But even if the original author or authors state in a paper that data and/or code is available upon request, that does not guarantee that said data and code will actually be made available. This issue encountered and described by Da, and all the other forms in which it can be manifested, is an important factor that hinders replication and directly affects reproducibility indicators in science (i.e., the ability to recompute a data analysis and reobtain the same results that had been obtained previously) and motivated me to investigate data and code availability in the Digital Humanities.

It is my hope that, by the end of this thesis, the need for and the relevance of the incremental nature of scientific knowledge acquisition is clear, and that setbacks and incorrect results are, not only normal occurrences during experiments and observations, but also desired. They are stepping stones towards more refined and robust scientific theories. The focus of this thesis is on replication, as it is a crucial tool for ensuring, or at least improving, the confidence in our claims and theories.

The belief that motivates this thesis is the idea that, when we define Digital Humanities as a method of critical and scholarly inquiry that embraces "computationality" (Berry, 2012, p. 6) and, thus, quantitative methods to understand large-scale cultural, social and political processes via massive quantities of data (p. 13), for this sub-field of the Humanities to succeed, it needs to abide by the best practices promoted by open science initiatives and corroborated by current research. If, instead of only one, several scholars can achieve the same results and reach the same conclusion when answering the same research question, then our confidence that that answer is true increases. This ensures that the knowledge produced by the Digital Humanities does not break under pressure.

1.1. Goal

The goal of this thesis is to bring current advances being made in the field of metaresearch to the Digital Humanities with the expected outcome of eventually improving how research is conducted in it. Metaresearch is the use of scientific methods to study scientific

research practices and this thesis will focus on the theme of reproducibility, or the process of verifying research practices by evaluating the sharing of data, methods and code as well as the obstacles and methods to encourage and improve sharing (Ioannidis et al., 2015).

To achieve the above goal, two projects that complement each other are proposed. The first is a survey of papers published in Digital Humanities journals with the objective of evaluating the presence of scientific transparency and openness indicators in both the surveyed papers as well as the journals they were published in. And the second project is a replication of a Digital Humanities study with the objective of, not only verifying the published results and claims, but also identifying and better understanding the difficulties associated with conducting a replication project in DH.

1.2. Research Questions

To advance towards the previously proposed goal, the following research questions are posed. RQ 1 through 3 are to be answered via the survey of DH journals and papers while RQ 4 is to be answered with the replication effort:

(1a) *Do Digital Humanities focused journals encourage scientific openness and transparency?*
(1b) *Do literary criticism focused journals encourage scientific openness and transparency?*

(2a) *Does published Digital Humanities research follow principles of scientific openness and transparency?*
(2b) *Does published literary criticism research follow principles of scientific openness and transparency?*

(3) *Do journal policies that encourage open science practices correlate with an increase in data and code shared by researchers in published papers?*

(4a) *Does the direct replication of a published Digital Humanities research project confirm the original findings?*

(4b) *Does the conceptual replication of a published Digital Humanities project, using methods other than the ones used by the original author, confirm the original findings?*

2. Literature Review

The key term for this thesis is “replication” which the Open Science Collaboration (2014) defines as "an attempt to replicate the original observation using the same methods of a previous investigation but collecting unique observations" (p. 300). Peng (2020) defines “replication” as “the independent investigation of a scientific hypothesis with newly collected data, new experimental setups, new investigators, and possibly new analytic approaches” (p. 4) and Jasny et al. (2011) as “the confirmation of results and conclusions from one study obtained independently in another”. Each of these definitions ascribe a slightly broader or narrower scope to replication. According to Nosek and Errington (2020), most definitions of the term focus on the repetition of technical methods and in the reobservation of previously made observations as a verification method. Instead, they propose a new definition that better fits the role of replication as a method for advancing scientific knowledge and confronting existing theories and understanding with new evidence. To that end, the authors propose replication as “a study for which any outcome would be considered diagnostic evidence about a claim from prior research” (p. 2). Based on their definition, two possible outcomes exist, in the first results consistent with a previously made claim increases the confidence in that claim, and in the opposite case, inconsistent results decrease the confidence.

The literature makes a distinction between two types of replications: exact and conceptual (Hudson, 2021; Stroebe and Strack, 2014). In the first, an experiment is rerun while keeping everything as similar as possible to the original result. By operationalizing all variables in the exact same way in both the original experiment and the replication, the results are expected to be the same, and, therefore, the first result observed was likely not due to chance or any unforeseen causes. Conceptual replication goes one step further by operationalizing the variables using

different measures to test the same hypothesis as the original study. A conceptual replication can be useful in cases where an exact replication might replicate an observed effect which was, in fact, an illusory effect (i.e., the replication simply replicated a systematic error). Feest (2019) argues that investigating systematic errors in original studies (at least in the social psychology context studied by the author) is a more effective way to develop a better understanding of the concepts being investigated. Stroebe and Strack (2014), similarly, argue that effects observed in an original study may fail to replicate due to the idiosyncrasies of the original study (including cultural context, participants, experimental setting, etc.). Hudson argues, in response, that exact and conceptual replications are useful tools for diagnosing the presence of systematic errors and that at least one replication is necessary to understand if a systematic error exists in the first place. For Nosek and Errington (2020) conceptual replications usually fail to be a study to fall under their definition of replication and are, instead, a study to attempts to test the generalizability of claim. This is due to these studies not being designed to revise confidence in a claim in case of inconsistent results being found, instead of their expected lessening in confidence, inconsistent results are used as a way of delimitating the boundaries of a claim.

A term that is close to that of ‘replication’ and that was coined by computer scientist Jon Claerbout (Goodman et al., 2016; Claerbout and Karrenbach, 1992) is ‘reproducible research’, in which authors of original research make available all data and code required to re-run an analysis (Barba, 2018). According to Rougier et al. (2017), to replicate a result is to write new code and obtain a result that is similar enough to what was originally published. This new code is written based on the description of the models and methods as provided by the original author. And to reproduce a result is to run the same code and data and obtain the same results. The goal of a

reproduction is the verification that a computation was performed correctly by the original research team.

The pair of terms “replicability” and “reproducibility” are, therefore, directly related to the concepts of “replication” and “reproduction” described above. Reproducibility is achieved when data and code are available for a second researcher to recompute an analysis and obtain the same results and the term can be used interchangeably with the term “computational reproducibility” (Open Science Collaboration, 2014, p. 45). Reproducibility is, then, dependent on code and data being, ideally, publicly available. Replicability is achieved when “consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data” (p. 46) are obtained. These two terms, “replicability” and “reproducibility”, are used with reversed meanings in certain disciplines or with no distinction in meaning whatsoever (Barba, 2018), which causes confusion when a formalized terminology across scientific disciplines is attempted. This thesis will use the terms as defined in the beginning of this paragraph and a final note should be added that more work is needed until the broad scientific community agrees on what reproducible research is and how the different concepts linked to it relate to each other (Peng, 2020 and Goodman, 2016. Also note that Goodman designates “reproducibility” as “methods reproducibility” and “replicability” as “results reproducibility”).

This literature review will be organized as follows: sections 2.1 and 2.2. focus on evaluations of the self-correcting mechanisms used by science. Based on Vazire and Holcombe’s (2021) taxonomy of these mechanisms, section 2.1 will review transparency, with a focus on open data, code, materials and methods, while section 2.2. will focus on indicators of critical appraisal in science, with a focus on reproducibility and replicability. Section 2.3 will review

what has been called the “reproducibility crisis”, a consequence of less-than-optimal transparency and reproducibility indicators. Section 2.4 will review efforts being made towards improving transparency and reproducibility and the last section, 2.5., will briefly go through the Open Science movement, which aims to improve transparency in science.

2.1 Transparency

Based on a sample of 250 articles published between 2014 and 2017 in social science journals, more specifically, journals classified as "Economics, Econometrics and Finance", "Psychology", "Business, Management and Accounting" and "Social Sciences", Hardwicke et al. (2020) evaluated the sampled articles based on transparency and reproducibility indicators. They found that 40% of the papers were available as Open Access and did not have access to 6% of the total articles, even though they had access privileges via their academic institution. The researchers encountered two broken URLs where they were supposed to find electronic supplementary material, eight broken URLs where raw data was supposed to be available and one broken URL where an analysis script should have been digitally accessed. The vast majority of the relevant 156 papers did not provide material (135 out of 156), data (126 out of 156) and analysis script (154 out of 156) availability statements and none of the studies were pre-registered. Only 2 out of the 156 articles were self-identified as a replication of another study.

To evaluate data availability in high-impact journals, Alsheikh-Ali et al. (2011) reviewed the public availability and data sharing policies in the 50 journals with the highest impact factors as calculated by the Journal Citation Reports. The authors analyzed data availability in the first 10 articles published by these journals in 2009 for a total sample size of 500 research articles.

Most of the journals (88%) had a statement with instructions to authors on public availability and sharing of data. These statements varied greatly on how strict the transparency policies were, ranging from requiring all primary data to be made available to only requiring an author statement indicating that data can be made available to other researchers upon request. 149 of the total 500 papers were not subject to any data availability policy and none of these papers made their data fully available. Of the remaining 351 papers, 59% did not fully follow the policies from the journal they were published in. Concerningly, only 9% of the 500 papers made all data publicly available online.

Iqbal et al. (2016) evaluated 500 randomly selected papers published from 2000 to 2014 with a PubMed indicator with the goal of evaluating transparency in published biomedical literature. Out of all 500 selected papers, 441 were published in biomedical journals and were then evaluated. Of these 441, less than a fifth were published with full open access. 268 of these papers contained empirical data and of these, only one provided a link to a full study protocol. None of them provided full access to raw data, two provided a broken link to where data was supposed to be found, and four provided partial access to data. Only four articles clearly stated that the research was a replication attempting to validate previous published results. Roughly 45% of the articles were unclear whether their results were novel or a replication. Iqbal et al. reported a decrease in articles that contained no statement of conflict from 94.4% in 2000 to 34.6% in 2014. The authors acknowledge that, in theory, it may be possible to obtain data, protocols, code and clarification from authors by communicating with them, however this potential communication should not replace transparency in published research. A continuation of Iqbal et al.'s study was published in 2018 by Wallach et al. evaluating transparency indicators in biomedical literature published from 2015 to 2017. In a sample of 104 papers that contained

empirical data, only one article provided a link to a full study protocol. 19 articles mentioned some level of public data availability while none of them mentioned any sharing of code. Only five out of 94 articles published in biomedical journals were replications attempting to validate previous published results. When comparing the results of their research and Iqbal et al.'s, the authors observed an increase in the availability of public data and total attempts at replication, while there were no changes in the availability of full protocols. Wallach et al. recognizes an increased awareness among scientists of a need for improved research transparency and reproducibility. Many scientists still are, however, unaware of how they can, in practical terms, improve transparency indicators and more efforts are still needed by journals, funders and researchers for transparency indicators to continue to improve. In biomedical sciences, the demand for these efforts may be perceived as critical when viewed through the lens of Ioannidis (2005) findings that most published research is false, with an increased likelihood of a finding to be false in studies conducted in smaller fields, when effect sizes are small, and in fields where there is greater flexibility in study design, analytical modes and outcomes. In many fields, several research findings are measures of the prevailing biases in that field, with bias being defined as the "combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced" (p. 697). To improve this situation, Ioannidis suggests better powered evidence via larger studies, as well as concentrating efforts on testing relationships with higher probabilities of being true. The author also suggests that research questions should be addressed by multiple research teams and to focus our attention on the totality of the evidence collected and analyzed instead of the results of a single study. This totality can be facilitated by the registration or networking of data collections. It can, then, be

concluded that these suggestions all benefit from improved transparency indicators and journal policies.

In an attempt to quantify reproducibility in business computing papers, Raghupathi et al. (2022) measured documentation transparency in three categories: the sharing of data used, documentation of the research methods and description of the experiment in text and code forms. In a sample of 100 papers that relied on empirical data and were published in the proceedings of the 2019 International Conference on Information Systems, the team observed that 67% of the papers were reproducible with 95% of them being method reproducible, 42% data reproducible, and 28% experiment reproducible. They attributed part of the irreproducibility to be due to the publications of short papers in which authors have limited space to describe their research. Similarly, Collberg and Proebsting (2016) were able to retrieve and build experimental code in 54% of 402 papers backed by code and published in ACM Computer Systems conferences. The two authors identified several factors that could result in code not being retrievable and buildable including the sharing of the wrong version of the code, code that is not shareable because it is not clean enough or was not developed with the intent of being shared, code in which the person responsible for the system is not available anymore (either left the research team or passed away), code that not properly backed up and was lost, and, lastly, code not being open-source (a reason that is more frequent in research conducted outside of the academic context).

To conclude this section, the table below summarizes the main points described in the previous paragraphs. The emphasis is on issues most relevant to this thesis, including open access, access to data, code, protocol and material, the amount of replication studies being published and presence of broken URLs in published papers.

Author(s)	Paper Year	Year Range	Field(s)	Total Sample Size	Relevant Sample Size (research based on empirical data)	Open Access	Broken URLs
Hardwicke et al	2020	2014-2017	Economics, Econometrics and Finance, "Psychology", "Business, Management and Accounting", "Social Sciences"	250	156	40%, no access to 6%	2 for supplementary material, 8 for raw data, 1 for code
Alsheikh-Ali et al	2011	2009	High-impact journals	500	500	N/A	N/A
Iqbal et al	2016	2000-2014	Biomedical	441	268	33 (7.5% of the 441 total) PMCOA PubMed Central Open Access	2
Wallach et al	2018	2015-2017	Biomedical	149	104	37 (24.8% of the 149 total) PMCOA PubMed Central Open Access	N/A
Raghupathi et al	2022	2019	Information Systems	125	100	N/A	N/A
Colberg and Proebsting	2016	2011-2012	Computer Systems	601	402	N/A	N/A

Author(s)	Paper Year	Access to Data	Access to Code	Acceess to Study Protocol	Access to material	Pre-registration	Replication
Hardwicke et al	2020	30 (19%)	2 (1%)	N/A	21 (13.5%)	0 (0%)	2 (1%)
Alsheikh-Ali et al	2011	47 (9%) full data publicly available	N/A	N/A	N/A	N/A	N/A
Iqbal et al	2016	0 full access to data (0%), 5 partial access to data (1.9%)	N/A	1 (0.4%)	N/A	N/A	4 (1.5%)
Wallach et al	2018	19 (18.3%)	0	1	N/A	N/A	5 (5.2%)
Raghupathi et al	2022	42 (42%)	28 (28%)	N/A	N/A	N/A	N/A
Colberg and Proebsting	2016	N/A	54%	N/A	N/A	N/A	N/A

Table 1 *Summary of literature review findings on transparency.*

2.2 Reproducibility

The Reproducibility Project was a large-scale initiative designed and conducted by the Open Science Collaboration (OSC) that reproduced key findings of 100 studies published in Psychology journals (Open Science Collaboration, 2014 and Open Science Collaboration, 2015). The project began in late 2011 and the results were published in 2015. A total of 270 contributing authors participated in the project that was run in an open project format: “project discussion, design, materials, and data are all available publicly” (OSC, 2014, p. 29). The OSC emphasizes that there is a lack of incentives by journals for authors to engage in replication studies, with novel findings and positive results being incentivized instead. Although there do exist journals that specialize in the publication of replications and null results, these are not the majority nor do they carry the same prestige as major publications specialized in the publication

of novel research. The voluntary participation of individual researchers in the Reproducibility Project was incentivized by the individual's own interest in the project's research question, their understanding of the importance of the project towards the overall confidence in the discipline they are all part of (i.e., a sense of duty) as well as the learning and training experience associated with participating in a large-scale open science project, and, finally, the chance of being published as co-author in a project that could (as was understood at the time) have an impact on the field of psychology. All of this serves to illustrate the difficulties associated with the execution of a large-scale replication project.

Out of the 100 original studies, 3 contained null results while the remaining 97 had significant results with $p < 0.05$. The replication effort resulted in significant results in 36% of the studies. It was also observed that the mean effect size of the replications was half of that of the original studies. The Reproducibility Project introduced a subjective "yes" or "no" answer to the question "did it replicate", which had to be answered by each replication team. In the end, 39% of the responses were positive. The project results showed a significant decline in the strength of the evidence published in the original studies and the authors provide that a potential explanation for inflated effect sizes in the original studies are bias in publication, selection, reporting as well as other types. These biases were reduced in the replications conducted by the OSC because "replication preregistration and pre-analysis plans ensured confirmatory tests and reporting of all results" (OSC, 2015, p. 6). The authors, as well as Rodgers and Collins (2021), also note that a successful replication does not necessarily guarantee a true positive result, but simply serves as a confirmation of that result's reliability. Similarly, failure to replicate does not necessarily indicate a false positive. Multiple replications and diverse methods for testing a research question, all pointing in the same direction, should be the hallmark of a true positive

result. In other words, it's the sum of all the collected evidence that matters (Ioannidis, 2005), which implies a need for more incentives from journals and funding agencies towards replication efforts.

The results of a second Reproducibility Project, this time in the field of cancer biology, were published in 2021 (Errington et al., 2021a). The research team was initially set to attempt to replicate 193 experiments published in 53 papers, but several obstacles decreased these numbers and resulted in the replication of 50 experiments published in 23 papers. The obstacles included difficulties in obtaining the original paper's data and code, and difficulties in developing the replication protocols via reading and extracting all the necessary information from the reporting in the original papers. In fact, the authors state that simply by reading a paper and its supplementary information, none of the replication studies could be designed, requiring them to contact the original authors for additional information or clarification (the original papers were published from 2010 to 2012, and improved reporting and transparency guidelines have been implemented since then. Section 2.4 describes some of these advances and improvements). After all the setbacks, the team was able to collect data to measure the replicability of 158 effects (Errington et al., 2021b).

Out of the original 158 effects, 136 were positive while the remaining 22 were null effects. These effects were published both as numerical values (117) and as images/graphs (41) and the first allowed the researchers to apply a more comprehensive replication methodology. Although the original number of null effects was small, there was a clear difference in replicability between positive and null effects: of the effects published with numerical values, 40% of the positive ones replicated while 80% of the null effects did so. The replication effect sizes of positive results were also significantly smaller than the original ones, with the median

effect size being 85% smaller. The authors attribute these rather low replicability results to how the original studies were reported, the complexity of the phenomena, and how the original and replication studies were conducted. The design of the Reproducibility Project in cancer biology does not have the means to determine what the cause(s) for the low reproducibility may be, but the authors reinforce that "there is substantial evidence of how the present research culture creates and maintains dysfunctional incentives and practices that can reduce research credibility in general" (p. 20). Rodgers and Collins (2021), in an editorial commenting the results of Errington et al.'s Reproducibility Project, advocate for more input on studies from statistics experts, especially before data is collected, with the goal of reducing bias, and an increase in preprinting registration for additional scrutiny before peer review. They conclude by pushing for a science that is rigorous, not eye-catching.

In parallel to the Reproducibility Project, the Open Science Collaboration was also involved in another series of replication efforts, called the Many Labs (ML for short). This decade long project resulted in five Psychology replication projects and publications (ML1: Klein et al., 2014; ML2: Klein et al., 2018; ML3: Ebersole et al., 2016; ML4: Klein et al., 2022; ML5: IJzerman et al., 2020). The first project attempted to replicate 13 effects and aimed at better understating variations in reproducibility across different samples and settings. Data was collected across many labs within and outside the US. A total of 36 samples were collected. The aggregate results showed that 10 of the original effects were replicated, one showed weak support for the original effect and two effects failed to replicate. The team stated that their results suggested that replicability has stronger links to the effect itself rather than the sample and setting. The second ML study, similarly to the first, accounted for sample and setting to investigate the replicability of 28 effects and, this time, 125 samples were collected in 36

countries. Results showed that 15 studies replicated (54%) with a statistically significant effect in the same direction of the original study. The team concluded that variation across sample and setting is modestly linked to variations in replicability. The ML3 accounted for the time of the year a participant sample was selected to better understand variation in results across the school year in universities, a setting where many research participants in Psychology research are selected from) and ML4 accounted for the participation or not of the original author in a replication. ML3 showed that changes in sample characteristics across the school year did not affect detection of the effects being investigated and ML4, although inconclusive, showed no difference in author participation in a replication: in the case of the effects investigated, they failed to replicate both with and without the original author participating in the replication. The last Many Labs project conducted more replications on a failed replication conducted in the Reproducibility Project: Psychology (OSC, 2015). This replication project attempted to address comments made by the original authors of the failed replication who offered an explanation on why their study may have failed to replicate. After addressing the comments from the authors, teams from nine universities failed to replicate the effect observed by the original author. Brian Nosek, co-founder of the Center for Open Science and a collaborator on the Many Labs project, comments on the success of ML project by mentioning how the ML concept and ethos have been adopted by other areas of research and other "many" projects have been formed aimed at examining replicability, including "Many Dogs", "Many Birds", and "Many EEGs" (Williamson, 2022).

In a study that replicated the most statistically significant finding from 18 research papers published in two top-tier journals in economics, Camerer et al. (2016) found that the average effect size of the replications was 66% the size of the original studies and 11 replications resulted

in effects in the same direction as those in the original study. The replicability of 61.1% is lower than the expected 92% if all original effects were accurately estimated and lower than the average prediction market belief of 75.2%. The authors believe science will improve as consequence of the current period of self-reflection induced by the "replicability crisis" but mentions that replication is a time and resource-consuming task, even when data and code are readily available. Because of this, Camerer et al. push for the scientific community to facilitate replication by designing and documenting methods that follow good professional norms and journal policies that lead to improved replicability.

Camerer et al. published another replication effort in 2018. This time the team attempted to replicate findings from 21 social science studies published in the journals Science and Nature between the years 2010 and 2015. From each study, one statistically significant effect was chosen to replicate. The results showed that, with an average replication sample size five times higher than the original studies, the average replication effect size was half of that of the original studies and 13 studies showed a significant effect on the same direction of the original study. Replicability varied between 57% and 67%, a similar number encountered by the research team's previous replication study. The authors attribute false positives and inflated effect sizes as causes for the low reproducibility score they obtained. Following Camerer et al. study, one of the teams who had a study that failed to replicate attempted to replicate their own study. The study that did not replicated successfully by Camerer et al. also failed to replicate by the original authors, but their study's most important claim did replicate in a positive way. Camerer et al. commended the authors for their replication effort and mentioned that it helped provide additional insights and understanding of their study's effects.

A concerning trend identified by Serra-Garcia and Gneezy (2021), especially given the context of replicability and reproducibility issues being reviewed here, is that nonreplicable research papers are cited 153 times more than replicable research. This trend was identified, with minimal differences in the journals Nature and Science as well as journals in the fields of economics and psychology. The two authors also found that nonreplicable and replicable research have similar impact. Serra-Garcia and Gneezy accounted for the possibility that the citations of nonreplicable research could be due to the citations of a failed replication but found that not to be the case. Instead, they hypothesize that a possible explanation for the phenomena they observed is a trade-off faced by journal reviewers between accepting papers with interesting versus reliable results. The authors suggest that increasing the cost of publication of problematic data (e.g., publishing the name of paper reviewers and asking for comments on editorial decisions in case a study fails to replicate) as a possible solution for this trade-off.

2.3. The reproducibility crisis

The literature reviewed on sections 2.1. and 2.2. showed that a large number of published research suffers from reproducibility and replicability problems. This serious issue led to what has now been called the "reproducibility crisis". To further show the extent of this problem, a survey conducted with 1576 researchers showed that 70% of them have failed to reproduce other researchers' experiments and more than half of them failed to replicate their own experiments (Baker, 2016).

One of the earlier studies that already pointed to a potential crisis was the influential and provocatively titled "Why most published research findings are false" by Ioannidis (2005).

According to the models developed by the author for calculating the probability for a research claim to be true, a well-designed and well-powered study (the ideal scenario) has an 85% probability of being true, whereas an under-powered study has a probability of 23% and an under-powered and poorly designed study, a probability of 17%. The author suggests that relying on statistical significance as a metric for what good research is, is not only a limited metric, but also one of the causes of the reproducibility crisis. Similarly, Sterne et al. (2001) argues that statistical significance at p=0.05 may indicate correlation between variables that are linked by chance alone, an issue that is aggravated by a tendency to focus on and publish positive rather than null results. In the context of preclinical research, issues related to a reliance on statistical significance are aggravated by Gosselin's (2021) findings that there is "insufficient reporting of tests, sample size and software" (p. 2). Gosselin also identified that the majority of the statistical software packages used in the papers he sampled were proprietary and not open software, an issue that could lead to reduced effectiveness of code sharing policies.

The situation continued to be investigated over the years and, in an editorial, Pashler and Wagenmakers (2012) acknowledged a crisis of confidence in Psychology, due to two fraud incidents that occurred in 2011 (the publication of evidence of extrasensory perception) and 2012. The Open Science Collaboration's Reproducibility Project (2012) was underway at that time and the authors mentioned that it could shed more light on the state of reproducibility in Psychology. The project's results were published three years later (Open Science Collaboration, 2015) and, not only did it help solidify the idea of a crisis in the field, but it also encouraged the conduction of similar studies in other areas. Camerer et al. (2016), Camerer et al. (2018) and Errington et al. (2021) are examples of such studies in the fields of economics, social science and cancer biology, respectively.

In an editorial published on Nature, Collins and Tabak (2014) state that reproducibility in preclinical research areas is susceptible to problems, a risk that is increased in research with animals. According to the authors, there is "a troubling frequency of published reports that claim a significant result, but fail to be reproducible" (p. 613). In a previous Nature editorial ("Announcement: Reducing our irreproducibility", 2013), the author mentions that reproducibility issues start at laboratories, but journals are also responsible for compounding them when they don't scrutinize published results and don't provide enough resources for other researchers to verify published results. Collins and Tabak remembers the fact that science is founded on its self-correcting mechanisms, of which replication is an important one, but admits that, in recent times, the checks and balances that ensures higher confidence in scientific theories and results have not been effective. The need to develop better methods for ensuring these checks and balances are working properly, as well as enforcing effective scientific reporting policies by journals are made more urgent due to a study by Leslie et al.'s (2012). The team found that a large number of Psychology researchers have engaged in questionable research practices, including failing to report all dependent measures and conditions, only reporting studies that worked, excluding data after looking into the impact of doing so. These questionable research practices go all the way up to outright falsifying data. All of these practices reduce the likelihood of a result to be reproduced by another research team.

Questionable research practices are likely influenced by the general tendency in science for authors to publish and cite positive results as well as journal editors' preference for publishing these studies with positive results (which leads to publication and citation bias), a practice that, by itself, is problematic (Sterne et al., 2001; Landis et al., 2012; Duyx et al., 2017; Mlinarić et al., 2017), but this is made even more problematic when questionable research

practices may increase the chances of finding significant positive results that are considered more 'suitable' for publication and, therefore, incentivized, even if these results are, in fact, false positives (Munafò, 2017). The tendency to publish positive results is more prominent in US states that exhibited a higher academic publication per capita, that is, states with strong scientific competition and pressure to publish (a culture of "publish or perish"). These results suggest that publication pressure can lead to publication bias and, in turn, affect the integrity and objectivity of science and contribute to the reproducibility crisis. Two surveys have showed that researchers are largely aware of these issues, with more than 60% of 1576 surveyed researchers saying that pressure to publish and selective reporting are often or always contributing factors to reproducibility issues (Baker, 2016). In another survey, roughly 39% of 467 respondents said that they have been pressured by a principal investigator or collaborator to produce positive data and roughly 63% said that this pressure for positive results affects their research reporting (Boulbes et al., 2018). It has been noted that "competition for grants and positions, and a growing burden of bureaucracy takes away from time spent doing and designing research" (Baker, 2016, p. 454), suggesting that underlying economic issues in science may play a major role in how science is conducted and influence how researchers perform in academia, thus contributing to the reproducibility crisis[1].

A proposed remedy for the reproducibility crisis is the collection of more evidence, or, more specifically, for the scientific community not to focus on results published on single

[1] None of the papers reviewed for this thesis digs deep into the economic issues surrounding how science is conducted and performed in academia, with a few of them sometimes hinting at the topic. But it seems quite reasonable to infer that this is one of, if not *the,* root cause of the issues affecting reproducibility. Would publication bias exist if publication wasn't one of the main ways of accumulating scientific capital by universities and research centers and there wasn't major pressure to publish and keep the scientific economy running? These issues are outside the scope of this thesis, but hopefully this will be taken into more consideration by scholars when the causes of the reproducibility crisis are reviewed, analyzed and discussed.

studies, but, instead, on the totality of the evidence that has been collected to corroborate claims and theories (Ioannidis, 2005; Goodman et al., 2016). Replication is an effective tool for collecting more evidence and making sure that theories are grounded in enough observations so they don't become a flawed description of reality. It is by challenging claims via additional evidence that we can be confident in our scientific truths (Simmons, 2014). To this effect, measures and policies that have been implemented by journals with the goal of improving reproducibility and better scientific reporting will be discussed in section 2.4.

It should be noted that even if reporting and transparency practices improve, that accounts mostly for how researchers conduct their research before the peer-review evaluation. Allison et al. (2016) describe issues they encountered when conducting post-publication peer-review. Despite this practice being an important mechanism for science to self-correct and handle errors in published research, the authors encountered several difficulties when contacting journals, reviewers and authors. These included: editors being unable or unwilling to take action; difficulties in locating where concerns should be directed to, including being unable to contact editors directly; journals and authors being reluctant to issue retractions (in one case, authors refused to retract an article even after a statistical error had been confirmed by an external statistical review); authors being charged expensive fees to publish manuscripts reviewing errors identified in published papers; a lack of standardization on how to request raw data to authors (this relates directly to issues of data transparency discussed previously); informal communication channels being often overlooked, including platforms that allow authors to comment on published research not being moderated by editors. Allison et al. conclude by stating that scientists engage in post-publication peer-review out of a sense of duty, but there is little incentive for them to do so. Addressing this lack of incentive is a good starting point to

encourage more scientists to be engaged in ensuring the self-correcting mechanisms of science are running effectively.

Lastly, with the emergence of machine learning and its use in various disciplines, reproducibility concerns caused by its use have recently come into light. The gravity of the issue has led a data scientist at Mayo Clinic's to state that he is "somewhat surprised that there hasn't been a crash in the legitimacy of machine learning already. But I think it could be coming very soon" (Gibney, 2022, p. 250). Kapoor and Narayanan (2022) have identified data leakage as a widespread issue in ML and one that has led to reproducibility issues. 'Leakage' has been defined as "the introduction of information about the target of a data mining problem that should not be legitimately available to mine from" (Kaufman, 2012, p. 1), an issue that is "a source for poor generalization and overestimation of expected performance" (p. 19). In Kapoor and Narayanan's leakage taxonomy in ML, leakage can occur when there is no proper separation between training and test data; when a ML model contains illegitimate features; and when a "test set is not drawn from the distribution of scientific interest" (p. 5). The authors surveyed and compiled the results from papers that identified leakage in various scientific areas and provided evidence that the issue needs to be addressed. All papers analyzed in computer security, radiology and satellite imaging suffered from leakage whereas papers in other areas suffer from the issue in varying degrees. The duo also provide a case study on published research on civil war prediction, in which the predictions fail to reproduce due to leakage. They also found that, once the errors were corrected, the ML models did not perform better than older logistic regression models, raising concerns about claims on the superiority of ML models. With such evidence available, researchers need to avoid a (new) crisis in confidence fueled by the improper

use of machine learning (Gibney, p. 251). The ML model info sheets[2] proposed by Kapoor and Narayanan can be an effective way to increase reproducibility in ML-based research.

2.4. Towards improved transparency and reproducibility

Calls for establishing a culture of reproducibility and to bring this topic to the forefront of scientific conversations have become stronger due to the rise of computational science, the need to access large bodies of data and the lack of norms that encourage the replication of existing studies (Peng, 2011, King 1995). Recent efforts to improve reproducibility and replicability can be grouped into three categories: social, statistical and methodological (Romero, 2019). This thesis is directly related to the methodological category, specifically the understanding of the need for and the potential effects caused by an increase in transparency indicators. Current journal policies aimed at improving transparency via methodological reforms will be reviewed below.

In 2012, the American Political Science Association introduced new policies related to data access and research transparency as well as updated its ethics guidelines (Lupian and Elman, 2014). The updated guidelines state that "researchers have an ethical obligation to facilitate the evaluation of their evidence-based knowledge claims through data access, production transparency, and analytics transparency so their work can be tested or replicated" (p. 21). These three transparency elements are further promoted by establishing that researchers who produced their own data should provide access to that data or explain why it cannot be done, they

[2] Kapoor and Narayanan (2022) model info sheet can be accessed via the following website: https://reproducible.cs.princeton.edu/#model-info-sheets

should also fully explain the procedures that enabled the generation of that data, and, finally, they should clearly explain the relationship between that data and the conclusions drawn from it. Whenever a replication leads to the challenge of previously made claims, the replication researcher is bound to the same transparency guidelines followed by the original author (i.e., replication data should also be made available). Exceptions to these guidelines are made in cases where data sharing would result in legal issues such as sharing confidential governmental data. The guidelines emphasize “data access and research transparency as an indispensable part of the research endeavor” (p. 21).

The US funding agency National Institutes of Health proposed in 2014 measures to improve reproducibility in pre-clinical research (Collins and Tabak, 2014). These measures have the goal of ensuring that the checks and balances that enable the self-correcting mechanisms of science are allowing the conduction of replications of prior research in an effective way by attempting to remedy an array of factors that have led to a lack of reproducibility. The first proposed measure is mandatory training for post-doctoral fellows on "enhancing reproducibility and transparency of research findings, with an emphasis on good experimental design” (p. 613). A checklist for reviewers was proposed with the intent of ensuring they check for any areas related to experimental design that could lead to irreproducibility. The agency also proposed a method for improving data transparency. A Data Discovery Index would allow researchers to search for unpublished data and, should they use it, the original creator and owner of the dataset would then be cited. Lastly, an online forum was launched with the goal of facilitating comments and discussions on published articles.

The editor-in-chief for the journal Science, Marcia McNutt, published two editorials (2014a and 2014b) in which she mentions a lack of universal transparency guidelines that applies

to all fields and, in the first of the cited editorials, she announces that the journal will ask editors and reviewers to identify papers that followed steps that led to levels of transparency that were high enough that it inspired improved confidence in the published results. These papers would then inform future transparency guidelines. McNutt also announced the addition of editors coming from the statistics community with the goal of reducing the probability of papers with statistical errors to slip into the journal. The inclusion of additional editors with statistical training is an important step towards solving an issue identified by Hardwick and Goodman (2020). In leading biomedical journals, 34% of their surveyed journals (37 journals) never uses specialized editors to review statistical methods and only 23% has a specialized editor for reviewing statistical methods on all papers. These results are not a dramatic change from the results found on a survey published by Goodman et al. (1998) 22 years prior, in which the team also found that, based on the surveyed journals editors' judgement, an important change was made on a manuscript about half the time research papers were submitted for statistical review. The number of changes made in manuscripts and the comparatively low number of editors with statistical training indicate that there is still a large room for improvements in this area, at least in biomedical sciences. Although no equivalent study was found in the context of the Digital Humanities, a post published on the Critical Inquiry blog containing responses to Da's article (2019), included a response by Prof. Taylor Arnold stating that leading DH journals, including Cultural Analytics, Digital Humanities Quarterly, Digital Scholarship in the Humanities among others, have zero trained statisticians on their boards or as editors. He acknowledges this as an issue for a field that relies on statistical methods.

The second editorial published by McMutt (2014b) refers to a set of guidelines for preclinical biomedical research developed after editors from 30 major journals in that field as

well as members from funding agencies assembled to discuss matters related to reproducibility. The set of guidelines ("Principles and Guidelines in Reporting Preclinical Research") requires that, at a minimum, authors should make available upon request by editors or other researchers any data used in a research project (unless ethics won't allow for that) and recommends that data be deposited on public repositories where it can be easily accessed and cited. It is also encouraged that authors share any software used or, at least describe how it can be obtained. Besides guidelines on code and material sharing, the document also suggests that journals should impose no limits to the length of a manuscript's methods sections in order to ensure clear and transparent reporting. Authors should follow a checklist to make sure complete reporting is provided, including a full description of statistics used, number of times experiments were replicated, how sample size was calculated as well as other topics. These reporting guidelines were adapted from Landis et al. (2012), who established a core set of guidelines and suggests that authors should report, at a minimum, on "randomization, blinding, sample-size estimation and the handling of all data". The first step towards implementing these proposed guidelines is for journals and funding organizations to provide clear guidance to reviewers as to what the required standards should be when evaluating a study as well as training scientists and students on proper study design and reporting.

Measures aimed at improving reproducibility like the ones described above were also implemented in 2013 by the journal Nature ("Announcement: Reducing our irreproducibility"). The journal abolished limits to the length of manuscripts' methods section, so authors could be as detailed as needed when reporting their research and authors were also required to be more precise when describing their use of statistics. Lastly, the journal introduced a checklist that authors were now required to fill with information on experimental and analytical design

elements ("Reporting standards and availability of data, materials, code and protocols"). The checklist requires authors to include a data availability statement and inform how other researchers can have access to the "minimum dataset" should they require it to review and extend the results of a study. It also requires statements on availability of materials and code and reviewers have the right to refuse a paper if code required for reproducing results cannot be shared. Certain journals that are part of Nature publishing even engage in peer-review of custom-code and algorithms used in studies. These requirements as well as requirements on reporting replication, trials, protocols and any methodological decisions that may introduce bias in a study were designed with the goal of improving reproducibility. To review the effectiveness of the checklist, a survey was conducted and published five years later ("Checklists work to improve science"). With the vast majority of respondents in agreement that a problem of poor reproducibility in science exists, about half of them agreed that the checklist improved the quality of research published in Nature. Finally, almost 60% of respondents believe that the researchers designing and conducting studies are the ones with the greatest responsibility to improve reproducibility, which implies that better investments in training and education might be required. The editorial concludes that the role journals play in improving reproducibility is to demand transparency from authors. And to that effect, the checklist introduced by Nature was observed to have improved the levels of reporting and transparency in preclinical animal studies (Han et al., 2017), an area that is prone to reproducibility issues (Collins and Tabak, 2014). A significant increase in reporting on randomization, blinding, sample size calculation, and reporting calculation was also observed: an increase that was much higher than the control group of other journals that didn't introduce the checklist (The NPQIP Collaborative group, 2017). The checklist has been considered a simple and practical way to improve reporting transparency.

In 2015, the journal Social Psychological and Personality Science introduced methodological changes aimed at improving reproducibility while keeping undesirable side effects away and costs low (Vazire, 2015). Along with the new changes, the journal started to accept replication studies and novel research accompanied by a replication now has an increased chance of being accepted for publication. Changes intended to improve transparency include requiring authors to explain how sample size was determined, what and how decisions on data exclusions were made and what were the measures for a study's research question variables. All the implemented changes have the goal of ensuring the journal keeps up with the evolving standards and best practices in psychology research.

Joelle et al. (2020) describe a program aimed at improving reproducibility in the Machine Learning focused conference Neural Information Processing Systems (NeurIPS), which is the premiere conference of its kind. The program was comprised of three components: "a code submission policy, a community-wide reproducibility challenge, and the inclusion of the Machine Learning Reproducibility checklist as part of the paper submission process" (p. 3). The code submission policy does not mandate authors to submit code but encourages them to do so after their manuscript is reviewed and accepted. It was observed that this policy of voluntary code submission led to an increase of 25% of authors spontaneously submitting code, going from half of the authors submitting code to nearly 75% doing so in a period of roughly one year. When inquired via a survey, a high number of reviewers answered that having access to code was important in their review process, an indication that code availability is relevant not only to other researchers willing to engage in replication studies, but also to reviewers willing to better understand and assess the manuscript under review. The Reproducibility Challenge component allows independent researchers to verify claims made in the papers published at NeurIPS and

selected high-quality replication reports are then reviewed and published. Joelle et al. notes that there was an increase in the number of individuals voluntarily participating in these replication efforts. And, lastly, the Reproducibility Checklist was first introduced at the NeurIPS 2018 conference "in response to findings of recurrent gaps in experimental methodology found in recent machine learning papers" (p. 11). The checklist includes information such as ensuring authors provide download links to codes for the models and algorithms used in the research and download links for the datasets, complete description of the data collection process, explanation about data exclusion and other points that ensure a clear and thorough reporting. Based on the survey conducted with the reviewers, a third of them considered the checklist useful for their review process.

To help promote openness, transparency and reproducibility, Nosek et al. (2015) describe a set of author guidelines developed during a committee meeting in 2014 at the Center for Open Science. Named the Transparency and Openness Promotion (TOP), the guidelines consist of eight standards with three levels for each, which work in a modular fashion: a journal can pick and choose which standards and levels to adopt. Standards in the guidelines include citation; data, analytic methods, research materials, and design and analysis transparency; preregistration of studies and analysis plans; and replication. Each level increases the stringency of each standard. Recent research has shown that journals have been adopting TOP guidelines at a slow pace. Patarčić and Stojanovski (2022) have analyzed the TOP factor (a metric created by the Center of Open Science to evaluate and measure the degree to which journals adhere to TOP guidelines) of 2000 journals in the physical, social, life, health and multidisciplinary sciences and roughly a fourth of them (n = 455) have not adopted any of the guidelines, another fourth (n = 561) have adopted only one of the guidelines with 70% of them having adopted the Data Citation

standard and only 78 journals have adopted all eight TOP guidelines. Out of 4661 identified TOP guidelines adoptions, Data Citation is the most adopted standard, followed by Data transparency, and Replication is the least adopted standard, followed by Analysis Plan Preregistration and Study Preregistration. Kianersi et al. (2023) evaluated 339 journals in the behavioral, social and health sciences and concluded that TOP "has not been implemented widely by journals in the behavioral, social, and health sciences despite journal endorsement and widespread community support" (p. 18) and most of the TOP standards have not been adopted by the majority of the evaluated journals. The authors also mention that there is currently no standardized way for evaluation TOP implementation by journals. It should also be mentioned that the mere presence of guidelines does not mean that they will be followed. Gabelica et al. (2023) analyzed 3416 articles that contained a data availability statement with 1792 of them stating that authors are willing to share their data upon request. However, 93% of the authors did not respond or declined to share their data.

Social and statistical reforms aimed at remedying the reproducibility crisis are outside the scope of this thesis, but a few initiative examples will be briefly reviewed for an understanding of what types of changes may arise from their implementation. The "Basic and Applied Social Psychology" journal banned null hypothesis significance testing from being included in papers published in it and, instead, asked authors to include strong descriptive statistics (Trafimow and Marks, 2015). The ban, as well as widespread misuse of p-Value, prompted the publication of a statement by the American Statistical Association clarifying and articulating principles related to the interpretation of quantitative results (Wasserstein, 2016). An approach for tackling the "reproducibility crisis" without banning misused statistical methods proposed by Peng (2015) is the increase in data analytics literacy. More researchers trained in data analytics and the proper

usage of statistical methods would ideally lead to a decrease in the "epidemic of poor data analysis" (p. 31) contributing to the current crisis.

Regarding social incentives for researchers to engage in conducting more replication studies, Koole and Lakens (2012) propose three rewards or incentives for psychology researchers willing to engage in replication: copublication as a way of making replication results more visible and accessible to the broader academic community; cocitation as an incentive for authors to cite studies other than those that break new scientific grounds; and elevating replication into becoming a common scientific practice. The third incentive could be achieved by teaching replication as an integral part of the academic curriculum. Koole and Lakens paper eventually led to creation of a three-million-euro investment by a Dutch funding agency, NWO, granted to fund the replication of nine influential studies (De Vrieze, 2017). This pilot project closed in 2022 and funded a total of 24 replication projects via three rounds of funding ("Replication Studies").

Machine learning has been proposed as a tool that can predict replicability and there currently is research being conducted in this space. Altmejd et al. (2019) used replication data from the Reproducibility Project: Psychology (OSC, 2015), Many Labs 1 and 3 (Klein et al., 2014; Ebersole et al., 2016), and replication on experimental economics (Camerer et al., 2016) to train a model that attempts to predict replicability. The model shows an accuracy of 71% in predicting replicability in the test data, a result that is on par with accuracy on predicting replicability by humans. The author's model shows that the two features most relevant for predictability of reproducibility are the original experiment's p-value and effect size and predictability increases by adding variables such as number of authors, paper length and lack of performance incentives, although a simpler model that includes only p-value and effect size tend

to show a level of performance that is on par with the model trained with the full feature set. Gordon et al. (2020) conducted a survey to evaluate the human ability to predict replicability and encountered an accuracy of 73%, on par with Almejd et al. machine learning model, and the authors also found a relationship between an original study's p-value and its replication prediction, a relationship that was also observed on Almejd et al.'s ML model. Gordon et al.'s survey also used replication data from experimental psychology, economics and social science (specifically, OSC, 2015; Camerer et al. 2016 and Canerer et al. 2018; and Many Labs 2, Klein et al., 2018), which helps relate Almejd et al. and Gordon et al.'s results to each other. Fraser et al. (2023) proposed a method that brings replicability prediction to 84% by relying on a team of human evaluators and aggregating their predictions, an accuracy that is significantly higher than Altmejd et al.'s machine learning model. The ML model, the authors claim, could see an increase in accuracy once trained with more data. When comparing a ML model against human subjectivity for predicting replicability, the usual ML benefits apply: speed and cost-efficiency. Altmejd et al. comment that potential synergies between human evaluators and the ML model could be explored and concede that the model's predictability could be affected by changes in research practices and, should the model be employed for use in pre or post publication peer-review, researchers could potentially exploit ways to manipulate the algorithm and increase their study's replicability prediction.

2.5. Open Science

Open science is a movement that aims to improve transparency in the current scientific landscape by addressing issues that have, historically, made transparent and accessible research

difficult to achieve even if a researcher wanted to attain it (Munafò et al., 2017). It has been defined as “transparent and accessible knowledge that is shared and developed through collaborative networks” (Vicente-Saez and Martinez-Fuentes, 2018), a definition that was created by reviewing 75 studies, and reviewing and synthesizing how the term “Open Science” was used by their authors. The goal of a unified definition is for the scientific community to have a clear and shared understanding of what Open Science is, so it can collaborate on achieving what the movement strives to bring to science. To contrast, the opposite of an open scientific environment, a lack of openness, has been said to “reduce the efficiency and veracity of knowledge construction” (Nosek and Bar-Anan, 2012).

When describing a “Scientific Utopia”, Nosek and Bar-Anan (2012) have suggested a scientific status-quo in which all published literature is open access and, therefore, fully accessible to anyone. A scenario that, they argue, would bring financial benefits as well as the free flow of knowledge and information. Access to knowledge is an important contributing factor to scientists and scholars becoming experts in their fields of study and there is a current (albeit timid) shift towards open access with the emergence of respected Open Access (OA) journals such as Public Library of Science (PLoS). PLoS charges a publishing fee from authors rather than a subscription fee from libraries or universities. Open access should not only be associated with a human right that makes the circulation of knowledge possible, but also as something that provides benefits for those who partake in it. OA has been associated with an increase in research impact and research published as OA sees an increased number in citations when compared to similar non-OA work (Willinsky, 2009, p. 22)

Open Data, similar to Open Access, also offers benefits and Barnett et al. (2012), when discussing them, mentions that data collected from 1987 to 2000 on 108 US cities and their

associated daily mortality counts, air pollutants and weather and that was made public on the internet led to the publication of 67 studies that analyzed that data and, as a result, an improved understanding of the health effects of air pollution and heat waves. The data was taken down in 2011 due to privacy issues, even though the mortality information was anonymized. This decision was, according to the authors, negative for reproducible research and the benefits generated by the publicly available data outweighed data-security concerns.

There are many reasons why data may not be made publicly available (a full and thorough review of these reasons is outside the scope of this literature review), with copyright being one of them and one of concern for Digital Humanities researchers. Thompson and Carrera (2021) developed a Digital Humanities course to teach Afrofuturism to students and, as part of the coursework, students were required to engage with culturally significant music. However, US copyright laws restricted the type of music they could use in the classroom and, should they be fined for infringement, the fees could add to US$900.000. The authors suggest four ways of dealing with US Copyright laws when using published recorded music in a pedagogical context. First, to treat the use of audio sample as a form of quotation. This applies if a derivative work is generated from the original music. Second, to rely on databases of copyright-free music. Third, contact artists who would be willing to license their music and, finally, to pay the necessary royalties. This illustrates how DH faces copyright challenges not only when attempting to study human culture, but also teach it.

The United Nations Educational, Scientific and Cultural Organization (UNESCO, 2021) have proposed a series of actions its Member States can take to promote open science in their territories. These include (i) promote the understanding of what open science is and what its benefits are; (ii) “developing an enabling policy environment for open science”; (iii) “investing

in open science infrastructures and services"; (iv) "investing in human resources, training, education, digital literacy and capacity building for open science"; (v) "fostering a culture of open science and aligning incentives for open science"; (vi) "promoting innovative approaches for open science at different stages of the scientific process"; (vii) "promoting international and multi-stakeholder cooperation in the context of open science and with view to reducing digital, technological and knowledge gaps" (p. 6). To achieve these, UNESCO recommends that public funding agencies and governing bodies for the sciences be guided by open science values and principles. Some of UNESCO's proposed actions should help address issues that have been identified as responsible for obstructing the growth of open science, and that will be reviewed next.

What are some of the barriers preventing open science from growing? Arthur et al. (2021) reviewed them in the context of humanities scholarship in Australia and identified barriers in various levels of academia. In the case of humanities researchers, some of these barriers included lack or limited knowledge of open science principles, reasons for archiving their work in repositories, and the quality of open access publishing as well as its benefits (e.g., increased readership and citations) as well as limited training on how to share data and code as well as limited incentives to do so. There is also a lack of understanding of the value of open science and its associated practices at an institutional level, which incentivizes scholarly output via traditional means. The authors state that, although their review focuses on open science in Australia, many of these issues are similar to what is encountered in other countries. In conclusion, for open science to flourish and scientific output to be made available with the public good as its main goal rather than profit, changes need to occur in the many layers that academia are made of.

3. Overview of the state of Open Science practices in the Digital Humanities

3.1. Methodology

For the open science practices survey component of this thesis aimed at answering research questions 1, 2 and 3, I collected data to create two datasets. The first one consists of a list of Digital Humanities and Literary Criticism journals and the second is a list of all the papers published in these journals during the year 2021. The journals were evaluated using the Transparency and Openness Promotion (TOP) guidelines developed by the Open Science Foundation (Nosek et al., 2015). There is currently no standardized way of evaluating TOP adoption (Kianersi et al., 2023), and I checked each journal's publication submission guidelines via their websites. See Table 1 for the complete list of evaluation topics and their respective criteria. Information about the journals' open access practices, date they were founded, and their impact scores were also included.

	Level 0	Level 1	Level 2	Level 3
Citation Standards	Journal encourages citation of data, code, and materials, or says nothing	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used consistent with journal's author guidelines.	Article is not published until providing appropriate citation for data and materials following journal's author guidelines.
Data Transparency	Journal encourages data sharing, or says nothing	Article states whether data are available, and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Analytic Methods (Code) Transparency	Journal encourages code sharing, or says nothing	Article states whether code is available, and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Research Materials Transparency	Journal encourages materials sharing, or says nothing	Article states whether materials are available, and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Design and Analysis Transparency	Journal encourages design and analysis transparency, or says nothing	Journal articulates design transparency standards	Journal requires adherence to design transparency standards for review and publication	Journal requires and enforces adherence to design transparency standards for review and publication
Preregistration of studies	Journal says nothing	Article states whether preregistration of study exists, and, if so, where to access it.	Article states whether preregistration of study exists, and, if so, allows journal access during peer review for verification.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing	Article states whether preregistration with analysis plan exists, and, if so, where to access it.	Article states whether preregistration with analysis plan exists, and, if so, allows journal access during peer review for verification.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies, or says nothing	Journal encourages submission of replication studies	Journal encourages submission of replication studies and conducts results blind review	Journal uses Registered Reports as a submission option for replication studies with peer review prior to observing the study outcomes.

Table 2 *Transparency and Openness Promotion (TOP) guidelines matrix*

The Papers dataset included a field for identifying if a paper refers to quantitative and/or computational research and fields for Open Access, Open Data, Open Code, Open Software, and

a field to identify if a paper referred to a book review. This last field was for reference only: by their nature, book reviews are not quantitative nor computational and a large quantity of them would lead to a higher quantity of non-computational papers in the dataset. I debated whether the book reviews rows should be included in the results or not and, in order for the survey to include a sample of all papers published in 2021, they were left included in the dataset.

Each Open Science practice field in the Papers dataset as well as the 'quantitative/computational' and 'book review' fields, were labeled with either an "Y" for yes, "N" for no and "N/A" for not applicable. There was no differentiation between articles that did not specify if the code and data used were available and articles mentioning that data and code were not available. In both cases the entry was labeled with "N" in the Open Data and Open Code fields.

To identify if authors provided direct access to data and code or instructions on how to obtain them, and to avoid the need to read each article in its entirety, I manually reviewed each article in the database that relied a computational or quantitative methodology and (1) visually scanned them in order to check if there was a section of the paper dedicated to providing information about data and code availability and/or (2) used the Adobe Acrobat Reader's search functionality to search for "data", "corpus", "code", "software" and "available". If a paper indicated that a specific search term could lead to positive results, additional keywords could be used. For example, if it was clear that the authors used Python as their programming language of choice "Python" was used as a keyword.

The collected data was analyzed in a few different ways: all papers; papers from journals not focused on Digital Humanities research; and papers from journals focused on Digital Humanities research. The third category was also broken down by individual journals for a more

comprehensive analysis of Open Science practices in these journals. The data visualizations were generated in Microsoft Excel and customized to the form presented here using Adobe Illustrator. There was no tweaking of the bar graphs' proportions in Adobe Illustrator, and this step served only to adjust colors, typography, and the overall aesthetic of the information visualization[3].

3.2. Results

3.2.1 Results: Open Scholarship practices in DH journals

Nine English language journals were included in the survey (see Table 1). Four of them focuses on the publishing Digital Humanities research - Digital Humanities Quarterly (DHQ), Journal of Cultural Analytics (JCA), Digital Scholarship in the Humanities (DSH) and International Journal of Digital Humanities (IJDH) - and five on publishing literary criticism research – New Literary History (NLH), Modern Language Quarterly (MLQ), American Literary History (ALH), Critical Inquiry (CI) and PMLA. No journals on Humanities disciplines other than literary criticism was included. Information on the year each journal was founded and their respective Impact Score[4] are included in the dataset, but that data was not included, nor would it affect the data analysis for the purposes of this thesis. Information about the peer review process for each journal was also included but that data was also not included in the data analysis. The

[3] It should be acknowledged that the use of Microsoft Excel and Adobe Illustrator, both proprietary software, doesn't align with the overall ethos of this thesis. Due to my experience using these tools in the past, I opted to use them and rely on a workflow that I am already familiar and efficient with. In the future, ideally, I will rely on open software for creating and editing spreadsheets as well as for the creation and manipulation of vector graphics.
[4] Impact Score was collected through the Resurchify website (https://www.resurchify.com/), a portal containing data and information on research-related topics such as journals and conferences. I did not have access to the paid Web of Science site to access Impact Score data and had to rely on Resurchify to relay that data.

presence of a peer review process in all journals (whether single blinded or double blinded) indicates that none of the journals included in this survey, open access or not, are predatory publishers.

Journal	Journal Acronym	URL	Submission URL	Impact Score (2022)	Open Access	Founded
Digital Humanities Quarterly	DHQ	http://www.digitalhumaniti	http://www.digita	0.34	Y	2007
Journal of Cultural Analytics	JCA	https://culturalanalytics.org	https://culturalan	0.57	Y	2016
Digital Scholarship in the Humanities	DSH	https://academic.oup.com/	https://academic.	1.06	Optional	1986
International Journal of Digital Humanities	IJDH	https://www.springer.com/	https://www.sprir	N/A	Optional	2021
New Literary History	NLH	http://newliteraryhistory.or	http://newliterary	0.43	N	1969
Modern Language Quarterly	MLQ	https://read.dukeupress.ed	https://read.duke	0.32	N	1940
American Literary History	ALH	https://academic.oup.com/	https://academic.	0.2	Optional	1989
Critical Inquiry	CI	https://criticalinquiry.uchic	https://criticalinq	1.37	N	1974
PMLA	PMLA	https://www.mla.org/Public	https://www.mla.	0.57	N	1883

Journal	Peer Review
Digital Humanities Quarterly	Single blind. Reviewers can decline to review if they detect a conflict of interest.
Journal of Cultural Analytics	Double blind. Reviewers names are disclosed and published
Digital Scholarship in the Humanities	Peer Reviewed. Process is not described
International Journal of Digital Humanities	Single-blind peer review
New Literary History	Double blind
Modern Language Quarterly	Peer Reviewed. Process is not described
American Literary History	Single-blind reviewed. Reviewers and editor know author, author doesn't know reviewers.
Critical Inquiry	Peer Reviewed by the editors-in-chief in consultation with the editorial team (not mentioned if blind)
PMLA	Reviewed by two readers (blind). Upon recommendation, reviewed by members of the editorial board (blind).

Table 3 *List of journals included in the Open Science practices survey.*

From the DH focused journals, two of them (DHQ and JCA) are fully open access while the other two give researchers the option to publish their research as open access but do not require them to do so. IJDH mentions on their website that they are currently in the process of switching the journal to a full open access policy. In the case of non-DH journals, five of them are not open access while one, ALH, gives researchers the option to publish as open access and requires authors that choose to do so to pay an open access charge of $4385.68 US Dollars[5]. Two of the optional open access journals, DSH and ALH, are published by Oxford University Press and IJDH is published by Springer.

[5] For a list of Open Access fees broken down by journals published by Oxford University Press, see https://academic.oup.com/pages/open-research/open-access/charges-licences-and-self-archiving

Journal	Journal Acronym	Citation Standards	Data Transparency	Analytic Methods (Code) Transparency	Research Materials Transparency	Design and Analysis Transparency	Preregistration of studies	Preregistration of analysis plans	Replication
Digital Humanities Quarterly	DHQ	0	1	1	1	0	0	0	0
Journal of Cultural Analytics	JCA	0	2	2	2	0	0	0	0
Digital Scholarship in the Humanities	DSH	0	0	0	0	0	0	0	0
International Journal of Digital Humanities	IJDH	1	1	0	2	0	0	0	0
New Literary History	NLH	0	0	0	0	0	0	0	0
Modern Language Quarterly	MLQ	0	0	0	0	0	0	0	0
American Literary History	ALH	0	0	0	0	0	0	0	0
Critical Inquiry	CI	0	0	0	0	0	0	0	0
PMLA	PMLA	0	0	0	0	0	0	0	0

Table 4 *Journal policies on Open Science practices measured using OSF's TOP guidelines.*

Each of the nine surveyed journals was evaluated using the Transparency and Openness Promotion (TOP) guidelines created by the Open Science Foundation (Nosek et al., 2015). TOP includes criteria for data citation, the sharing of data, analytic methods, and research materials, preregistration of the research and journal policies relating to submission and publication of replicated research. Each of these topics is evaluated on a scale of zero to four, or levels as per the TOP nomenclature. See Table 1 for a detailed description of each topic and their levels. For the non-Digital Humanities focused journals, all the criteria on all journals are at level 0, meaning that no journal mentions anything relating to the citation of data, the sharing of the data and code used for research, preregistration of research or the publication of replicated research. An interpretation of the meaning of this finding will be discussed in section 4.3. For the Digital Humanities focused journals, the two journals with open access policies, DHQ and JCA, included details on data, code, and research materials sharing. The JCA includes details on where researchers should share their research materials[6] in order for their article to be published while the DHQ asks authors to submit their research materials along with their final paper submission, if the paper includes any supplementary material, but the journal does not require authors to do so before their research can be published. The IJDH, which gives authors the option to publish

[6] The Journal of Cultural Analytics requires all researchers to publish their data, code used and research materials on Dataverse, which is hosted on Harvard University's server. The full repository can be accessed via the following URL: https://dataverse.harvard.edu/dataverse/culturalanalytics

their research as open access and is in the process of switching to a fully open access policy, requires authors to include a data availability statement in their article, and recommends authors to share any data used in their research in a data repository, but does not require them to do so in order for their research to be published. The journal also provides instructions and examples on how to cite any dataset used. IJDH provides information on how to submit supplementary material and states that any file submitted to them will be made available to readers by the journal as submitted by authors (no modifications or editing of any submitted material). The fourth DH focused journal, the DSH, does not provide instructions or guidelines on the submission of data, code, and research materials nor on how to cite the data used in the research. Lastly, none of the four DH focused journals included instructions or guidelines on the preregistration of research and analysis plans, or guidelines and policies relating to the submission and publication of replicated research.

3.2.2 Results: Open Scholarship practices in published DH papers

This section will present a series of stacked bar charts showing the results of the survey conducted on 526 papers published in the year of 2021 on the nine journals previously described. Figures 1 through 4 refer to the results of all types of papers whereas figures 5 and 6 focus specifically on papers that contain quantitative or computational research. This distinction was quantified via the "Computational/Quantitative" field in the Papers database and to determine if a research project was computational and/or quantitative or not, I visually scanned each paper looking for indicatives of the display of quantitative analysis results, i.e., data visualizations, and confirmed it by checking the methodology section. The results for "Book Review" will mostly

be ignored as they don't add much to this thesis' research project and questions. Book Review, as mentioned in the Methodology section, was established for reference only and as a potential way of explaining the results for the number of computational and quantitative papers in the dataset. Whenever a book review was detected, the "Computational/Quantitative" field for that paper was labeled as N/A. Any published introduction for a journal issue was also labeled as N/A.

Papers from all journals

	Yes	No	Total
Open Access	208	318	526
	40%	60%	100%

	Yes	No	Total
Book Review	83	443	526
	16%	84%	100%

	Yes	No	N/A	Total
Computational/Quantitative	110	278	138	526
	21%	53%	26%	100%

	Yes	No	N/A	Total
Open Data	86	36	404	526
	16%	7%	77%	100%

	Yes	No	N/A	Total
Open Code	42	68	416	526
	8%	13%	79%	100%

	Yes	No	N/A	Total
Open Software	85	31	410	526
	16%	6%	78%	100%

Table 5. *Results of the open science practices survey for papers from all journals*

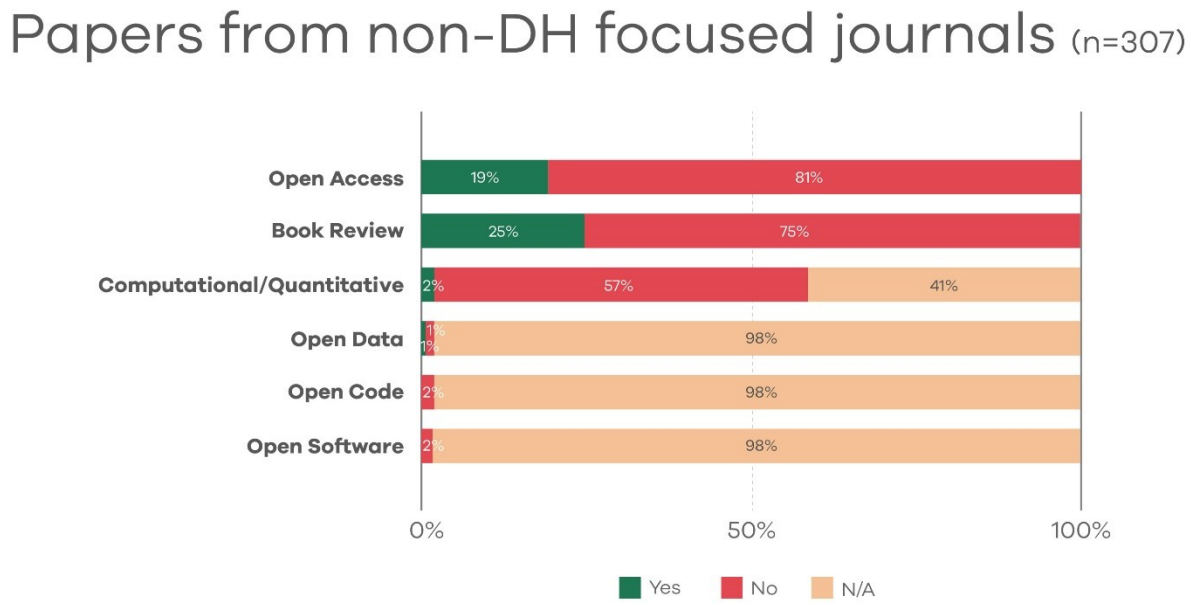


Figure 1. *Results displayed as stacked bar graphs for all papers published in 2021.*

Out of the 526 papers surveyed, 208 were published as open access, that is, 40% of the total. The figures for open data, code and software are very low when viewed in this context, but that is a consequence of the low rate of computational and quantitative papers which are 110 papers, or 21% of the total. The figures that are more representative of the state of open science practices in the Papers database will be presented in figures 5 and 6.

Papers from non-DH focused journals

	Yes	No	Total
Open Access	59	248	307
	19%	81%	100%

	Yes	No	Total
Book Review	76	231	307
	25%	75%	100%

	Yes	No	N/A	Total
Computational/Quantitative	6	174	127	307
	2%	57%	41%	100%

	Yes	No	N/A	Total
Open Data	2	4	301	307
	1%	1%	98%	100%

	Yes	No	N/A	Total
Open Code	0	6	301	307
	0%	2%	98%	100%

	Yes	No	N/A	Total
Open Software	1	5	301	307
	0%	2%	98%	100%

***Table 6**. Results of the open science practices survey for papers from non-Digital Humanities focused journals.*

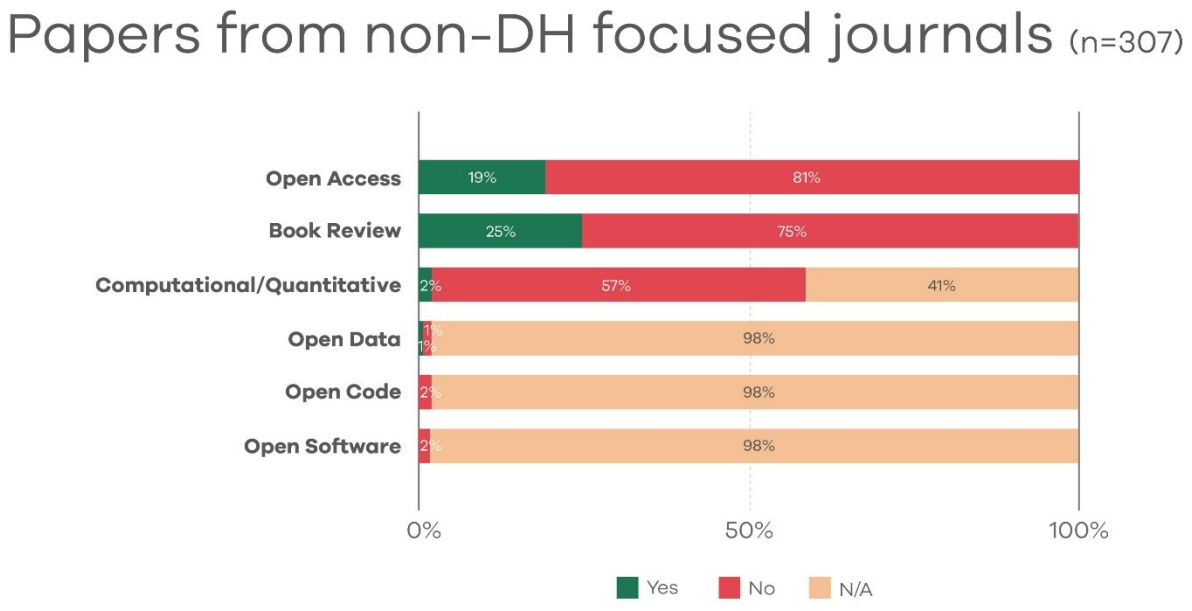


***Figure 2**. Results displayed as stacked bar graphs for papers published in 2021 in non-Digital Humanities focused journals.*

The next graph, shown in figure 2, represents the results for the papers from the five journals that are not focused on Digital Humanities. The percentage of open access papers decreases in this context to 19% of the 307 papers. Only six papers were from research that utilized computational and/or computational methods, a very low 2% of the total. The lack of quantitative research published on the literary criticism-focused journals correlates to the 98% of surveyed papers being labeled N/A on Open Data, Code and Software. From these six papers, two of them shared the data and one used open software, while none of them shared the code used to generate the results.

Papers from DH focused journals				
	Yes	No	Total	
Open Access	149	70	219	
	68%	32%	100%	
	Yes	No	Total	
Book Review	7	212	219	
	3%	97%	100%	
	Yes	No	N/A	Total
Computational/Quantitative	104	104	11	219
	47%	47%	5%	100%
	Yes	No	N/A	Total
Open Data	84	32	103	219
	38%	15%	47%	100%
	Yes	No	N/A	Total
Open Code	42	62	115	219
	19%	28%	53%	100%
	Yes	No	N/A	Total
Open Softwar	84	26	109	219
	38%	12%	50%	100%

Table 7*. Results of the open science practices survey for papers from Digital Humanities focused journals.*

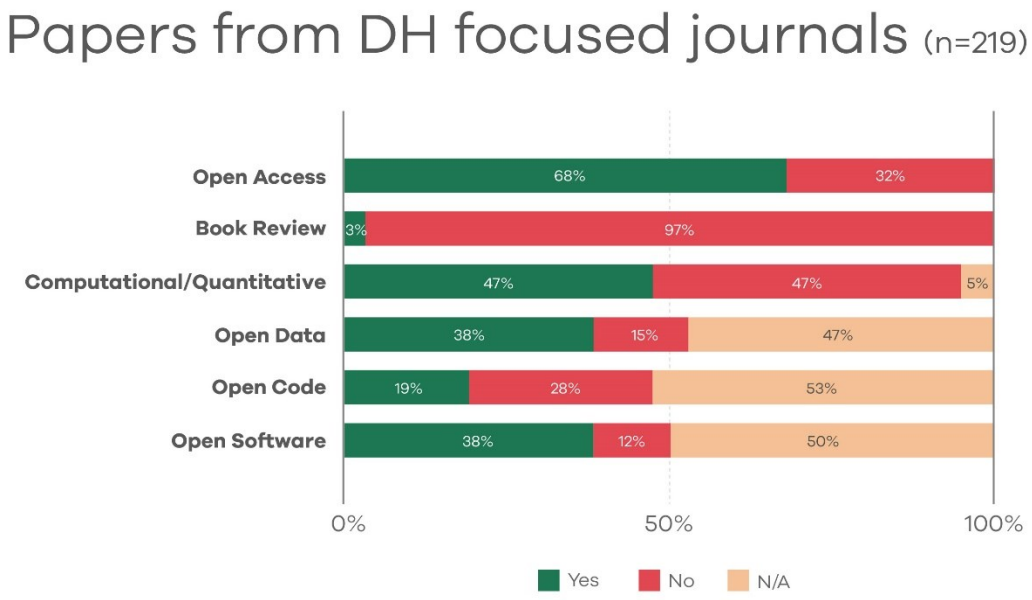


Figure 3. *Results displayed as stacked bar graphs for papers published in 2021 in Digital Humanities focused journals.*

Compared to the literary criticism-focused journals, 68% of the 219 papers from the Digital Humanities-focused journals were published as open access (see Figure 3), that is 49% more than the literary-criticism's 19%. The presence of two journals with fully open access policies (JCA and DHQ) partially explains this overall higher number of open access papers. Almost half of the papers published in the four DH journals were computational and/or quantitative research, or, more specifically, 47% of the total. The Open Data, Code and Software numbers are affected by the presence of non-computational papers, as observed with roughly half of them being labeled as N/A. The more accurate numbers will be described below, in figure 5, when the results of only computational and quantitative papers are analyzed.

DH papers broken down by journal

	Open Access		Book Review		Comp./Quant.			Open Data			Open Code			Open Software			
	Yes	No	Yes	No	Yes	No	N/A	Yes	No	N/A	Yes	No	N/A	Yes	No	N/A	Total
Digital Humanities Quarterly	68	0	7	61	25	33	10	17	10	41	9	13	46	19	7	42	68
Journal of Cultural Analytics	29	0	0	29	20	8	1	26	2	1	18	2	9	20	0	9	29
Digital Scholarship in the Humanities	45	65	0	110	57	53	0	40	19	51	12	46	52	41	19	50	110
International Journal of Digital Humanities	7	5	0	12	2	10	0	1	1	10	3	1	8	4	0	8	12

	Open Access		Book Review		Comp./Quant.			Open Data			Open Code			Open Software			
	Yes	No	Yes	No	Yes	No	N/A	Yes	No	N/A	Yes	No	N/A	Yes	No	N/A	Total
Digital Humanities Quarterly	100%	0%	10%	90%	37%	49%	15%	25%	15%	60%	13%	19%	68%	28%	10%	62%	100%
Journal of Cultural Analytics	100%	0%	0%	100%	69%	28%	3%	90%	7%	3%	62%	7%	31%	69%	0%	31%	100%
Digital Scholarship in the Humanities	41%	59%	0%	100%	52%	48%	0%	36%	17%	46%	11%	42%	47%	37%	17%	45%	100%
International Journal of Digital Humanities	58%	42%	0%	100%	17%	83%	0%	8%	8%	83%	25%	8%	67%	33%	0%	67%	100%

Table 8. *Results of the open science practices survey for papers from DH focused journals broken down by journal*

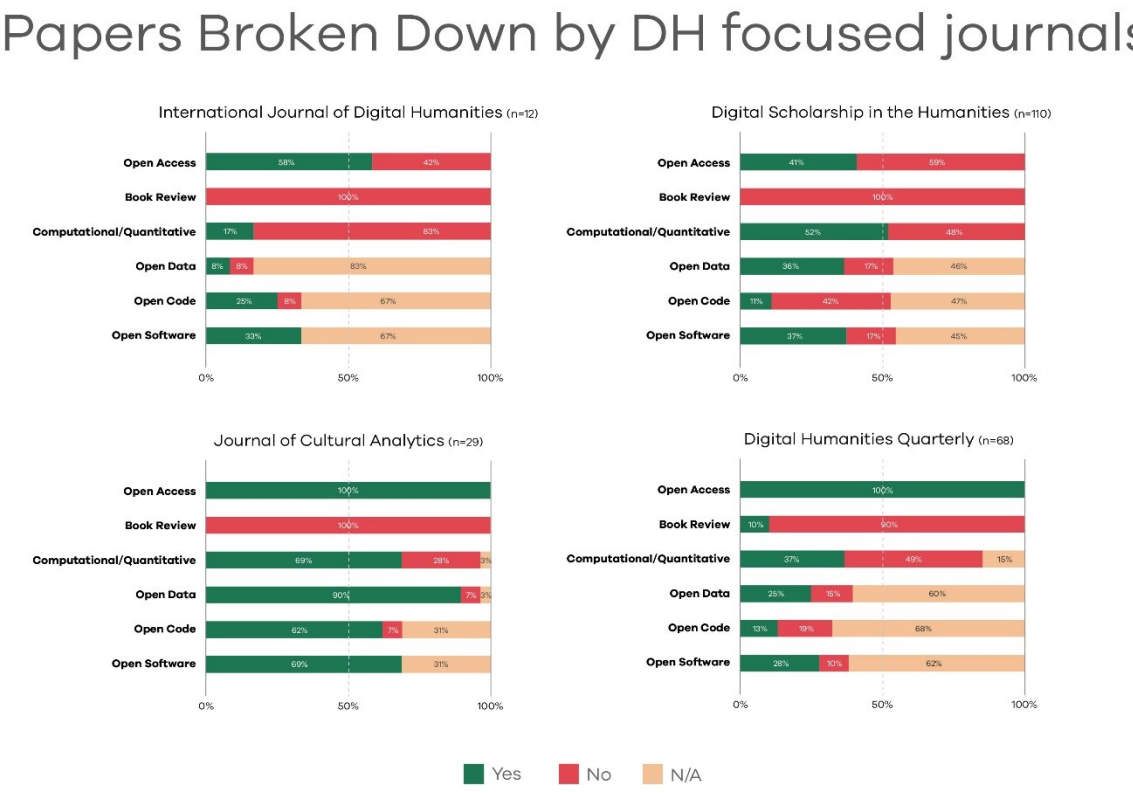


Figure 4. *Results displayed as stacked bar graphs for papers published in 2021 in Digital Humanities focused journals and broken down by journal.*

Figure 4 above breaks down the results of the 219 papers published in DH-focused journals. The journal with the most published papers was the Digital Scholarship in the Humanities, with a total of 110 papers (almost half of the total of 219). Of these, 41% were published as open access, a number that is largely explained by the publication of two supplemental issues in the surveyed year of 2021. These issues were fully published as open access, whereas the regular four issues published in that year had papers published using the standard DSH policies described previously. A total of 41 papers were published in these two supplemental issues as open access (37% of the total of 110). It follows that the vast majority of the remaining 71 papers published in the four regular issues were not published as open access.

By visiting the DSH website, it can be observed that only three supplemental issues were published since 1986, the year it was founded.

The Journal of Cultural Analytics is the journal that published, proportionally, the highest number of computational and quantitative research: 69% of its total of 29 papers. On the other end is the International Journal of Digital Humanities, with only 17%. The IJDH had a small number of total papers published in 2021 though, only 12. Digital Scholarship in the Humanities had an almost even split between computational and non-computational published research, 52% and 48% respectively. And Digital Humanities Quarterly was the only DH-focused journal that published book reviews: seven out of its 68 published papers, or 10% of the total. When broken down by journal, it is possible to observe which journals exhibited higher and lower levels of open science practices. These numbers will be analyzed below as figure 6 breaks down the results of only published quantitative research and are better suited for this analysis.

Quantitative papers from all journals

	Yes	No	Total
Open Access	59	51	110
	54%	46%	100%

	Yes	No	Total
Book Review	0	110	110
	0%	100%	100%

	Yes	No	N/A	Total
Computational/Quantitative	110	0	0	110
	100%	0%	0%	100%

	Yes	No	N/A	Total
Open Data	72	36	2	110
	65%	33%	2%	100%

	Yes	No	N/A	Total
Open Code	40	67	3	110
	36%	61%	3%	100%

	Yes	No	N/A	Total
Open Software	80	30	0	110
	73%	27%	0%	100%

***Table 9**. Results of the open science practices survey for quantitative research papers from all journals*

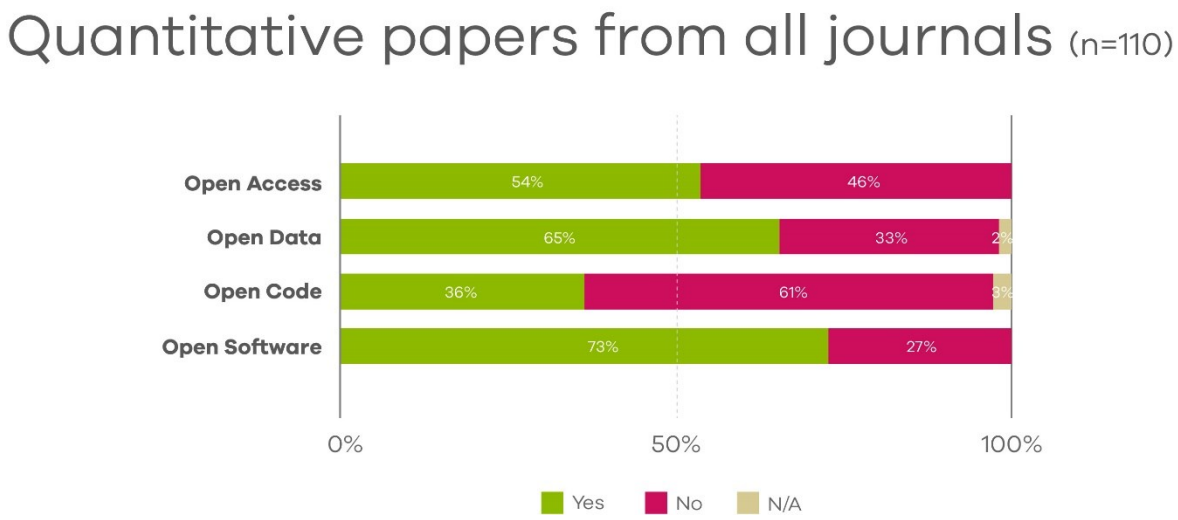


Figure 5. *Results displayed as stacked bar graphs for quantitative papers published in 2021 and published in all surveyed journals.*

Figure 5 shows the overall open science practices for published computational and/or quantitative research published in all the surveyed journals. These numbers are broken down by journal in the next image. There was an almost even split between open access and non-open access published research, 54% and 46% respectively. The results show that researchers tend to publish the data used more often than the code used: 65% of the papers were open data while only 36% were open code. The majority of the published research relied on the use of open software and, although possible explanations for the use of non-open software were not quantified, the use of specialized software in some projects, such as motion capture or audio and video editing tools partially explains the 27% of the projects that were not labeled as open software. This will be further examined in the Discussion section.

Quantitative papers broken down by journal

	Open Access		Book Review		Comp./Quant.			Open Data			Open Code			Open Software			
	Yes	No	Yes	No	Yes	No	N/A	Yes	No	N/A	Yes	No	N/A	Yes	No	N/A	Total
Digital Humanities Quarterly	25	0	0	25	25	0	0	14	10	1	9	13	3	18	7	0	25
Journal of Cultural Analytics	20	0	0	20	20	0	0	18	2	0	18	2	0	20	0	0	20
Digital Scholarship in the Humanities	14	43	0	57	57	0	0	37	19	1	12	45	0	39	18	0	57
International Journal of Digital Humanities	0	2	0	2	2	0	0	1	1	0	1	1	0	2	0	0	2
MLA ALH CI	0	6	0	6	6	0	0	2	4	0	0	6	0	1	5	0	6

	Open Access		Book Review		Comp./Quant.			Open Data			Open Code			Open Software			
	Yes	No	Yes	No	Yes	No	N/A	Yes	No	N/A	Yes	No	N/A	Yes	No	N/A	Total
Digital Humanities Quarterly	100%	0%	0%	100%	100%	0%	0%	56%	40%	4%	36%	52%	12%	72%	28%	0%	100%
Journal of Cultural Analytics	100%	0%	0%	100%	100%	0%	0%	90%	10%	0%	90%	10%	0%	100%	0%	0%	100%
Digital Scholarship in the Humanities	25%	75%	0%	100%	100%	0%	0%	65%	33%	2%	21%	79%	0%	68%	32%	0%	100%
International Journal of Digital Humanities	0%	100%	0%	100%	100%	0%	0%	50%	50%	0%	50%	50%	0%	100%	0%	0%	100%
MLA ALH CI	0%	100%	0%	100%	100%	0%	0%	33%	67%	0%	0%	100%	0%	17%	83%	0%	100%

Table 10. *Results of the open science practices survey for quantitative research papers from all journals broken down by journal*

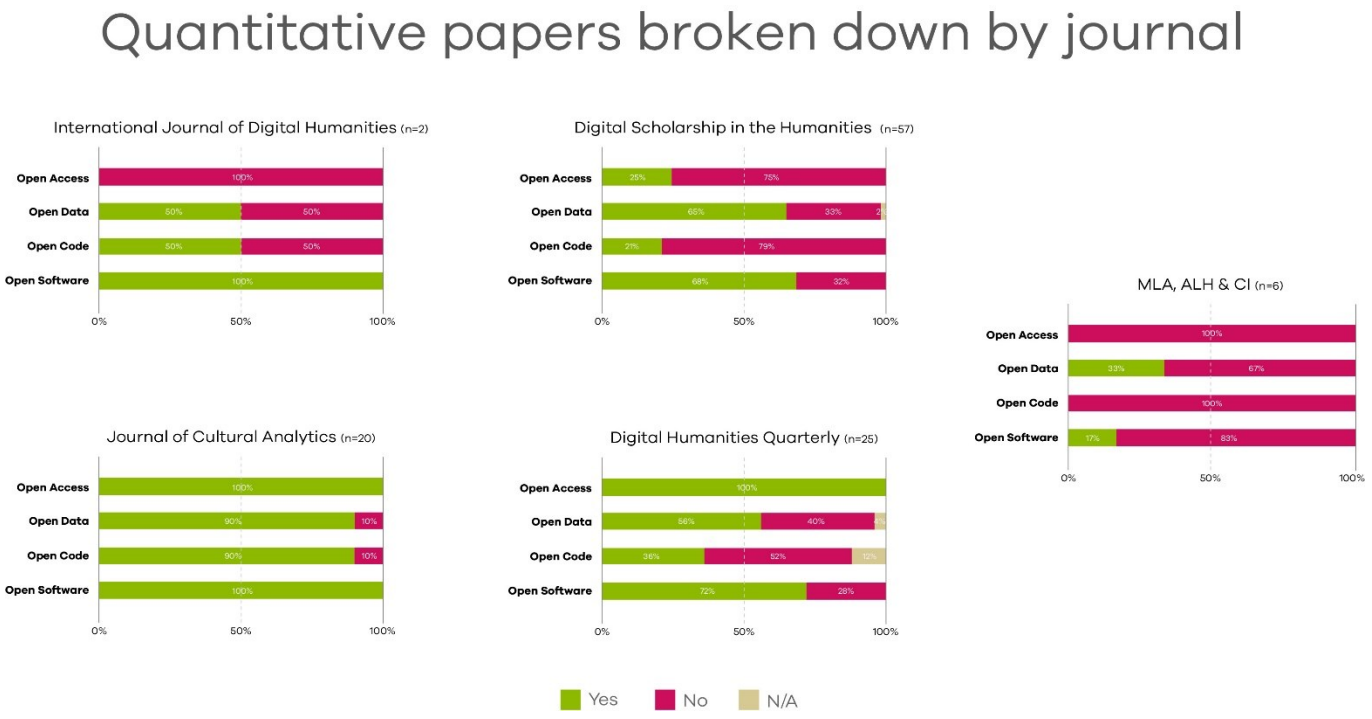


Figure 6. *Results displayed as stacked bar graphs for quantitative papers published in 2021 and broken down by journal.*

The last figure breaks down the results of all computational and quantitative papers by journal. The Journal of Cultural Analytics and Digital Humanities Quarterly both have 100% open access papers (20 and 25 papers respectively), which is explained by their fully open access policies as previously mentioned while the International Journal of Digital Humanities had its two published quantitative papers not published as open access and from the 57 quantitative papers published in the Digital Scholarship in the Humanities, 25% of them were published as

open access. There were six quantitative papers published in the non-DH focused journals and none of them were published as open access. For a review of the open access policies of each journal, see Table 2.

The JCA had the highest numbers of open data and open code among its published papers, both at 90% of the total of 20 quantitative and/or computational papers. Three papers were part of the remaining 10% and, of these three, one mentioned a DOI that contained both the data and code used but wasn't available anymore. Section 4.3 of this thesis will address the issue of broken URLs. Regarding the remaining two papers, one described the corpora used, but did not provide access to it and the other mentioned the use of two Natural Language Processing tools but did not provide the code that generated the published results. The IJDH only had two published quantitative papers: one provided access to its data and code while the other did not. In both the DSH and DHQ, the numbers for open data were higher than the numbers for open code, 65% and 56% respectively against 21% and 36%. Lastly, of the six quantitative and/or computational papers published in the non-DH focused journals, two shared the data used and none of them shared the code. The last open science practice that was measured, open software, resulted in higher numbers than the other three measurements. Both the IJDH and JCA had a 100% use of open software in their published quantitative and/or computational papers, the DSH and DHQ shared similar results, at 68% and 72% respectively of the published papers using open software and, lastly, out of the six papers published in the non-DH focused journals, one of them used open software. A possible explanation for these numbers, although not quantified during the research process of this thesis, is the widespread use of open-source Natural Language Processing libraries in Python or R. It was also observed that the use of specialized software, such as Adobe Premiere for video editing, Kaleidoscope Pro for sound analysis and

Photosounder for converting images into sound, contribute to the use of non-open software. Digital Humanities is a large field and deals with textual, sonic, audiovisual and imagetic data, among others. Because of the myriad of software available for handling that vast amount of data and data types it is a challenge to ensure that all software is available to all researchers.

3.3. Discussion

To begin this discussion, the dataset containing the list of journals and papers as well as the annotations and data analysis are available on the McGill University Dataverse[7] repository (Tiefenbach Keller, 2023). This came with its own challenges that reflected issues faced by the authors of the surveyed papers when they attempted to share their own datasets or described why data couldn't be made available. How to share all the papers that were reviewed and formed the basis of the database created for this survey? Many of the papers are restricted by copyright laws so sharing them could lead to legal issues. Many of these papers are also not open access so if a researcher would like to confirm if the findings presented here are accurate, they would need to have access to each non-open-access paper via a library that grants access to them or by purchasing them. This second option is financially prohibitive to most people who don't have academic credentials and access to a library system. In the end my best option was to simply provide the list of all papers and instructions on how to recreate the database.

[7] To access the two survey datasets, see https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/KRPCBL

This section will now focus on describing how the data collected in the survey contributes to answering research questions 1 through 3.

(RQ 1a) *Do Digital Humanities focused journals encourage scientific openness and transparency?*

Four Digital Humanities focused journals were reviewed. From these four, one of them, Journal of Cultural Analytics, had stricter guidelines on how authors should make their data and code available to others; another, the Digital Humanities Quarterly, asked authors to include their data and code as supplementary material when submitting their articles, but the journal did not make it clear how they would enforce authors to submit their data and code; the third journal, International Journal of Digital Humanities, requires authors to include a data availability statement and encourages them to share and/or cite any data used. Instructions on how the code should be shared are not as clear. And, lastly, the Digital Scholarship in the Humanities does not provide any instructions on how data and/or code should be shared. The first two journals, the JCA and DHQ are fully open access and the IJDH is currently transitioning to becoming a full open access journal. The DSH journal gives authors the option to publish as open access.

Based on what was just reviewed, the answer to RQ1 cannot be a simple yes or no because there is no consistency in openness policies between the four DH-focused journals.

(RQ 1b) *Do literary criticism focused journals encourage scientific openness and transparency?*

At the time the data presented in this thesis was collected, none of the literary criticism focused journals contained any policies related to data or code sharing and one out of five

journals had an open access policy that allowed authors to publish their articles as open access while the other four journals did not have an open access policy. It should be noted that, as of August of 2022, the American Literary History updated its data and code openness policies. The journal now encourages authors to make their data and code available and requires authors to include a data availability statement in their article.

Given the very low number of articles published in these journals in 2021 that included the use of quantitative data, six papers out of a total of 307, the presence of data and code openness guidelines do not seem like a necessity at this time since there is not a real demand for it, and literary criticism articles published in the five journals reviewed in this thesis rely largely on methods that are not quantitative or computational. The recent move made by the American Literary History, though, is a move in the right direction for that journal to be ready, should that demand arise, to handle the submission of humanities research that rely on quantitative methods. In the year 2021, the ALH published two quantitative and/or computational research articles, and it remains to be quantified how many quantitative articles were published in that journal in 2022. Was there an increase that prompted the update in the journal policy? Or that update was enforced by its parent publisher, Oxford University Press? A positive answer to the second question leads to the discussion of the relevance of large academic publishing groups in ensuring that published research follows principles of openness. The Oxford University Press is the publisher of two of the journals reviewed in this thesis, the ALH and Digital Scholarship in the Humanities. Both of them give authors the option to publish their research as open access and the ALH requires authors to include a data availability statement. The OUP also has a page on its

website dedicated to its open access policies[8]. Delving more deeply into this discussion is outside the boundaries of this thesis, but an understanding of the role of the large academic publishing companies in ensuring that the scientific knowledge is grounded in open practices, is important for the comprehension of the current state of openness in science.

(RQ 2a) *Does published Digital Humanities research follow principles of scientific openness and transparency?*

The data collected indicates a correlation between journal policies on openness and the presence of papers that share data and code used as well as articles published as open access. All papers published in the two open access journals, Journal of Cultural Analytics and Digital Humanities Quarterly, were published as fully open access. The Digital Scholarship in the Humanities gives authors the option to publish their papers as open access and only 41% of the authors did so. This figure drops to 25% when we look only at the quantitative and/or computational papers published in that journal. Lastly, the International Journal of Digital Humanities, which also gives authors the option to publish as open access, had 59% of its 12 papers published as open access. It should be noted that the IJDH had a significantly lower total number of papers published in 2021.

The trend described above is also observed on the sharing of data and code used by authors. The JCA has strict guidelines on how authors should share their data and code, a policy that led to 90% of the papers including the data and code on the designated repository. The DHQ

[8] For information on Oxford University Press' open access statement and types of open access policies options for its journals, see https://academic.oup.com/pages/open-research/open-access

encourages authors to submit their data and code as supplementary material, which led to 56% of the papers sharing the data and 36% sharing the code. The DSH provides no guidelines on how authors should share their data and code, an editorial decision that led to 65% of the authors sharing the data used and 21% sharing the code. Lastly, the IJDH only had two quantitative papers published in 2021 and one of them shared the data and code used while the other did not.

Based on the results published in this thesis, stricter openness guidelines lead to more authors publishing their paper as open access and sharing the data and code used in their research. It was also observed that there was a higher tendency to share the data used rather than the code. This tendency was present in the two journals that did not have stricter guidelines on how data and code should be shared, the DHQ and the DSH.

(RQ 2b) *Does published literary criticism research follow principles of scientific openness and transparency?*

Because there were only six papers published in the literary criticism that relied on quantitative and/or computational methods, there is little data available on data and code sharing practices. Out of these six papers, none of them shared the code used, and two shared data. But, similarly to what was discussed in RQ 1b, most of the papers published in these journals rely on methods that are not quantitative, such as close reading, therefore requiring these articles to share any data or code used is unfair. If, in the future, the number of quantitative and/or computational papers published in literary criticism journals increases, then this research question can be revisited, and an answer will help understand the state of openness practices in the literary criticism space.

But one of the openness metrics is applicable to the five literary criticism journals reviewed in this thesis, open access. And, overall, based on data from the year 2021, these journals publish the majority of their papers as not open access. Out of 307 papers, 81% were not open access and this number increases to 91% if a supplemental issue containing short essays, poems, and images created during the 2020 COVID pandemic and published in the Critical Inquiry is excluded. Most of these journals are fairly old: the youngest of them, CI, was founded in 1974, while the oldest, the PMLA, was founded in 1883. Future research should investigate how older and established institutions and publishers can better adapt to the demands of openness so their publications can reach a wider audience, especially an audience outside the walls of academia.

(RQ 3) *Do journal policies that encourage open science practices correlate with an increase in data and code shared by researchers in published papers?*

Contrary to what was hypothesized, and in the context of the data collected for this thesis, papers published in 2021 in Digital Humanities focused journals, encouraging authors to share their code and data did not lead to an increase in them being shared. In the case of data sharing, the Digital Scholarship in the Humanities had a slightly higher number of papers that shared the data used in its published projects compared to the numbers observed in the Digital Humanities Quarterly. Even though the DHQ encourages authors to share data and code, 56% of authors shared their data while 65% of the authors who published on DSH did so. These numbers are reversed when we look at the sharing of code, 21% shared their code on the DSH and 36% on the DHQ. Pineau et al. (2020) observed an increase in code sharing in papers submitted to the Neural Information Processing Systems conference when a code sharing policy based on voluntary

participation was included: from less than 50% before the policy was implemented to close to 75% after it. These numbers are similar to those observed and described here on data sharing on both the DHQ and DSH but the inclusion or not of a policy that encouraged authors to share their data did not affect the quantity of authors doing so in a significant way. Pineau et al. numbers, though, are significantly higher than the numbers observed here for the sharing of code on both the DHQ and DSH journals. It is not clear, at this time, why the numbers for code sharing are so low in these two journals, and this is a topic that can be investigated further in future research. Contrastingly, a journal policy that enforces data and code sharing as a requirement for a paper to be published, such as that of the Journal of Cultural Analytics, leads to numbers that are virtually 100%, when we exclude issues such as a broken URL. The evidence collected here suggests that the only effective journal policy that brings the code and data sharing levels to a significantly high number is one that obligates authors to follow open science practices. The mere encouragement of authors to follow open science practices does not seem sufficient to result in data and code sharing levels that are different to what is observed when such encouragement is not made. For this analysis, the results for data and code sharing from the International Journal of Digital Humanities as well as from the journals that are not focused on DH were not included since the number of relevant papers was very low: two for the IJDH and six for all of the non-DH focused journals.

While collecting the data for this thesis, a few phenomena were observed and will be described and discussed below. But before delving into these discussions, it is important to note that these observations are beyond what was described in the methodology section and, therefore, are topics that could be further explored in a systematic way in the future. All

observations as well as anything deemed worth looking further into were recorded and are available for review in the "Notes" field of the Paper database.

First, only one out of 110 computational/quantitative papers reviewed was a replication of another study (Rizvi, 2021). This very low number reflects results that were encountered previously by other researchers: in a study conducted by Iqbal et al. (2016), four out of 268 biomedical papers clearly stated that they referred to a replication of another study and Wallach et al. (2018), when extending the research conducted by Iqbal et al., identified five biomedical papers out of 94 as being replications; Hardwick et al. (2020) observed that two out 156 social science papers self-identified as a replication of another study.

The second topic that will be discussed is the presence of broken URLs in the reviewed papers. Although I was not systematically searching for broken URLs, just by trying to access the data and code provided by authors so the URLs could be included in the Papers database, I came across six hyperlinks that didn't lead me to where they were supposed to. Interestingly, Rizvi (2021), author of the only published paper among all the papers reviewed for this thesis that was a replication of another study, mentioned that he could not access the supplementary material provided by the author of the original research by following the URL included in the paper. Hardwick et al. (2020) have also encountered broken URLs when studying transparency: two broken URLs when trying to access supplementary material, eight broken URLs when trying to access data and one broken URL when trying to access code.

The broken URL issue has been identified by the National Academies of Sciences, Engineering, and Medicine (2019) and described as a problem that can lead to failure to reproduce research even if the original authors did their due diligence of recording and reporting relevant information (pp. 69). The authors suggest that researchers should share their digital

artifacts in a way that they are "searchable by providing a unique global identifier for the deposited artifact, has a stated guarantee of long-term preservation, and is aligned with a standard set of data access and curation principles" (pp. 120) to reduce the risk of URLs becoming obsolete as well as preserving data quality and integrity over time. On the topic of sharing code, the authors suggest providing information about the unique identifiers for any library used and the code itself and to cite the DOI of a deposited source code instead of citing a link to a Git repository (pp. 123). To understand the magnitude of this problem, Klein et al. (2014) checked for what they called "reference rot" in approximately 1.85 million articles published from 1997 to 2012 on arXiv, Elsevier, and PubMed Central (PMC). Link rot increased significantly over time: for articles published in 2012, 13%, 22% and 14% of the total links were broken on articles published on arXiv, Elsevier, and PMC respectively. These numbers increase to 18%, 41% and 36% in 2005 and to 34%, 66% and 80% in 1997. It should be noted that articles published in 1997 had less references to digital artifacts than articles published in 2012, but their findings confirm trends identified by other researchers. Klein et al. also points out to fact that, besides the issue of broken links, the contents of an URL can also change over time, sometimes to the point that the new content is not at all what it was when that URL was cited.

Lastly, the third topic that will be discussed here are a few potential reasons for why Digital Humanities data may not be made available by authors. Note that these are just a few reasons that I identified while scanning the 110 computational/quantitative papers, so this list is not at all exhaustive and further investigation in this topic can lead to a potential taxonomy of reasons and help journal editors address them when reviewing manuscripts and publishing them. (a) Earhart et al. (2021) mention that the data they used had been deposited in the author's university repository but is embargoed until the paper has been published. (b) Have and

Enevoldsen (2021) state that the audio Media Collection used by them is legally protected and can only be accessed via on-site Dutch library computers. University faculty and students are only allowed to stream the content and cannot download it. (c) Martin (2021) recorded 100.000 .wav files of 1 minute each for their study. That is a large amount of data and if we assume that each minute of .wav requires 10Mb of hard drive space, then it adds up to around 1Tb. (d) Gittel (2021) shared the filtered data that was analyzed in the paper, but access to the original database needs to be requested to libraries located in Germany. Based on these four examples, an initial list of reasons for why DH data may not be easily made available include data embargoes; copyright and legal reasons; file size and/or large volume of files; data availability restricted to specific parts of the world (e.g., library systems in one country).

4. Replicating a DH project

For this section, I conducted a replication that aimed to serve as a case study on the topic in the context of the Digital Humanities. In the beginning of the process, I was faced with the challenge of selecting which study or studies to replicate. To attempt to measure replicability in the Digital Humanities in a similar way to what, for example, the two Reproducibility Projects (Open Science Collaboration, 2015; Errington et al., 2021a) did, a large amount of time and resources are needed. Instead, I opted to conduct a single replication project and provide an understanding of the challenges and benefits associated with conducting such a project. The selection process was straightforward: although I could have opted for a randomized approach, I chose a research project that I was interested in conducting myself. Peter Meindertsma (2019) paper on changes in lexical and emotional diversity in top charting songs in the USA was a project that I had intended to conduct myself as a research project for my thesis. His choice of methodology and the results obtained appeal to my personal and scholarly interests, so it made perfect sense to conduct a replication of Meindertsma project.

Peter Meindertsma research was published in 2019 in the Digital Humanities Quarterly journal. He studied how popular top charting Billboard songs changed over the period of 1956 to 2016 and the concept of "change" was analyzed and measured in four different ways: (1) changes in hit diversity, (2) changes in word usage, (3) changes in sentiment and (4) changes in lexical complexity. The author's database comprised of 27.108 songs, an average of 444 songs per year, and that represents all songs that charted during the analyzed period. Although Billboard has charts for specific genres (e.g., Hot Country Songs, Dance Club Songs, Hot Christian Songs, and, among many others, Hot R&B/Hip-Hop Songs), the list of songs used for his research refers to the all-genre list of Billboard Hot 100, Hot 40, Hot 10 and Hot 1, which are

tracked weekly for each year. From now on, “hot” will be referred to as “top”, since that is the word used by Meinderstma to refer to Billboard’s Hot charting songs.

Returning to the four metrics used by Meindertsma to analyze change, (1) changes in hit diversity refers to the frequency that top performing songs appear on the Billboard weekly charts. The results presented in the paper show a decline over the years in the number of top performing songs, a result that, according to the author, indicates a loss of market competition and diversity. In the second (2) analysis conducted by Meindertsma, he tracked changes in the popularity of specific word usage. Although the results reflect interesting trends in US popular music and culture (e.g., the sudden growth in the use of profanities during the 90s; the increase of the use of the word “money” starting from the 80s; the constant decrease of the use of the word “darlin’” over the years; and the peak use of the word “disco” during the 70s, the period when Disco music was very popular in the US. All these trends can be visualized in the graphs provided by the author), this metric seems more appropriate for an exploratory analysis of word usage and a starting point for further investigation on how word usage corresponds to cultural trends and Meindertsma concludes by saying that this metric is not ideal for understanding “changes in the homogeneity of popular lyrics” (p. 5).

The next two metrics refer to what he used to understand these changes. The first of them (3) refers to change in sentiment in popular songs lyrics. To measure sentiment, Meinderstma relied on the Affective Norms for English Words (ANEW) list of words (Bradley and Lang, 1999). The list contains 1034 words along with valence, arousal and dominance values attributed to each word. Meinderstma added up the valence values for all ANEW words that were present in popular song lyrics, calculated the average valence score for each song and calculated the average valence score per year. The result shows a decrease in valence over the years. The author

also calculated the standard deviation of the valence scores and encountered short-term fluctuation and a tendency to decrease over time. The second metric (4) are changes in lexical complexity. Meindertsma results show an increase in the number of words used in popular music, when averaged by year, as well as an increase in the average length of popular songs. By dividing these two results, he also found an increase in the average number of words per second per year in popular music. Lastly, he calculated the average Type Token Ratio (TTR) of popular songs lyrics. To do so, he calculated the TTR of randomly sampled words from popular songs with a sample size of 75, repeated the sampling process 50 times and averaged the results. In the end, it is possible to observe an increase in TTR noticeably from the late80s/early 90s onwards. A trend that, the author notes, coincides with a growth in the presence of hip hop songs on the Billboard charts starting in that same time period. Meinderstma also calculate the standard deviation of the sampled TTR and encountered fluctuating sinewave pattern, with a global tendency to decrease over time, a trend that, he concluded, indicates a tendency of popular songs lyrics towards homogeneity.

4.1. Methodology

In order to attempt to replicate Meinderstma (2019) research and its results, the first step involved locating and retrieving the data and code used in the original project. To do so, I visited the Digital Humanities Quarterly website[9] where the paper was published as well as the author's personal website[10] but was unable to locate neither the data nor the code. Without direct access to them, I could either email the author directly or rely on the directions given on the research

[9] URL used to access Peter Meinderstma original research paper: https://www.digitalhumanities.org/dhq/vol/13/4/000440/000440.html
[10] URL to access Peter Meinderstma personal website: https://www.petermeindertsma.com/

paper to attempt to recreate the dataset and code myself. Since attempting to obtain the data and code from a single study would provide less material for discussion, at least when compared to, for example, Stodden et al. (2018) who emailed 180 authors attempting to retrieve the data used by them in their respective published papers and quantified the type of responses obtained, I opted to, instead, recreate Meinderstma dataset from scratch and then describe the process.

4.1.1 Data collection

In his paper, Meindertsma mentions that he retrieved his list of top charting Billboard songs from a website called "Bullfrogs Pond" and, according to his Works Cited section, that website was accessed in May 2013. However, when I tried to access that website, it was not available anymore[11]. An alternative approach was, therefore, needed. Wikipedia contains lists of all Billboard's Year End Hot 100 singles, and I scraped the lists from the years 1959 to 2018, ending with a list of 6000 songs with attributes that included song title, artist, chart position and year. This list differed from Meinderstma's list in one significant way: each year contains only the 100 best performing singles while Meinderstma's included the top performing songs from each week. This limitation restricted me from replicating the author's first analyzed change in popular songs: "changes in hit diversity". To retrieve the lyrics associated with each song in the list of 6000 songs, I used the Genius[12] API and a Python script that iterated through each entry in the list. The script couldn't retrieve all the lyrics and roughly 500 songs returned no results from the API. These were, then, individually searched for on Google and added manually to the

[11] According to a post on an online forum titled "American Top 40 Fun & Games Site", the Bullfrog Pond went through legal issues and had to be taken down. The forum thread can be accessed via the following URL: https://at40fg.proboards.com/thread/4623/websites-peak-positions-year-hot

[12] For more information about Genius, see https://genius.com/Genius-about-genius-annotated . They claim to be "the world's biggest music encyclopedia".

database. After the full lyrics database was compiled, data cleanup was required. Genius lyrics include song structure information in them including “intro”, “verse”, “pre-chorus” and “chorus” and “outro”. These words were all deleted from the lyrics. Another unnecessary data included in the lyrics were “you may also like...” followed by the name of an artist like that of the lyrics and “get tickets for as low as...” promoting concerts. These phrases were deleted. Note that the name of the artist that came after “you may also like” was not deleted as well as the name of artists that were included along with song structure information (e.g., “Verse 1: Drake”). The Genius API retrieved several songs in a language other than the lyrics original song (it retrieved a translated version of that song), and in these cases I manually retrieved the correct lyrics in English and added it to the database.

Meinderstma database included the duration of each song, and he retrieved that information from the “Bullfrogs Pond” website. As already mentioned, that website is not available anymore and I retrieved that information using the Spotify API. With a Python script, it was possible to retrieve not only the duration of each song, but also their genres and song features[13] that are used by Spotify to better classify songs and fine tune their recommender algorithm. These attributes include “energy”, “valence”, “instrumentalness”, “danceability” among others. It should be noted that Spotify does not describe in detail how these values are calculated. Tables 10 and 11 below show how the lyrics and Spotify datasets were structured. Only the first 10 entries of the total of 6000 are shown.

[13] For more information about the song features calculated and provided by Spotify, see https://developer.spotify.com/documentation/web-api/reference/get-several-audio-features

Chart	Year	Artist	Song	Lyrics	duration_sec
1	1959	Johnny Horton	The Battle of New Orleans	The Battle of New Orleans In 1814	151.507
2	1959	Bobby Darin	Mack the Knife	Mack the Knife Oh the shark babe l	184.333
3	1959	Lloyd Price	Personality	Personality Over and over I tried to	155.213
4	1959	Frankie Avalon	Venus	Venus Hey Venus! Oh Venus! Venu	144.483
5	1959	Paul Anka	Lonely Boy	Lonely Boy I'm just a lonely boy Lo	151.107
6	1959	Bobby Darin	Dream Lover	Dream Lover Every night I hope ar	150.706
7	1959	The Browns	The Three Bells	The Three Bells Les Trois Cloches T	169.573
8	1959	The Fleetwoods	Come Softly to Me	Come Softly to Me Doo dooby doo	144.933
9	1959	Wilbert Harrison	Kansas City	Kansas City I'm going to Kansas Cit	150.333
10	1959	The Fleetwoods	Mr. Blue	Mr Blue Our guardian star lost all l	145.133

***Table 11**. First 10 entries of the Genius lyrics dataset*

Chart	Year	Song	Artist	Genre 1	Genre 2	Genre 3	Spotify URL	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_	duration_	time_signature
1	1959	The Battle of New	Johnny Horton	cowboy western			https://open.	0.704	0.743	9	-12.069	1	0.129	0.805	0	0.205	0.92	177.362	151507	151.507	4
2	1959	Mack the Knife	Bobby Darin	adult standards	easy listening	lounge	https://open.	0.549	0.529	3	-12.291	0	0.108	0.76	0	0.206	0.464	82.755	184333	184.333	4
3	1959	Personality	Lloyd Price	adult standards	doo-wop	louisiana blues	https://open.	0.604	0.313	5	-13.495	1	0.0375	0.801	0	0.342	0.86	129.182	155213	155.213	4
4	1959	Venus	Frankie Avalon	adult standards	bubblegum pop	doo-wop	https://open.	0.576	0.347	10	-18.755	1	0.0317	0.852	0.599	0.124	0.85	114.883	144483	144.483	4
5	1959	Lonely Boy	Paul Anka	adult standards	canadian pop	easy listening	https://open.	0.623	0.605	0	-9.923	1	0.0306	0.352	0	0.101	0.804	106.163	151107	151.107	3
6	1959	Dream Lover	Bobby Darin	adult standards	easy listening	lounge	https://open.	0.526	0.774	5	-1.692	1	0.0459	0.715	0	0.437	0.715	131.715	150706	150.706	4
7	1959	The Three Bells	The Browns	deep adult standards	nashville sound	vocal harmony group	https://open.	0.47	0.159	7	-14.795	1	0.0403	0.893	0	0.124	0.418	106.488	169573	169.573	4
8	1959	Come Softly to Me	The Fleetwoods	adult standards	doo-wop	rock-and-roll	https://open.	0.601	0.095	1	-18.478	1	0.0374	0.93	0.000285	0.116	0.672	122.975	144933	144.933	4
9	1959	Kansas City	Wilbert Harrison	rhythm and blues	rock-and-roll		https://open.	0.727	0.665	1	-6.453	1	0.0352	0.653	0.233	0.349	0.912	114.979	150333	150.333	4
10	1959	Mr. Blue	The Fleetwoods	adult standards	doo-wop	rock-and-roll	https://open.	0.374	0.0889	4	-17.475	1	0.0356	0.931	6.88E-06	0.111	0.211	96.677	145133	145.133	4

***Table 12**. First 10 entries of the Spotify dataset*

Meinderstma mentions in his paper that instrumental songs were excluded from his analysis. I did not permanently exclude instrumental songs from the replication database due to their potential source for additional insights, e.g., answering future research questions such as which genres and time periods contain the most instrumental songs in popular top 100 songs? The presence of instrumental songs does not affect Type Token Ratio since the sampling method used in that calculation excludes songs with word count lower than the sample size (in the lyrics database, instrumental songs have only one word, “Instrumental”). If the presence of instrumental songs could potentially affect the results of an analysis, such as calculating the average ANEW scores in which an instrumental song would add a zero to the calculation, then these songs were removed from that calculation.

4.1.2 Choice of original results to replicate.

Changes in Hit Diversity (1)	Changes in Word Usage (2)	Changes in Sentiment (3)	Changes in Lexical Complexity (4)
Total top 100, 40, 10 and number 1 hits in the Billboard Hot 100 per year	Relative frequencies of song lyrics that contain selected terms on the Billboard Hot 100	Average valence scores of songs on the Billboard Hot 100 per year	Average number of words per song on the Billboard Hot 100 per year
		Standard deviation of average valence of Billboard Hot 100 songs per year	Average length in seconds of songs on the Billboard Hot 100 per year
			Average words per second of songs on the Billboard Hot 100 per year
			Average Type Token ratio of songs on the Billboard Hot 100 per year
			Standard deviation of sampled Type Token ratio

***Table 13**. Original Meinderstma's measurements*

Five of Meinderstma's published results were chosen for replication, namely the average valence scores of songs on Billboard Hot 100 per year (the green item on Table 13), average number of words per song per year, average song length per year, average words per second per song per year and, average type token ration (TTR) per song per year (the yellow items on Table 13). The same methods used by Meinderstma were implemented for the replication, following the author's descriptions on the original paper.

Changes in Hit Diversity and Changes in Word Usage were chosen not to be replicated because, in the case of the first, the reconstructed dataset didn't allow for that replication to be accomplished – it requires weekly Billboard Top 100, Top 40, Top 10 and Top 1 data per year while the reconstructed dataset only lists year-end Top 100 data. And in the second case,

Meinderstma didn't draw any conclusions from the examination of how word usage changed over time.

4.1.3 Direct replication

With all lyrics added to a CSV file (see Table 10), a Python script would then iterate through the 6000 lyrics, remove all upper cases and tokenize the words. It would them count and store the counts as well as the words, unique words, words excluding stop words and unique words excluding stop words. The TTR for each song was also calculated (this would not be included in the results, though. The sampling method used by Mederstma was also implemented and will be described shortly). By using the song lengths scrapped from Spotify, the words per second, unique words per second and TTR per second were calculated. All the results were added to a CSV file (see Table 11). The tokenization and unique words identification were implemented using both the NLTK Python library as well as vanilla Python. The compared results showed very little variation and the NLTK approach was favored due to its slightly increased time efficiency and streamlined code. All subsequent methods that rely on these word counts will retrieve the NLTK word counts.

Year	Chart	Words	Unique words	Words - no stopwords	Unique words - no stopwords	Word count	Unique Word Count	Type Token Ratio	Word count - no Stopwords	Unique Word Count - no Stopwords	Type Token Ratio - no Stopwords	Words per second	Unique words per second	TTR per second	Words per second - no Stopwords	Unique words per second - no Stopwords	TTR per second - no Stopwords
1959	1	the battle c	Counter({'the': 35	battle orleans	Counter({'fire	514	141	27.4319	184	80	43.47826087	3.39258	0.9306501	0.18106	1.214465338	0.528028408	0.286971961
1959	2	mack the k	Counter({'the': 10	mack knife sh	Counter({"'s":	248	133	53.629	131	79	60.30534351	1.34539	0.7215203	0.29094	0.710670363	0.428572204	0.327154354
1959	3	personality	Counter({'over': 6	personality pr	Counter({'pers	291	49	16.8385	78	21	26.92307692	1.87484	0.3156952	0.10849	0.502535226	0.135297945	0.173458904
1959	4	venus hey \	Counter({'venus':	venus hey ver	Counter({'ven	175	71	40.5714	66	32	48.48484848	1.21122	0.4914073	0.2808	0.456801146	0.221479344	0.335574763
1959	5	lonely boy	Counter({'i': 16, 't	lonely boy m	Counter({'lone	168	66	39.2857	58	30	51.72413793	1.11179	0.4367766	0.25999	0.383833972	0.198534813	0.342301402
1959	6	dream love	Counter({'yeah': 4	dream lover r	Counter({'drea	304	79	25.9868	94	32	34.04255319	2.01717	0.5241994	0.17243	0.623730973	0.212333948	0.225887179
1959	7	the three b	Counter({'the': 17	bells trois clo	Counter({'jimr	232	106	45.6897	111	64	57.65765766	1.36814	0.6250995	0.26944	0.654585341	0.377418575	0.340016734
1959	8	come softh	Counter({'dum': 9	softly doo do	Counter({'dum	314	51	16.242	222	25	11.26126126	2.16652	0.3518867	0.11207	1.531742253	0.172493497	0.077699773
1959	9	kansas city	Counter({'i': 21, "'	kansas city m	Counter({"'m"	221	71	32.1267	85	31	36.47058824	1.47007	0.4722849	0.2137	0.565411453	0.206208883	0.242598686
1959	10	mr blue ou	Counter({'blue': 1	mr blue guard	Counter({'blue	206	84	40.7767	99	42	42.42424242	1.41939	0.5787795	0.28096	0.68213294	0.289389732	0.292312861

Table 14. *First 10 entries of the word count dataset*

Another TTR calculation method, besides the one described above, was implemented using Python. Randomly selected words from a predetermined sample size would be used to

calculate a TTR value for. The sampling method would be repeated and also predetermined number of times and all TTR values would then be averaged and assigned to each song. The same sampling size of 75 and sampling repetition of 50 chosen by Meinderstma were used. Any song with less words than the sampling size was removed from the analysis, which essentially removed all instrumental songs and a few songs with very short lyrics.

The last result that was replicated, the average ANEW scores per year per song, was calculated by adding up all the scores for each ANEW word's attributes and dividing each sum by the total number of ANEW words identified on a song. The ANEW word dictionary used was the same used by Meinderstma: the original Bradley and Lang's (1999) list of 1034 words.

4.1.4 Conceptual replication

To explore potential new ways to further validate the original claims published by Meinderstma, additional ways of measuring "Changes in Lexical Complexity" were included in this replication. The first was counting the hapaxes (words that appear only once in a text) present in each song and averaging the sums by year and the second was calculating Dale-Chall scores for each song and averaging the scores by year. Dale-Chall is a readability score that takes into account the presence of words taken from a list of 3000 words considered easy to understand by a fourth-grade student. The words outside of the list that were present on each song were stored in a CSV file and the most repeated words per year were analyzed in an attempt to understand how language evolved in popular music lyrics. Lastly, k-means clustering was calculated for the 6000 lyrics, clustering together lyrics based on the total number of words and total number of unique words. Analyzing the songs that were clustered together due to having a

large quantity of words could provide insights into what these songs have in common, especially if they share any genre commonalities.

To further explore "Changes in Sentiment", while Meinderstma's study used only the 'valence' scores associated with each ANEW word, for this replication, 'Arousal' (an attribute that associates each word with a value that ranges from calm to excited) and "Dominance" (associates each word with a value that ranges from a large, dominating figure to a small, less dominating figure) were included. Additionally, the discrete emotion scores that were appended to the 1034 words were also calculated to validate and extend the original findings (Stevenson et al., 2007). These include individual scores for happiness, anger, sadness, fear and disgust.

Lastly, all data and code used in this replication is available on the McGill University Dataverse[14] repository (Tiefenbach Keller, 2023). In addition to the full code being made available, I added all relevant code snippets to the end of this document, in Appendix A. This should help in case, in the future, the URL shared in this thesis becomes broken, which is a common issue as has been discussed previously. Code snippets are also an easy way to share any relevant pieces of code necessary for rerunning a calculation, and since computer code is essentially text, they can easily be included in academic and scientific documents.

[14] To access all data and code used for the replication of Meinderstma's research, see https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/KRPCBL

4.2. Results

4.2.1 Results: Direct replication

The results below refer to the replication of the results of Meinderstma's sub-sections (3) and (4), "changes in sentiment" and "changes in lexical complexity", respectively. Meindersta's paper includes graphs with no values or accompanying table, therefore exact values cannot be retrieved, and comparisons need to be made relying on visual cues and the values displayed on the y axis of the graphs. Red and green dashed lines were added to the original and replicated result graphs to aid in the visual analysis. These dashed lines are especially useful for visualizing and contrasting any difference in magnitude of the results. For example, Figure 7 shows a dashed red line on Y value 6.6 on both the original and replication results graphs. The line shows how that value is higher than all observations made on the replication, but, in the original results, all observations made before the 1980s were higher than it. The blue dashed line is placed on a lower value, 6.2, and all observations made on the original study are higher than it, whereas on the replication the observations oscillate going above and below it. Figure 10 is the only one with just one dashed line: all original observations are higher than the Y value of 1.0 while all replication observations (apart from a few towards the end of the examined period) are below it.

Besides what is shown in the graphs below, it was attempted to produce a visualization that overlayed the replication results on top of the original ones, but with no access to the original raw data and results the only possible way of doing so was by manually manipulating vector images of the graphs. Since this method cannot guarantee 100% of precision, it is not included in this section, but is available at the end of this document on Annex B. I decided to keep it in this thesis because it poses some interesting research questions on Data Visualization

and replication when raw data is not available. Advances on this issue can help in the visual comparison of results in replication efforts.

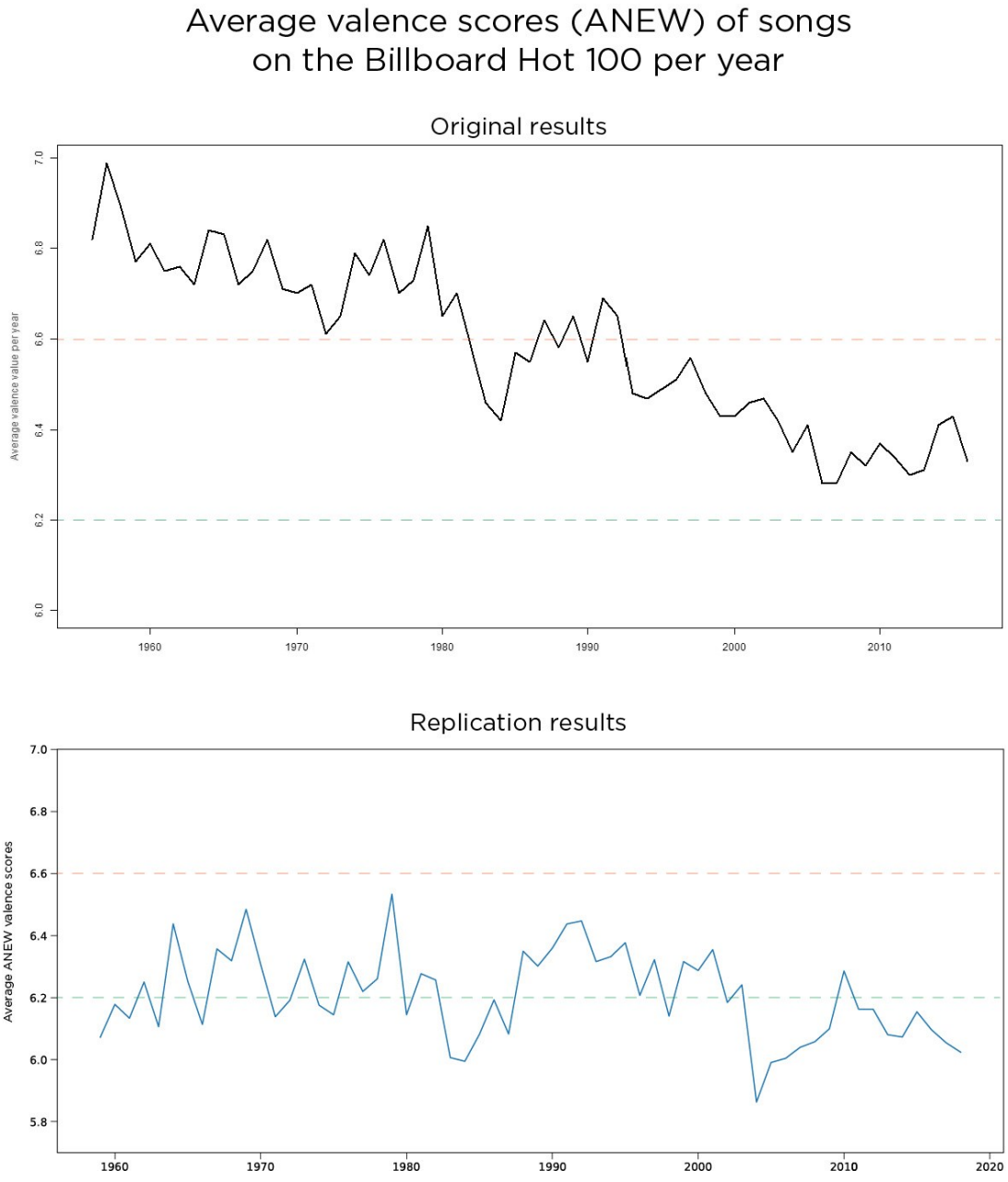


Figure 7. *Results of the original and replicated average valence scores per song on the billboard Hot 100 using the ANEW word list.*

This first result refers to Meinderstma's analysis on changes in sentiment (4). The original study results showed a somewhat constant decline in the average ANEW valence scores over the years. The replication showed a different direction in the first half of the analyzed period, with ANEW scores remaining relatively stable over time until it catches up with the original tendency to decrease over time at around the mid-1980s. There is a pronounced dip in ANEW scores in the mid 2000s that wasn't observed in Meinderstma's results. Also note that the average scores are, overall, lower on the replication. While the original had a maximum value of almost 7 and a minimum value of roughly 6.3, the replication had a maximum of slightly above 6.5 and a minimum of around 5.85 during the dip observed in the mid-2000s.

Figure 8. *Results of the original and replicated average number of words per song on the billboard Hot 100.*

The next set of results all relate to (4) *changes in lexical complexity*. Both the original and the replication results showed a similar trend in the average number of words per song. There is an increase in words per. song over time with a steeper increase around the late 80s/early 90s. Original and replication results differ in the magnitude of the observed word count: whereas the original had a peak of around 400 average words per song, the replication had

a peak of roughly 600. These peaks were both observed in the mid 2000s. The lowest values were also slightly different: the original's lowest average words per song is roughly, in the 180s in 1956, while the replication's lowest score is roughly 200 in 1960.

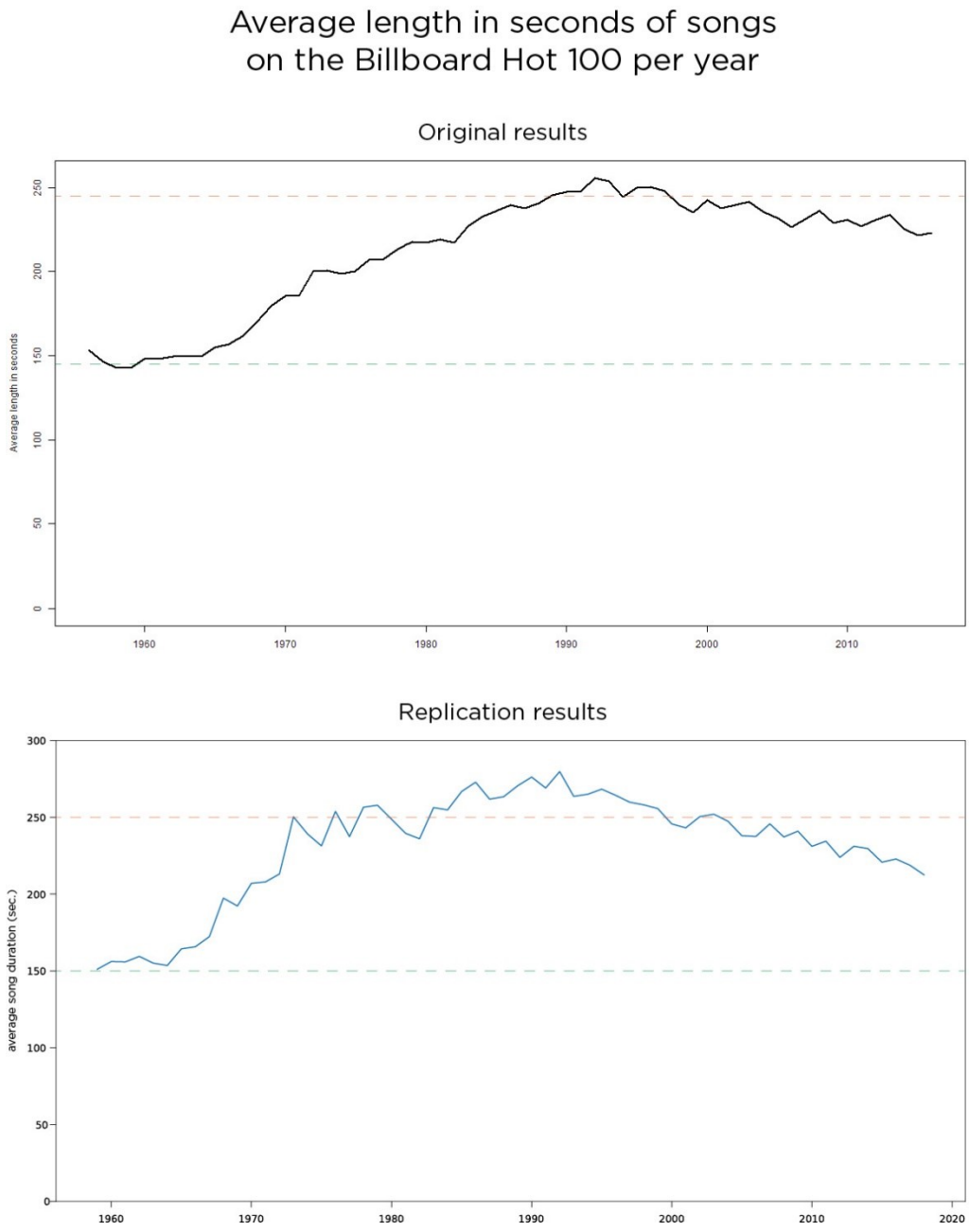


Figure 9. *Results of the original and replicated average song length per song on the billboard Hot 100.*

Average song length also showed a similar trend in both the original and replication results. There is a gradual increase in average song length, and it plateaus in the early 90s. While not very remarkable, average song length shows a slight decrease from the early 90s onwards. Compared to the previous result, the difference in the magnitude of the values is less strong here, both results show a starting point of around 150 seconds (2.5 minutes) in the early 1960s while the highest values are in the 250 seconds (or roughly 4 minutes) mark. These observed values were all expected given that popular music and singles have traditionally been recorded in 78rpm discs that were able to store up to 3 or 4-5 minutes of music depending on the disc being 10 or 12 inches in size ("The history of 78 RPM recordings").

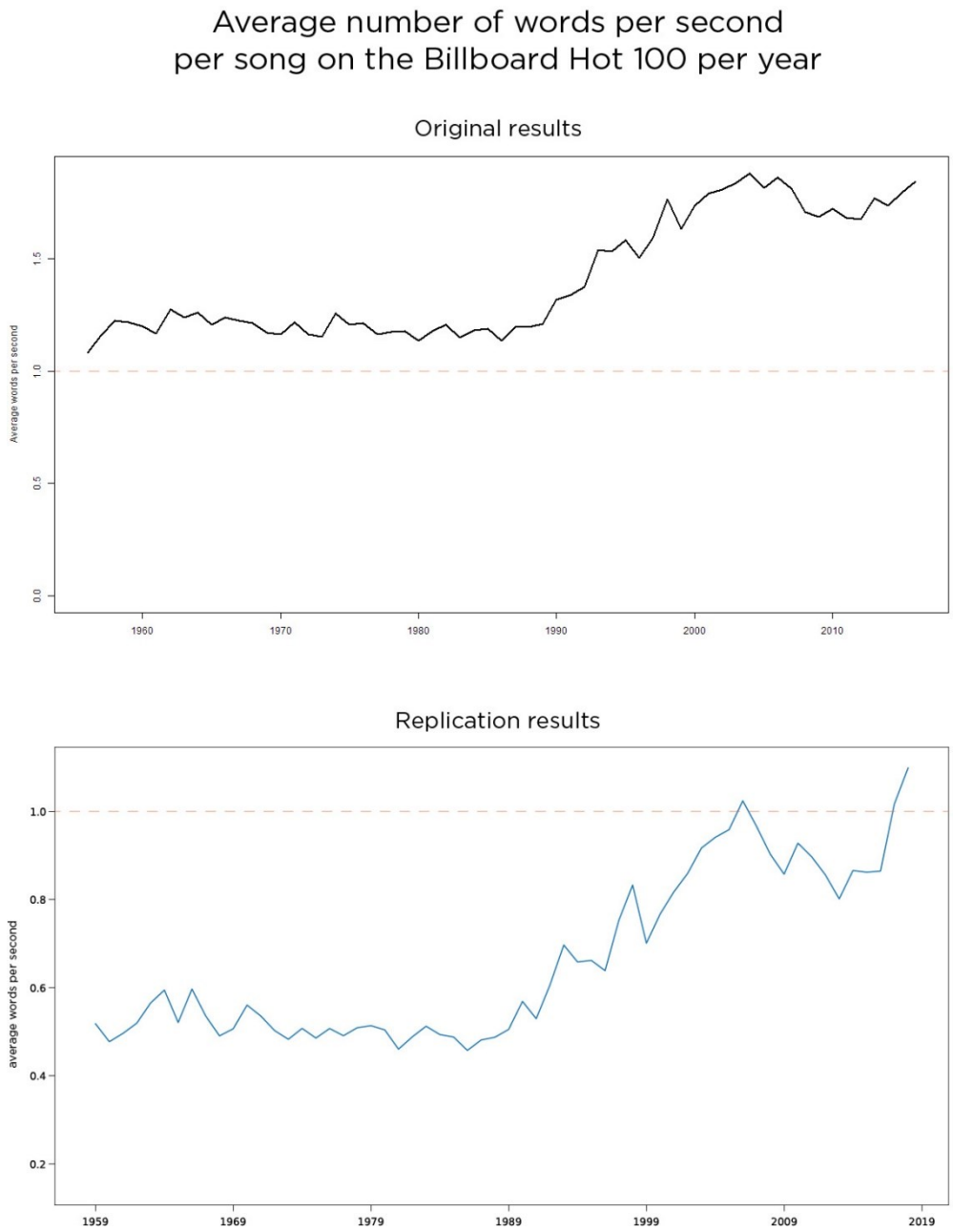


Figure 10. *Results of the original and replicated average number of words per second per song on the billboard Hot 100.*

The average number of words per second is the ratio between the two previous results. Based on the fact that both showed similar trends in the original and replication studies, it follows that the average number of words per second also shows a similar trend: stable values until the late 1980s/early 1990s, and then a steep increase is observed leading to another period

of stability. A difference in magnitude is also seen here: the original values range from roughly 1 to 2 while the replication ranges from 0.5 to 1, but the overall trend is similar.

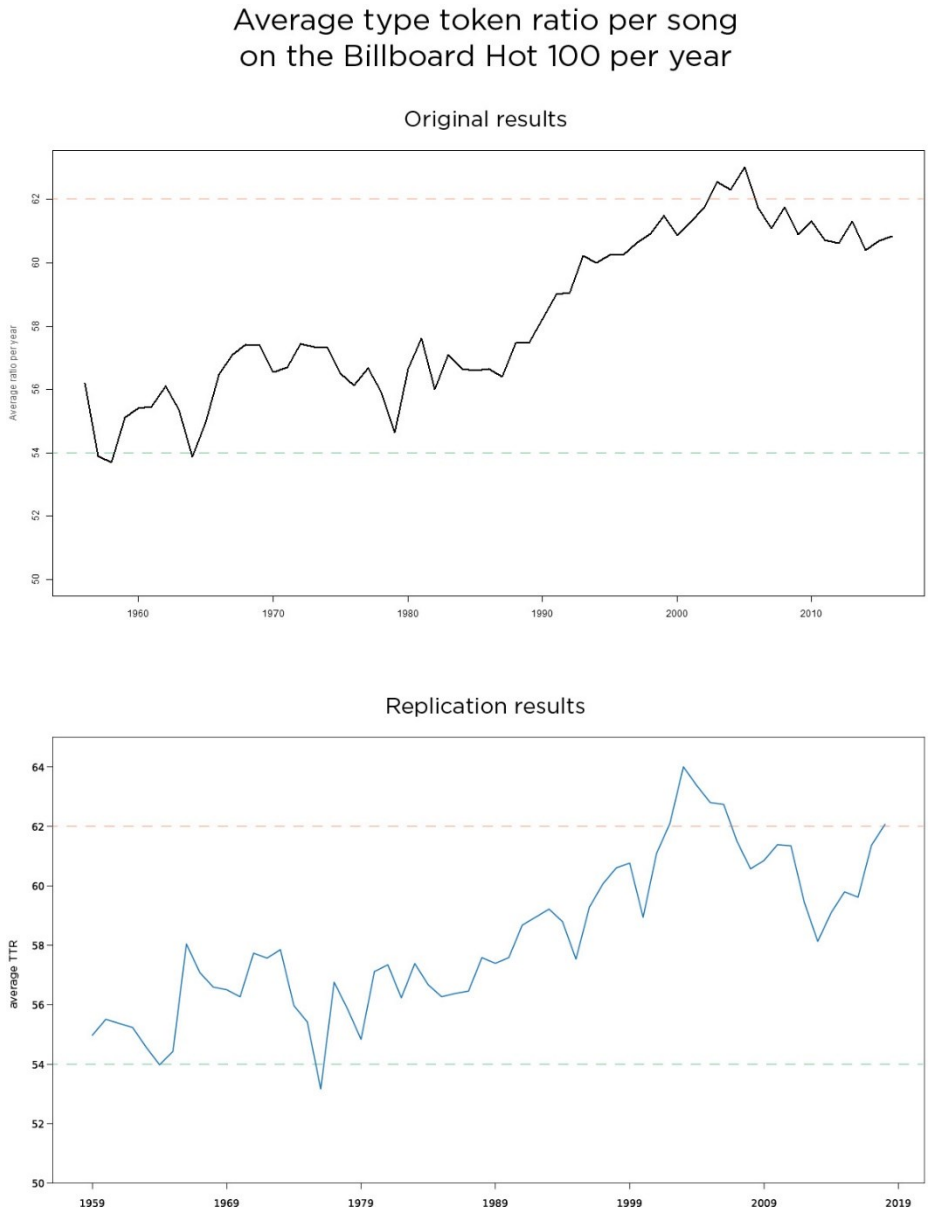


Figure 11. *Results of the original and replicated average type token ratio per song on the billboard Hot 100.*

The type token ratio of a text is the total number of words divided by the total number of unique words. The average TTR per song followed a similar trend in both original and replication studies. Before the late 80s/early 90s, the averages oscillated between around 54 and 58. It then steadily increases and peaks in the mid 2000s, at an average TTR of roughly 64. The magnitude of the values in both the original and replication studies remained much more similar than what was observed in the previous results in this set. There was a pronounced dip in the average TTR in the mid-2010s in the replication study that was not observed in the original study. TTR is a very sensitive metric and a closer look at the period of the line dip could reveal if there are songs in that period that didn't follow the overall post-late 80s/early 90s trend. As it will be discussed later, the introduction of hip-hop in the billboard charts is one of the main explanations for the increase in TTR (as well as the average number of words per song and average number of words per second), but it is not clear if the number of hip hop songs present in the Billboard Top 100 charts remains stable throughout the 90s and onwards.

4.2.2 Results: Conceptual replication

The following results include conceptual replications of Meinderstma's study and its claims as well as extensions that can help expand the boundaries of what his original results allowed to know.

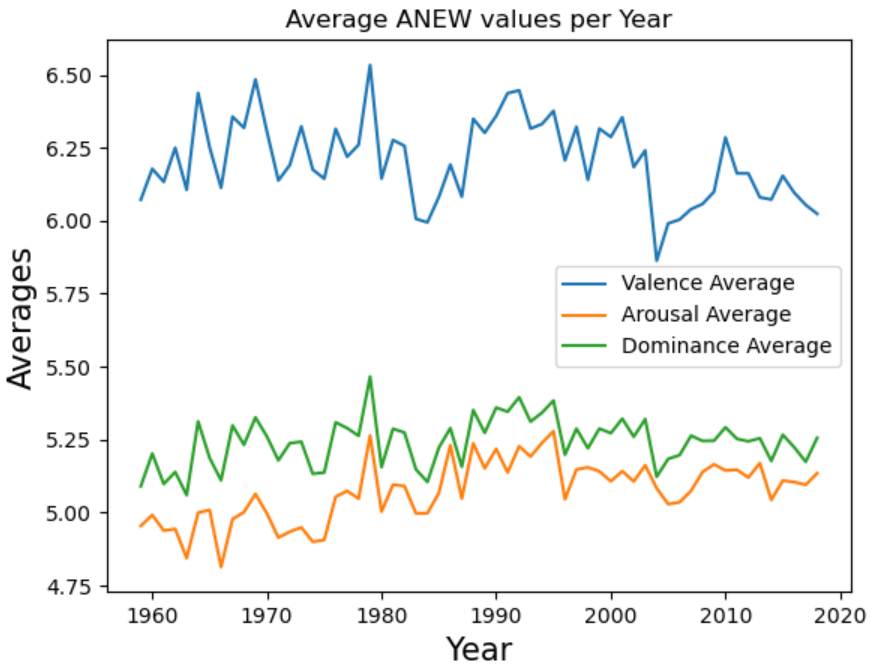


Figure 12. *Results of the average valence, arousal and dominance scores per song on the billboard Hot 100 using the ANEW word list.*

Meinderstma used ANEW Valence scores to quantify changes in sentiment. The blue line on Figure 11 is the same that was shown on Figure 7 when the original and replication average valence scores were compared. The orange and green lines refer to average ANEW arousal and dominance scores. Dominance oscillates over time, but no clear trend can be seen. Arousal also oscillates over time, especially before the mid 80s, when there is an overall increase in arousal

that persists towards the mid-90s when there is a slight decrease in the value that persists until the end of the analyzed period. It follows that lyrics appear to have become slightly less "calm" starting from the mid-80s.

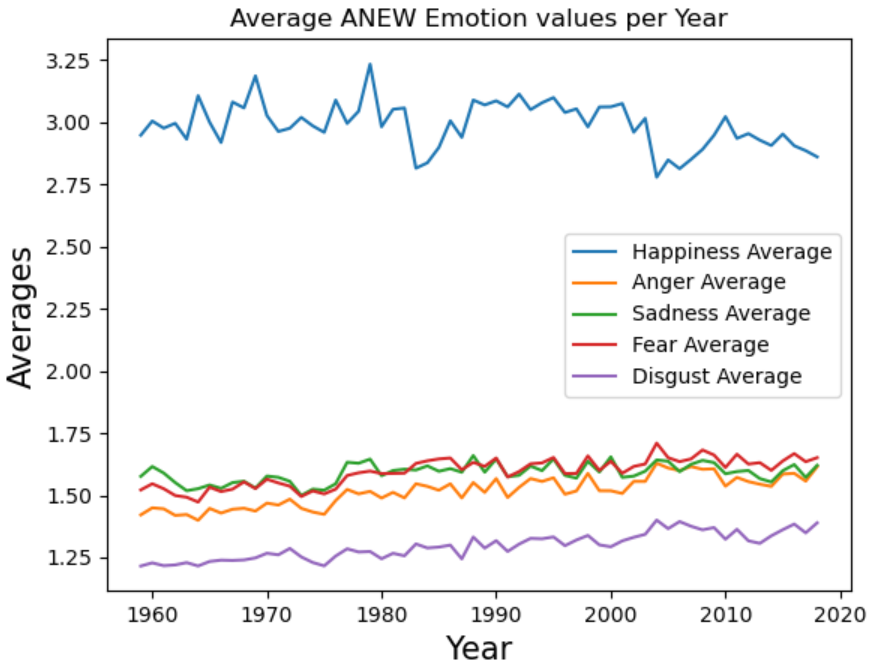


Figure 13. *Results of the average happiness, anger, sadness, fear and disgust scores per song on the billboard Hot 100 using the ANEW word list.*

Figure 12 shows the results of the average discrete emotion values and serve as a complement to the previous ANEW results. Besides oscillations over time, there is not a very strong trend that is observable. Meinderstma states that "songs on average are less "happy" than ever before" (p. 6), but happiness scores remained stable throughout the analyzed period. There is a dip in the line during the mid 2000s and, once it moves up again, an apparent slight decline over the next years before the end of the period. Not a strong enough decline at the end, nor a

strong enough decline over the decades to support the original claim. It is not clear why the original results showed a clear decline in valence while replication did not show the same trend. It should be noted that Meinderstma's database included weekly Top 100 songs, while the replication database that I used relied on yearly Top 100. Could it be the case that songs with lower valence tend to stay longer in weekly charts, therefore increasing their weight when averaged? The other four discrete emotions also showed a stable line over the analyzed period, with a very slight increase over time, but certainly not enough to support claims such as "there was an increase in disgust in Top 100 lyrics over the years". The initial values are lower than the final values, but the magnitude of the change is small. This applies to anger, sadness, fear and disgust average scores.

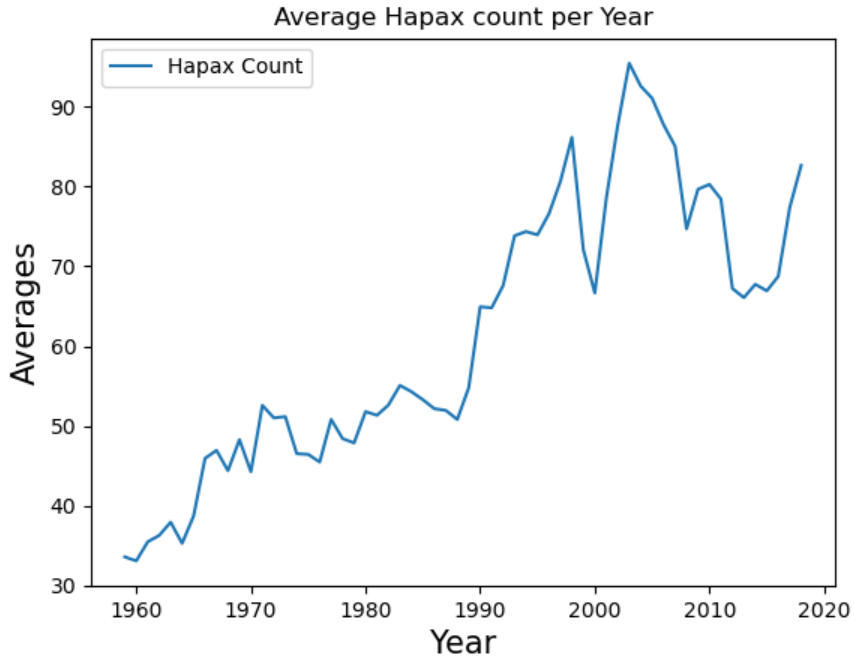


Figure 14. *Results of the average hapax counts per year on Top 100 Billboard songs.*

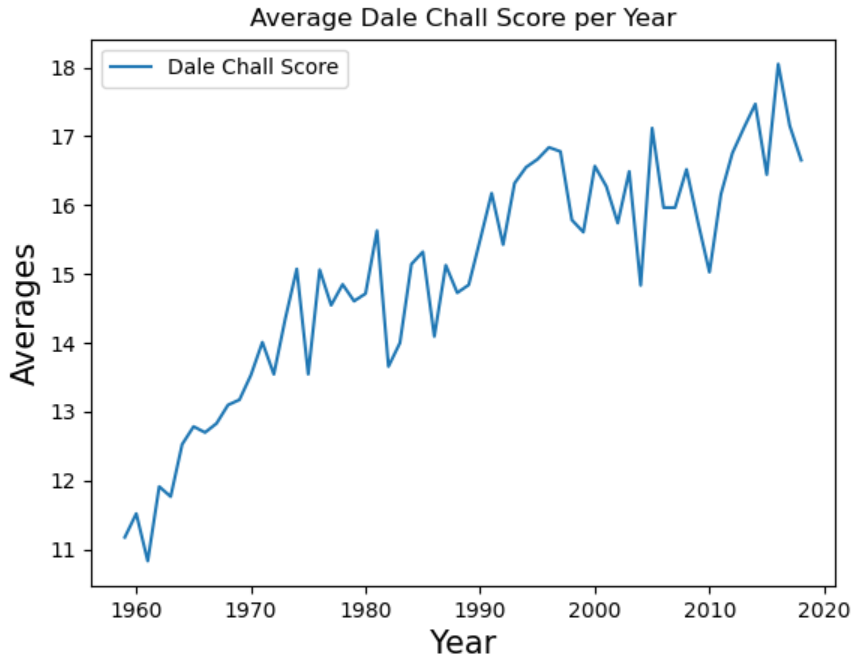


Figure 15. *Results of the average Dale Chall scores per year on Top 100 Billboard songs.*

The results shown in figures 13 and 14 validate Meinderstma's claims of an increase in lexical diversity over time in Billboard Top 100 songs (p. 8). There is a clear increase in hapax counts over the years, indicating that there was an increase in the use of unique words and, likely, a larger vocabulary. The Dale Chall scores also showed a constant increase over time, indicating that songs have not only become more lexically diverse, but more complex words have also become more commonly used. The hapax counts saw a steep increase in the late 80s/early 90s period, reinforcing the notion of a strong increase in lyrical complexity at that period. However, this was not observed in the Dale Chall results: the increase was steady over time, with an almost linear increase until the 70s and then, even though the increase continued, it became more oscillatory.

Although there was an increase in the Dale-Chall Score over the years, a closer inspection of that metric reveals a more nuanced understanding of that result. The Dale-Chall

Score takes into account the ratio between the presence of words considered difficult (difficult words are any word outside a list of 3000 words considered easy to understand by a fourth-grade student) and the total number of words in a text. By looking closely at what the Python script determined as being difficult words present in the lyrics and counting the top 10 of these words added up by year, it is possible to identify a few words that are clearly present in all years. One notable example is the word “yeah” a word that, even though it could be argued is not difficult, is not present in the Dale-Chall list of words. “Yeah” was the top 1 word for several years and was present as a top 5 in almost all years being analyzed. One difference between the early and final years of the range is the frequency “yeah” is present in lyrics. In 1959 it was sung 117 times whereas in 2018 that number jumped to 677. From 1959 to 1969 the average frequency was 143 and from 2008 to 2018 the average increased to 423.5. Other words considered difficult and that were heavily represented include “ooh”, “na”, “doo”, “la”, “uh”, “ai”/”ay”. All of these, again, wouldn’t be considered difficult words, but are arguably relevant for rhythmic or melodic singing by a performer. Depending on the song, some of these words can be over-represented and influence the statistical analysis of word frequency. For example, the song “Havana” by Camila Cabello, ranked top 97 in 2017 and top 4 in 2018, contains lyrics that can be exemplified as:

Havana, ooh na na (ayy)

Half of my heart is in Havana, ooh na na (ayy, ayy)

He took me back to East Atlanta, na na na, ah

Oh, but my heart is in Havana (ayy)

There's somethin' 'bout his manners (uh huh)

Havana, ooh na na (uh)

Ooh na na, oh na na na (ooh, ooh ooh ooh ooh ooh ooh)

Take me back, back, back like...

Ooh na na, oh na na na (yeah, babe)

Take me back, back, back like... (lyrics from Havana by Camila Cabello)

The song features the word "na" 88 times and the word "ooh" 56 times. That's roughly a third of the total number of occurrences of "na" and roughly a fourth of the total "ooh"s in 2018. Because of the way the Python tokenizer I used works, "na" is already overrepresented since it is the leftover part of "gonna" and "wanna" when these two words are stemmed. On the other hand, both "gonna" and "wanna" are not present in the Dale-Chall list of easy words ("want" and "going" are on the list though). In conclusion, the point is, many of the most frequent difficult words in the Dale-Chall sense, are syllables used for singing and word contractions such "'cause", "'bout", "gonna" and "wanna". Another word category that may increase the Dale-Chall score, especially around the 1990s, are curse words. Meinderstma had already presented in his results a sudden spike starting in the 1990s in the use of swear words (p. 5). The analysis of the presence of difficult words confirms that finding. The word "bitch" was even the top 8 difficult word used in 2017 and, although this was the only instance in which a swear word appeared in the top 10 most used words, swear words were frequently used from the 1990s onwards.

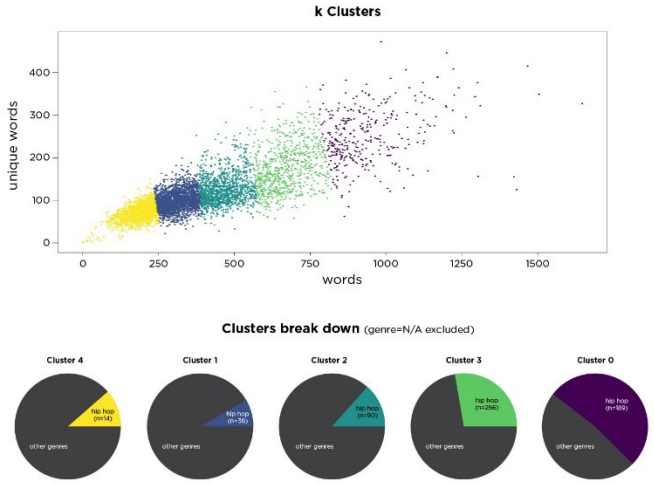


Figure 16. *Distribution of k nearest neighbors' clusters when plotting "total number of words" x "total number of unique words" in Top100 songs. Pie charts represent the proportion of hip-hop songs to the total for each cluster.*

The last result serves to test Meinderstma's statement that the increase in type token ratio in the late 80s/early 90s coincide with the increase in popularity hip hop songs (p. 9). By clustering together songs using k nearest neighbors based on their number of words and unique words, it was expected that songs in the hip hop genre would be more frequent in the clusters with higher word and unique word counts. This was indeed observed. In Cluster 0, the one with the highest counts, 186 out of 363 were hip hop songs (roughly 50%), and many of the remaining songs on that cluster were R&B or related genres. Only 13 songs on that cluster (less than 0.1%) were released before the 1990s which helps validate the idea that the introduction of hip hop in the Top 100 is a probable explanation for the increase in TTR in the late 80s/early 90s.

4.3 Discussion

(RQ 4a): *Does the direct replication of a published Digital Humanities research project confirm the original findings?*

There is currently no consensus on what a successful replication is (Cramerer et al., 2018, Open Science Foundation, 2015). For the replication project being described here, it was not possible to conduct a more rigorous statistical comparison between the original and replication results due to the raw numbers of the original paper not being readily available. What was possible to do was to compare the original and replication graphs and the claims made by Meinderstma, the original study's author.

This replication collected the Billboard Top 100 song lyrics data again and wrote new code with the intent of generating, ideally, results that coincided with the ones published by Meinderstma, therefore confirming and increasing the confidence that we can have in them. Replication was conducted on changes in sentiment and changes in lexical diversity, as studied by Meinderstma. The replication results diverged in the first and showed quite similar results in the latter. Based on the evidence gathered in this replication, the original changes in lexical diversity results replicated and the changes in sentiment results did not replicate.

(RQ 4b) *Does the conceptual replication of a published Digital Humanities project, using methods other than the ones used by the original author, confirm the original findings?*

The conceptual replication conducted for this thesis reinforced the direct replication results. The two methods used for testing changes in lexical diversity, hapax count and Dale Chall score, reinforced the validity of the original claim that there was an increase in lexical

diversity over time. The Hapax count also reinforced the idea that there was a sharper increase in lyrical diversity during the late 80s/early 90s period. On the other hand, the conceptual replication of changes in sentiment did not confirm the original claims. An analysis of the average discrete emotion scores did not show a steady decline in happiness as stated by Meinderstma, instead, happiness scores remained relatively stable over the analyzed period.

The collected data was also analyzed with the goal of testing Meinderstma's hypothesis that the increase in lexical complexity in the Billboard Top 100 songs can be explained, at least to some degree, by the emergence of Hip Hop as a popular genre from the 90s onwards. Evidence was gathered to show that roughly half of the most lexically complex (i.e., high type and token numbers) songs in the database were Hip Hop songs and a significant number of the other half were songs from genre with Hip Hop influences or an influence on Hip Hop, such as R&B. It should be noted that Meinderstma didn't test this hypothesis in his paper, therefore this part of the replication effort shouldn't be considered a direct or conceptual replication, but rather an extension of his original study.

The remainder of this section will focus on topics that are relevant to this thesis but do not relate directly to the two research questions discussed above. The first of them being my experience recreating the Billboard Top 100 dataset. As it was already mentioned in the methodology section, Meinderstma's paper didn't provide direct access to his dataset, and I refrained from contacting him. I only conducted one replication for this thesis and understanding if I'd be able to have access to the original study's dataset would not provide any sense on the overall trend in the Digital Humanities on weather DH researchers are willing to share their data (and code, for that matter) upon request. Instead, recreating the dataset gave me a sense of how

much work is involved in recreating a dataset in the context of a replication. Meinderstma provided information on a website he used to have access to the list of Billboard Top 100 songs (Bullfrog's Pond), but that website was not available on the internet anymore when I tried to access it, which essentially made me look for a solution myself – essentially starting from zero.

The process involved learning a few API's and Python libraries to eventually be able to have a usable lyrics database. This included being able to extract hundreds of Top 100 tables from Wikipedia, aggregate and use them to retrieve data from the Genius lyrics API and Spotify API. This process was time consuming both in terms of learning how to use the APIs as well as waiting for them to process all the 6000 requests. I tried my best to write Python code that could handle any errors, but the API would eventually encounter unforeseeable issues when retrieving data, which would require, not only supervision, but also the understanding of how to handle the error. The Spotify API was, overall, very effective in retrieving all the requests that were thrown at it and, except for a very few wrong songs being retrieved, the only time consuming and laborious part was manually consolidating music genres (e.g., West Coast Rap, East Coast Rap, Gangsta Rap, etc. were all changed to simply Hip Hop). On the other hand, Genius had difficulties retrieving many of the lyrics. It failed to retrieve 500 lyrics entirely and, although in the Methodology section I simply described using google to manually retrieve, clean up and store the lyrics in the database, that process actually took almost one week. I did not quantify the time and effort needed to recreate the database in order to provide a more objective assessment, but subjectively it became clear to me that any larger scale replication project would be quite hard if conducted by only one person and without direct access to the materials used in the original projects.

The Digital Humanities, and here I am mostly speaking about DH projects dealing with cultural artifacts turned into quantitative and computational data, lies in an interesting space when compared to other disciplines. Compared to, for example, psychology, there is no need to gather a new sample of people to try and replicate a previous finding, subjects who would, invariably, be different than the original ones. The lyrics contained in Meinderstma's Top 100 Billboard database are exactly the same as the lyrics contained in the database I recreated. What follows, I argue, is that the same logic applies to Shakespeare plays, 18th Century French Literature, Modernist Paintings or tweets posted immediately after the 2022 US Election.

But DH projects tend to rely on large quantities of cultural data, which leads to my next point: how can we be sure of the accuracy and integrity of the vast datasets being fed into our Python or R scripts and the results coming out of them? I did my due diligence of paying attention to all the details of my data collection process. Randomly checked data points to verify its integrity and tried to fix any issues I found along the way. At the very end of the project, when I was exploring the KNN graph and looking at one visible outlier I noticed that data point referred to the lyrics of a song that did not match its Top 100 chart and position. The data point referred to the lyrics of a Wu Tang Clan song (a Hip Hop collective) when in fact it should have been the lyrics of a song by the Isley Brothers (a Funk group). At that point in the process, I did not have enough time to fix it and rerun all the codes to obtain updated results. That was one song in one of the 60-year period analyzed, so at most it would increase slightly the average word count of one year in the 70s. The real problem starts when these types of problems compound and affect the overall results and findings of research. Remedying this exact issue is one of the main points of conduction replication: to increase our confidence in and validate scientific findings.

To add to the topic of replication and an increase in confidence, I'd like first to reiterate that a replication effort that obtains results that are consistent with the original ones, increases our confidence in the original claims made by the authors (Nosek and Errington, 2020), and I'd like to point out two replication scenarios. In the first, the researcher of the original study makes their data and code available, and that data has issues embedded in it. A second researcher (or team) then uses that data to rerun the code and, ideally, obtain the same results that were published. Simply rerunning the code by someone else cannot change the fact that there were issues already present in the data and this replication does not have the ability to change our confidence in the original results (or as goes the saying in computer science: garbage in, garbage out). It is, thus, necessary for the replication team to check for any potential data integrity issues if the replication effort in the first scenario has any goals of increasing confidence in the original results, otherwise it merely checks if the published results match what was obtained from the data analysis. In the second scenario, the replication researcher (or team) recreates the dataset used by the original researcher from the ground up. Although there is always a possibility that errors may still be present, it is expected that the ones present in the original dataset are now absent in the replication. The second scenario, as pointed out in the literature review (Feest, 2019; Hudson, 2021; Nosek and Errington, 2020; Stroebe and Strack, 2014), has a greater potential to validate the original claims.

Lastly, because the data and code used on Meinderstma research was not made public, a computational reproduction of the project was not possible. If I had contacted the author and requested them, there was a possibility I would have gotten access to the code and/or data. Instead, I performed a replication that required collecting the lyrics data again and rewriting the code for analyzing it. The type of data used in Digital Humanities presents some attributes that

differentiate it from the data discussed in the literature review presented previously. Compared to, for example, a group of humans and their behavior when in a group (such as what is studied in Psychology) which may vary greatly depending on context, the individuals forming the group, etc., song lyrics are static. The lyrics I collected for this replication project are the same as the ones analyzed by Meinderstma. The reconstructed dataset can be, in theory, exactly like the original dataset. In the case being discussed here, the only factor that limited a more exact reconstruction of the original dataset was the website where Meinderstma used to retrieve his Billboard song list from was not available anymore and my methodology for retrieving a new list, limited my dataset to contain one Top 100 list per year, instead of one Top 100 list per week. Despite this difference, the results in changes in lexical complexity were similar enough that the same conclusions could be drawn. Changes in sentiment, however, did not point to the same direction and, even though further work could help clarify what may have caused the difference in the original and replication results, corpus construction (Piper, 2022) and the resulting observations could have played a role here. To borrow from Derksen and Morawski's (2022) idea of replication as an enactment of a reality, there is a possibility that both Meinderstma and I observed unique constructed realities, and neither his nor mine results are necessarily right or wrong, but they can complement and inform each other. Future work could explore further the idea of corpus construction, enacted realities and how this relationship plays out in the context of DH.

5. Limitations and future work

The Digital Humanities journals and published papers survey conducted for this thesis focused on studies published in the year 2021. Given that all papers published that year in the surveyed journals were analyzed, the survey should be a snapshot of transparency indicators for that year. But because this analysis focused on only one single year, it is not possible to determine if there is an increase or decrease in the number of researchers and published research following openness and transparency guidelines. The analysis also did not include journals that specialize in humanities disciplines such as History, Archeology, Musicology and others. Future work should aim to fill these gaps and understand how research in the various humanities fields that rely on empirical data compare to each other in terms of adoption of transparency practices and if field-specific journal policies lead to increased indicators.

The second project conducted, the replication of a published DH research, was a case study that aimed to explore and understand potential difficulties and issues associated with conducting a replication. The project was limited by the lack of access to the original code and analysis scripts used by the original author which (although this was research design decision) restricted the type of statistical analysis that could be made to evaluate and compare the original and replication results. Future work related to the replication conducted here should aim to improve the statistical methods used for comparing the original and replication results as well as continue to explore what it is still possible to learn and understand about popular music. Meinderstma (2016) already suggested topics for further research such as conducting research on popular music in other countries and languages. Beyond additional work for this single replication, the Digital Humanities would greatly benefit from a larger scale replication study, similar in scope and scale to the studies conducted in other scientific fields, to evaluate how

much of what is published and claimed as truth is verifiable and confirmed by other researchers so our confidence in DH and its methods can be further solidified. A question that is also worthy of answering by the DH community is what are the parameters of a successful replication in the context of DH? That answer would guide what to strive for when performing replications in DH by accounting for the computational methods specified to the field as well as the specificities in the types of data examined by DH researchers.

6. Conclusion

With the aid of methods from the field of metaresearch, a discipline that focuses on conducting research on research itself (Ioannidis, 2015), and discussions actively happening in that field, this thesis attempted to understand if DH could potentially be suffering from similar issues observed in studies that led to the conclusion that science is currently going through a "Reproducibility Crisis". I attempted to understand this through the three main constituent elements of this thesis: a literature review which surveyed current literature investigating reproducibility and replicability issues in science as well as efforts being made to remediate these issues; a survey of literary criticism and Digital Humanities journals and papers aimed at measuring and understanding their transparency indicators; and the direct and conceptual replication of a published DH study.

The literature reviewed for this thesis shows that we are currently going through a "Reproducibility Crisis", a crisis that can potentially affect our confidence in science and in the claims made using its methods. The large-scale Open Science Collaboration effort "Reproducibility Project: Psychology" (OSC, 2015) was one of the initial replication efforts that demonstrated evidence of a crisis. It showed that only roughly 40% of the attempted replications confirmed the original finding. Similar studies were later conducted in other fields including cancer biology, economics and social sciences (Errington et al., 2021a; Camerer et al., 2016; Camerer et al., 2018), all of which resulted in less-than-ideal replication rates. A self-reporting survey on replication showed similar results, with 70% of the respondents answering that they have failed to replicate other researchers' experiments (Baker, 2016). These findings are made more significant when there is evidence that nonreplicable research tends to be cited much more than replicable ones (Serra-Garcia and Gneezy, 2021).

Although the causes of the Reproducibility Crisis are still being investigated, there is evidence suggesting the existence of a high degree of pressure on researchers to publish in the competitive academic field (a culture of "publish or perish") (Fanelli, 2010) a situation that can potentially lead researchers to engage in what has been labeled as "questionable research practices", ranging from only reporting studies that worked, to falsifying data (Leslie et al.'s, 2012). These practices increase the chances of a research manuscript being accepted for publication despite lowering the probability of said research being replicable (Munafò, 2017). This situation is aggravated by a tendency from researchers to publish and cite positive results (which can be more easily achieved via "questionable research practices"), as well as a tendency from journals and reviewers to accept and publish research showing positive and "interesting" results, a practice called "publication bias" (Sterne et al., 2001; Landis et al., 2012; Duyx et al., 2017; Mlinarić et al., 2017). The effects of the competitive landscape in the academic field could explain, at least to some degree, the low reproducibility numbers observed in metaresearch studies.

The low transparency indicators observed in different scientific fields, including lack of access to data and analysis code (Alsheikh-Ali et al., 2011; Collberg and Proebsting, 2016; Iqbal et al., 2016; Wallach et al., 2018; Hardwicke et al., 2020; Raghupathi et al., 2022), increase the difficulty of measuring reproducibility and hinders replication efforts. Deficient transparency levels limit the opportunities for investigating and understanding the "Reproducibility Crisis", its potential causes and effects, thus contributing to lowering the confidence in science. A proposed remedy for the "Reproducibility Crisis" is for the scientific community to switch the focus from novel and "interesting" results published in single studies, to the totality of evidence collected in, ideally, well-designed and well-powered studies conducted by multiple research teams with the

goal of corroborating or refuting claims and theories (Popper, 2002; Ioannidis, 2005; Simmons, 2014; Goodman et al., 2016). To conclude, replication is a crucial tool available for science to self-correct and achieve the aforementioned goal, and transparency ensures replication efforts can be conducted effectively.

This thesis gathered evidence, via a survey of journals and papers, to an understanding of transparency in literary criticism and Digital Humanities. By evaluating nine journals (four focused in Digital Humanities research and five focused in literary criticism) and 526 articles published by them in the year 2021, it was possible to create a snapshot of transparency levels during that year. The survey showed that 40% of all the articles were available as open access, a figure that is comparable to the open access levels in the social sciences, as measured by Hardwicke et al. (2020). Open access levels were significantly lower in the papers published in literary criticism journals (19%) and higher in the DH ones (68%). There is a very low number of studies reliant on quantitative and/or computational methods published in the surveyed literary criticism journals and, combined with the ones published in the DH journals and reliant on such methods, a total of 110 studies were surveyed. Of these, 65% shared the data used and 36% shared the code. It was observed that strict journals guidelines on data and code submission (clear instructions on how and where to submit data and code and requiring authors to do so in order to have their manuscript accepted) results in almost 100% of code and data being shared by authors. There does not seem to be much difference in results when a journal suggests authors to submit their data and code or says nothing, although more research is needed to validate this hypothesis.

Besides the quantified results reviewed above, the survey yielded some additional information relevant to the topic of this thesis, even though they were not collected and reviewed in a systematic manner. The first of them relates to the number of surveyed papers that were replication of another study: only one paper out of 110 was a replication (Rizvi, 2021). The second issue identified was several broken URLs that should have led to data and/or code. This issue affected Rizvi's replication as well as my own replication (described below). Lastly, a few reasons on why Digital Humanities data may not be readily or easily made available were identified and include data embargoes; copyright and legal reasons; file size and/or large volume of files; data availability restricted to specific parts of the world (e.g., library systems in one country).

The last component of this thesis was the replication of a Digital Humanities research: Meinderstma's Changes in Lyrical and Hit Diversity of Popular U.S. Songs 1956-2016, published in 2019. The focus of the replication was recreating the database of Billboard songs and attempting to replicate two out of four of the studies conducted by the original author: changes in lexical diversity and changes in sentiment. The replication results showed an increase in lexical complexity, a result that had also been observed by the original author. Besides using the same methods employed by Meinderstma to measure changes in lexical complexity, additional methods were also used to corroborate the original findings. Additionally, the replication gathered evidence to corroborate Meinderstma's hypothesis that hip hop songs were responsible, at least in part, to the increase of lexical complexity in popular US songs starting from the early 1990s.

When attempting to replicate changes in sentiment, the results differed from what was observed by Meinderstma: while he observed a decrease in valence and stated that popular songs are less happy than ever, the replication results showed oscillating results over time, with no clear trend towards a decrease or increase in valence. By using a more granular analysis approach in the replication, a decrease in happiness was also not observed. It is not clear at this time what could have caused this difference in results and further research could shed light on this. Additional ways of measuring and tracking changes in sentiment (e.g., machine learning or per sentence analysis rather than word counts) could also lead to a better understanding of the matter.

The main goal of this replication effort was not an attempt to measure reproducibility in DH, since that would require a much larger number of replications instead of only one, but to understand the challenges and struggles associated with the recreation of a database and to promote a culture of replication in DH, one that has a spirit of community at its core and in which researchers who choose to engage in replication efforts do so with the main goal of building larger and stronger evidences for or against theories. Increased transparency in DH could help save the time and energy of researchers willing to replicate DH projects: time and energy that could be better used to further study claims and theories and keep building on what had already been started by other authors. The journal and papers survey conducted for this thesis show that there is much room for improvement towards higher transparency indicators in DH. To keep with the spirit of everything that was discussed throughout this thesis, all data and code relevant to both the survey and replication components of this thesis are available on the

McGill University Dataverse[15] repository (Tiefenbach Keller, 2023), and code snippets for all calculations done for the replication project are also available at the end of this document, on Appendix A.

To conclude, it is my hope that the evidence gathered for this thesis can serve as a starting point for further investigation and an improved understanding of transparency and reproducibility issues in the Digital Humanities. These advances can eventually lead to refined guidelines and policies across DH journals aimed at promoting and facilitating replication efforts as well as an increase in interest by the DH community and its practitioners in the relevance and importance of replication as a tool for increasing the confidence in the knowledge being generated and shared in the field of the Digital Humanities.

[15] All data and code relevant to this thesis can be accessed via the following URL: https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/KRPCBL

Bibliography

Allison, David B. et al. "A tragedy of errors". *Nature*, Vol. 530, 2016.

Altmejd, Adam et al. "Predicting the replicability of social science lab experiments". *PLoS ONE*, Vol. 14, No. 12, 2019.

Alsheikh-Ali, Alawi A. et al. "Public Availability of Published Research Data in High-Impact Journals". *PLoS ONE* 6(9): e24357, 2011.

Altmejd, Adam et al. "Predicting the replicability of social science lab experiments". *PLoS ONE*, Vol. 14, No. 12, 2019.

"Announcement: Reducing our irreproducibility". *Nature*, Vol. 496, No. 7446, 2013.

Arthur, Paul Longley et al. "Open scholarship in Australia: A review of needs, barriers, and opportunities". *Digital Scholarship in the Humanities*, Vol. 36, No. 4, pp. 795-812, 2021.

Baker, Monya. "1,500 scientists lift the lid on reproducibility". *Nature* 533, 452–454, 2016.

Barba, Lorena A. "Terminologies for Reproducible Research". *arXiv*:1802.03311, 2018.

Barnett, Adrian G. et al. "Benefits of Publicly Available Data". *Epidemiology*, Vol. 23, No. 3, pp. 500-501, 2012.

Berry, David M. "Introduction: Understanding the Digital Humanities". *Understanding Digital Humanities*, edited by David M. Berry, Palgrave Macmillan, 2012.

Boulbes, Delphine R. et al. "A Survey on Data Reproducibility and the Effect of Publication Process on the Ethical Reporting of Laboratory Research". *Clin Cancer Res*, Vol. 24, No. 14, pp. 3447–3455, 2018.

Bradley, Margaret M. and Lang, Peter J. "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings". *Technical Report C-1*, The Center for Research in Psychophysiology, University of Florida, 1999.

Brown, Paul et al. "How the Word Adjacency Network (WAN) works". *Digital Scholarship in the Humanities*, Vol. 37, No. 2, 2022.

Camerer, Colin F. et al. "Evaluating replicability of laboratory experiments in economics". *Science*, Vol. 351, No. 6280, 2016.

Camerer, Collin F. et al. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015". *Nature Human Behaviour*, Vol. 2, No. 9, pp. 637-644, 2018.

"Checklists work to improve science". *Nature*, Vol. 556, pp.273-274, 2018.

Claerbout, Jon F. and Karrenbach, Martin. "Electronic documents give reproducible research a new meaning". *SEG Technical Program Expanded Abstracts*, pp. 601-604, 1992.

"Computational Literary Studies: A Critical Inquiry Online Forum". *Critical Inquiry*, https://critinq.wordpress.com/2019/03/31/computational-literary-studies-a-critical-inquiry-online-forum/, accessed 25 April 2023.

Collberg, Christian and Proebsting, Todd. "Repeatability in Computer Systems Research". *Communications of the ACM*, Vol. 59, No. 3, 2016.

Collins, Francis S. and Tabak, Lawrence A. "Policy: NIH plans to enhance reproducibility". *Nature*, Vol. 505, No. 7485, pp. 612-613, 2014.

Craig, Hugh. "Authorial Attribution and Shakespearean Variety: Genre, Form and Chronology". *Shakespeare Survey 70: Creating Shakespeare*, edited by Peter Holland, Cambridge University Press, 2017.

Da, Nan Z. "The Computational Case against Computational Literary Studies". *Critical Inquiry*, Vol. 45, No. 3, pp. 601-639 , 2019.

Derksen, Maarten and Morawski, Jill. "Kinds of Replication: Examining the Meanings of 'Conceptual Replication' and 'Direct Replication'". *Association for Psychological Science*, Vol. 17, No. 5, pp. 1490-1505, 2022.

De Vrieze, Jop. "'Replication grants' will allow researchers to repeat nine influential studies that still raise questions". *Scienceinsider*, 11 July 2017, https://www.science.org/content/article/replication-grants-will-allow-researchers-repeat-nine-influential-studies-still-raise, accessed 16 April 2023.

Duyx, Bram et al. "Scientific citations favor positive results: a systematic review and meta-analysis". *J Clin Epidemiol*. 2017.

Earhart, Amy E. "Citational politics: Quantifying the influence of gender on citation in Digital Scholarship in the Humanities". *Digital Scholarship in the Humanities*, Vol. 36, No. 3, pp. 581–594, 2021.

Ebersole, Charles R. et al. "Many Labs 3: Evaluating participant pool quality across the academic semester via replication". *Journal of Experimental Social Psychology*, Vol. 67, pp. 68-82, 2016.

Egan, Gabriel et al. "I would I had that corporal soundness': Pervez Rizvi's Analysis of the Word Adjacency Network Method of Authorship Attribution". *Digital Scholarship in the Humanities*, Vol. 38, No. 4, 2023.

Errington, Timothy M. et al. "Investigating the replicability of preclinical cancer biology". *eLife* 10:e71601, pp. 1-31, 2021a.

Errington, Timothy M. et al. "Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology". *eLife* 10:e67995, 2021b.

Fanelli, Daniele. "Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data". *PLoS One*, Vol. 5, No. 4, 2010.

Feest, Uljana. "Why replication is overrated". *Philosophy of Science*, Vol. 86, No. 5, pp. 895-905, 2019.

Fraser, Hannah et al. "Predicting reliability through structured expert elicitation with the repliCATS (Collaborative Assessments for Trustworthy Science) process". *PLoS ONE*, Vol. 18, No. 1, 2023.

Gabelica, Mirko et al. "Many researchers were not compliant with their published data sharing statement: a mixed-methods study". *Journal of Clinical Epidemiology*, Vol. 150, pp. 33-41, 2022.

Gibney, Elizabeth. "Is AI fuelling a reproducibility crisis in science?". *Nature*, Vol. 608, pp.250-251, 2022.

Gittel, Benjamin. "An institutional perspective on genres: generic subtitles in German literature from 1500-2020". *Journal of Cultural Analytics*, Vol. 6, No. 1, 2021.

Goodman, Steven N. et al. "Statistical Reviewing Policies of Medical Journals". *J Gen Intern Med*., Vol. 13, No. 11, 1998.

Goodman, Steven N. et al. "What does research reproducibility mean?". *Science Translational Medicine*, Vol. 8, No. 341, 2016.

Gordon, Michael, et al. "Predicting replicability - Analysis of survey and prediction market data from large-scale forecasting projects". *PLoS ONE*, Vol. 16, No. 4, 2021.

Gosselin, Daniel-Romain. "Insufficient transparency of statistical reporting in preclinical research: a scoping review". *Sci Rep* 11, 3335, 2021.

Han, SeungHye et al. "A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review". *PloS One*, Vol. 12, No. 9, 2017.

Hardwick, Tom E. and Goodman, Steven N. "How often do leading biomedical journals use statistical experts to evaluate statistical methods? The results of a survey". *PLoS One*, Vol. 15, No. 10, 2020.

Hardwicke, Tom E. et al. "An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017)". *Royal Society Open Science*, Vol. 7, No. 2, 2020.

Have, Iben and Enevoldsen, Kenneth. "From close listening to distant listening: Developing tools for Speech-Music discrimination of Danish music radio". *Digital Humanities Quarterly*, Vol. 15, No. 1, 2021.

Hermann, J. Berenike et al. "Response by the Special Interest Group on Digital Literary Stylistics to Nan Z. Da's". *Journal of Cultural Analytics*, Vol. 5, No. 1, 2020.

Hudson, Robert. "Explicating Exact versus Conceptual Replication". *Erkenntnis*, https://doi.org/10.1007/s10670-021-00464-z, 2021.

IJzerman, Hans. "Many Labs 5: Registered Replication of Förster, Liberman, and Kuschel's (2008) Study 1". *Advances in Methods and Practices in Psychological Science*, Vol. 3, No. 3, pp. 366-376, 2020.

Ioannidis, John P. A.. "Why most published research findings are false". *PLOS Medicine*, Vol. 19, No. 8, pp. 696-701, 2005.

Ioannidis, John P. A.. "How to Make More Published Research True". *PLoS Med*, Vol 11, No. 10, 2014.

Ioannidis, John P. A. et al. "Meta-research: Evaluation and Improvement of Research Methods and Practices". *PLoS Biol*, Vol. 13, No. 13, 2015.

Iqbal, Shareen A. et al. "Reproducible Research Practices and Transparency across the Biomedical Literature". *PLoS Biol*, Vol. 14, No. 1, 2016.

Jasny, Barbara R. et al. "Again, and Again, and Again...". *Science*, Vol. 334, No. 6060, p. 1225, 2011.

Landis, Story C. et al. "A call for transparent reporting to optimize the predictive value of preclinical research". *Nature*, Vol. 490, No. 7419, pp. 187-191, 2012.

Langer, Lars et al. "The Rise and Fall of Biodiversity in Literature: A Comprehensive Quantification of Historical Changes in the Use of Vernacular Labels for Biological Taxa in Western Creative Literature." *People and Nature*, vol. 3, no. 5, 2021.

Leslie, John K. et al. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling". *Psychological Science*, Vol. 23, No. 5, pp. 524-532, 2012.

Kapoor, Sayash and Narayanan, Arvind. "Leakage and the Reproducibility Crisis in ML-based Science". *arXiv*:2207.07048 [cs.LG], 2022.

Kaufman, Shachar et al. "Leakage in Data Mining: Formulation, Detection, and Avoidance". *ACM Transactions on Knowledge Discovery from Data*, Vol. 6, No. 4, pp. 1-21, 2012.

Kianersi, Sina et al. "Evaluating Implementation of the Transparency and Openness Promotion Guidelines: Reliability of Instruments to Assess Journal Policies, Procedures, and Practices". *Advances in Methods and Practices in Psychological Science*, Vol. 6, No. 1, 2023.

King, Gary. "Replication, Replication". *Political Science and Politics*, Vol. 28, pp. 444-452, 1995.

Klein, Martin et al. "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot". *PLoS One,* 9(12)*,* December 26, 2014.

Klein, Richard A. et al. "Investigating Variation in Replicability". *Social Psychology*, Vol. 45, No. 3, pp. 142-152, 2014.

Klein, Richard A. et al. "Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement". *Collabra: Psychology*, Vol. 8, No. 1, 2022.

Klein, Richard A. et al. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings". *Advances in Methods and Practices in Psychological Science*. Vol. 1, No. 4, pp. 443-490, 2018.

Koehler, Wallace. "Web Page Change and Persistence - A Four-Year Longitudinal Study". *Journal of the American Society for Information Science and Technology*, Vol. 53, No. 2, 2002.

Koole, Sander L. and Lakens Daniël. "Rewarding Replications: A Sure and Simple Way to Improve Psychological Science". *Perspectives on Psychological Science*, Vol. 7, No. 6, 2012.

Martin, Alison. "Hearing Change in the Chocolate City: Computational Methods for Listening to Gentrification". *Digital Humanities Quarterly*, Vol. 15, No. 1, 2021.

McNutt, Marcia. "Reproducibility". *Science*, Vol. 343, No. 6168, p. 229, 2014a.

McNutt, Marcia. "Journals unite for reproducibility". *Science*, Vol. 346, No. 6210, p. 679, 2014b.

Meindertsma, Peter. "Changes in Lyrical and Hit Diversity of Popular U.S. Songs 1956-2016". *Digital Humanities Quarterly*, vol. 13, no. 4, 2019.

Mlinarić, Ana et al. "Dealing with the positive publication bias: Why you should really publish your negative results". *Biochem Med (Zagreb)*, Vol.27, No. 3, 2017.

"More Responses to 'The Computational Case against Computational Literary Studies'". *Critical Inquiry*, https://critinq.wordpress.com/2019/04/12/more-responses-to-the-computational-case-against-computational-literary-studies/, accessed 25 April 2023.

Mulligan, John. "Computation and Interpretation in Literary Studies". *Critical Inquiry*, Vol. 48, No. 1, 2021.

Munafò, Marcus R. et al. "A manifesto for reproducible science". *Nat Hum Behav*, Vol. 1, 0021, 2017.

National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, 2019.

Nosek, Brian A. and Bar-Anan, Yoav. "Scientific Utopia: I. Opening Scientific Communication". *Psychology Inquiry*, Vol. 23, No. 3, 2012.

Nosek, Brian A et al. "Promoting an open research culture". *Science*, Vol. 348, No. 6242, pp. 1422-1425, 2015.

Nosek, Brian A. and Errington, Timothy M. "What is replication?". *PLoS Biology*, Vol. 18., No. 3, 2020.

Open Science Collaboration. "The Reproducibility Project A model of large-scale collaboration for empirical research on reproducibility". *Implementing Reproducible Research*, edited by Victoria Stodden, Friedrich Leisch and Roger D. Peng, CRC Press, pp.299-324, 2014.

Open Science Collaboration. "Estimating the reproducibility of psychological science". *Science*, Vol. 349, No. 6251, 2015.

Lupia, Arthur and Elman, Colin. "Openness in Political Science: Data Access and Research Transparency". *PS: Political Science and Politics*, Vol. 47, No. 1, pp. 19-42, 2014.

Pashler, Harold and Wagenmakers, Eric-Jan. "Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?". *Perspectives on Psychological Science*, Vol. 7, No. 6, 2012.

Patarčić, Inga and Stojanovski, Jadranka. "Adoption of Transparency and Openness Promotion (TOP) Guidelines across Journals". *Publications*, Vol. 10, No. 4:46, 2022.

Peng, Roger D. "Reproducible Research in Computational Science". *Science*, Vol 334, No. 6060, pp. 1226-1227, 2011.

Peng, Roger D. "The Reproducibility Crisis in Science: A Statistical Counterattack". *Significance*, Vol. 12, No. 3, pp. 30-32, 2015.

Peng, Roger D. and Hicks, Stephanie C. "Reproducible Research: A Retrospective". *arXiv*:2007.12210, 2020.

Pineau, Joelle et al. "Improving Reproducibility in Machine Learning Research". arXiv:2003.12206v4, 2020.

Piper, Andrew. *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data (Elements in Digital Literary Studies)*. Cambridge University Press, pp. 1-88 , 2020a.

Piper, Andrew. "Do we know what we are doing?". *Journal of Cultural Analytics*, Vol. 5, No. 1, 2020b.

Piper, Andrew. "Biodiversity is not declining in fiction". *Journal of Cultural Analytics*, Vol. 7, No. 3, 2022.

Popper, Karl. *The Logic of Scientific Discovery*. 2nd ed., Routledge, 2002.

"Principles and Guidelines for Reporting Preclinical Research". *NIH Grants & Funding*, https://grants.nih.gov/policy/reproducibility/principles-guidelines-reporting-preclinical-research.htm. Accessed 29 April 2023

Raghupathi, Wullianallur et al. "Reproducibility in Computing Research: An Empirical Study". *IEEE Access*, Vol. 10, pp. 29207-29223, 2022.

"Replication Studies". *NWO*, https://www.nwo.nl/en/researchprogrammes/replication-studies, accessed 16 April 2023.

"Reporting standards and availability of data, materials, code and protocols". *Nature Portfolio*, https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards, accessed 29 April 2023.

Rizvi, Pervez. "The use of the t-test in Shakespeare scholarship". *Digital Scholarship in the Humanities*, Vol. 36, No. 3, 2021.

Rizvi, Pervez. "An analysis of the Word Adjacency Network method—Part 1—The evidence of its unsoundness". *Digital Scholarship in the Humanities*, Vol. 38. No. 1, 2023a.

Rizvi, Pervez. "An analysis of the Word Adjacency Network method—Part 2—A true understanding of the method". *Digital Scholarship in the Humanities*, Vol. 38. No. 1, 2023b.

Rodgers, Peter and Collings, Andy. "Reproducibility in Cancer Biology: What have we learned?" *eLife* 10:e75830, 2021.

Romero, Felipe. "Philosophy of science and the replicability crisis". *Philosophy Compass*, Vol. 14, No. 11, 2019.

Schmidt, Stefan. "Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences". *Review of General Psychology*, Vol. 13, No. 2, pp. 90-100, 2009.

Serra-Garcia, Marta and Gneezy, Uri. "Nonreplicable publications are cited more than replicable ones". *Science Advances*, Vol. 7, No. 21, 2021.

Sterne, Jonathan A. C. et al. "Sifting the evidence—what's wrong with significance tests? Another comment on the role of statistical methods". *BMJ*, Vol. 322, No. 7280, pp. 226-231, 2001.

Simons, Daniel J. "The value of direct replication". *Perspectives on Psychological Science*, Vol. 9, No. 1, pp. 76-80, 2014.

Stevenson, Ryan et al. "Characterization of the Affective Norms for English Words by discrete emotional categories". *Behavior Research Methods*, Vol. 39, pp. 1020-1024, 2007.

Stodden, Victoria et al. "An empirical analysis of journal policy effectiveness for computational reproducibility". *PNAS*, Vol. 115, No. 11, pp. 2584-259, 2018.

Stroebe, Wolfgang and Strack, Fritz. "The alleged crisis and the illusion of exact replication". *Perspectives on Psychological Science*, Vol. 9, No. 1, pp. 59-71, 2013.

"The history of 78 RPM recordings". Yale University Library, https://web.library.yale.edu/cataloging/music/historyof78rpms, accessed 03 June 2023.

The NPQIP Collaborative group. "Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution". *bioRxiv* 187245, 2017.

Thompson. Tyechia L. and Carrera, Dashiel. “Afrofuturist Intellectual Mixtapes: A Classroom Case Study”. *Digital Humanities Quarterly*, Vol. 15, No. 1, 2021.

Tiefenbach Keller, Stephen. *Replication code and data for: "Towards a culture of replication in the Digital Humanities: an understanding of transparency and openness practices in published DH papers*. 2023, https://doi.org/10.5683/SP3/KRPCBL, Borealis, V1, UNF:6:xlZR+T0jMhFCAgbrVJHqWA==

Trafimow, David and Marks, Michael. "Editorial". *Basic and Applied Social Psychology*, Vol. 37, No. 1, 2015.

UNESCO. "UNESCO Recommendation on Open Science". *United Nations Educational, Scientific and Cultural Organization*, 2021.

Vazire, Simine. "Editorial". *Social Psychology and Personality Science*, Vol. 7, No. 1, 2015.

Vazire, Simine and Holcombe, Alex O. "Where are self-correcting mechanisms in science?". *Review of General Psychology*, Vol. 0, No. 9, pp. 1-12, 2021.

Wallach, Joshua D. "Reproducible research practices, transparency, and open access data in the biomedical literature, 2015-2017". *PLoS Biol*, Vol. 16, No. 11, 2018.

Wasserstein, Ronald L. "ASA Statement on Statistical Significance and P-Values". *The American Statistician*, Vol. 70, No. 2, 2016.

Willinsky, John. *The Access Principle: The Case for Open Access to Research and Scholarship*. The MIT Press, 2009.

Willianson, Eric. "After 10 Years, ‘Many Labs’ Comes to an End – But Its Success Is Replicable". *UVAToday*, May 23, 2022, https://news.virginia.edu/content/after-10-years-many-labs-comes-end-its-success-replicable

Appendix A

A list of code snippets with the main line of codes required to execute the data analysis performed in this thesis is provided below. All code is written in Python 3 language. Full code can be accessed via the McGill University Dataverse repository (Tiefenbach Keller, 2023).

Calculate word counts (types, tokens, TTR and types, tokens and TTR per second). Note that all lists in the code below need to be declared:

```
import nltk as nlp
from nltk.corpus import stopwords

# Tokenize
lyrics = "Text goes here"
lyrics = lyrics.lower()
tokens = nlp.word_tokenize(lyrics)
types = nlp.Counter(tokens)

# Remove stopwords
tokens_noStopwords = [word for word in tokens if not word in stopwords.words()]
types_noStopwords = nlp.Counter(tokens_noStopwords)

wordCount.append(len(tokens))
wordCount_noStopwords.append(len(tokens_noStopwords))
uniqueWordCount.append(len(types))
uniqueWordCount_noStopwords.append(len(types_noStopwords))
typeTokenRatio.append(len(types)/len(tokens)*100)
if len(types_noStopwords) > 0 and len(tokens_noStopwords) > 0:
        typeTokenRatio_noStopwords.append(
                        len(types_noStopwords)/len(tokens_noStopwords)*100)
else:
typeTokenRatio_noStopwords.append(0)

songLength.append(float("Song lenght goes here"))
words_perSecond = [value/songLength[index] for index,value in enumerate(wordCount)]
words_perSecond_noStopwords = [value/songLength[index] for index,value in
```

```
                                        enumerate(wordCount_noStopwords)]
UniqueWords_perSecond = [value/songLength[index] for index,value in
                                        enumerate(uniqueWordCount)]
UniqueWords_perSecond_noStopwords = [value/songLength[index] for index,value in
                                            enumerate(uniqueWordCount_noStopwords)]
TTR_perSecond = [value/songLength[index] for index,value in enumerate(typeTokenRatio)]
TTR_perSecond_noStopwords = [value/songLength[index] for index,value in
                                        enumerate(typeTokenRatio_noStopwords)]
```

Calculate ANEW scores for lyrics and save ANEW sums and averages to CSV file:

```
# Create ANEW dictionary
anew = {}
with open('anewList.csv', 'r', encoding="ISO-8859-1") as csvFile:
reader = csv.reader(csvFile)
next(reader)
for row in reader:
# 0 = Word// 1 = Valence// 2 = Arousal// 3 = Dominance// 4 = Happiness// 5 =
        Anger// 6 = Sadness// 7 = Fear// 8 = Disgust
anew[row[0]] = [row[1], row[2], row[3], row[4], row[5], row[6], row[7],
        row[8],]

# Count number of ANEW words present in a text and add up their attributes
wordCounter = 1
valence, arousal, dominance = 0,0,0
happiness, anger, sadness, fear, disgust = 0,0,0,0,0
words = nlp.word_tokenize(row[2])
for word in words:
if anew.get(word) != None:
wordCounter = wordCounter + 1
valence += float(anew.get(word)[0])
arousal += float(anew.get(word)[1])
dominance += float(anew.get(word)[2])
happiness += float(anew.get(word)[3])
anger += float(anew.get(word)[4])
sadness += float(anew.get(word)[5])
fear += float(anew.get(word)[6])
disgust += float(anew.get(word)[7])
if wordCounter == 0:
wordCounter = 1
```

```
# Save sums to CSV and averages by dividing the sum by total number of words
with open('sentimentANEW.csv', 'a', newline='') as wordsCSVfile:
        write = csv.writer(wordsCSVfile)
write.writerow([row[0], row[1], wordCounter, valence, valence/wordCounter, arousal,
arousal/wordCounter, dominance, dominance/wordCounter, happiness,happiness/wordCounter,
anger, anger/wordCounter, sadness,sadness/wordCounter, fear, fear/wordCounter, disgust,
disgust/wordCounter])
```

Calculate TTR by randomly sampling words from a text and append the result to a list:

```
from random import sample
sampleSize = 75
sampleTimes = 50
TTR = []
years =[]

tokens = nlp.word_tokenize(row[2])
TTRSampled =[]
if len(tokens) > sampleSize:
for i in range(sampleTimes):
tokensSample = sample(tokens,sampleSize)
types = nlp.Counter(tokensSample)
TTRSampled.append(len(types)/len(tokensSample)*100)
years.append(row[0])
TTR.append(sum(TTRSampled)/sampleTimes)
else:
pass
```

Calculate hapax count and store it into csv file:

```
from nltk.probability import FreqDist
tokens = nlp.word_tokenize("text goes here")
fdist = FreqDist(tokens)
with open('hapaxList.csv', 'a', newline='') as wordsCSVfile:
        write = csv.writer(wordsCSVfile)
        write.writerow([len(fdist.hapaxes()), fdist.hapaxes()])
```

Calculate Dale-Chall score and store it into csv file:

```
# https://pypi.org/project/py-readability-metrics/#dale-chall-readability
from readability import Readability
r = Readability("text goes here")
dc = r.dale_chall()

# Save Dale Chall readability scores to CSV
with open('lexicalReadability.csv', 'a', newline='') as wordsCSVfile:
write = csv.writer(wordsCSVfile)
write.writerow([dc.score, dc.grade_levels])
```

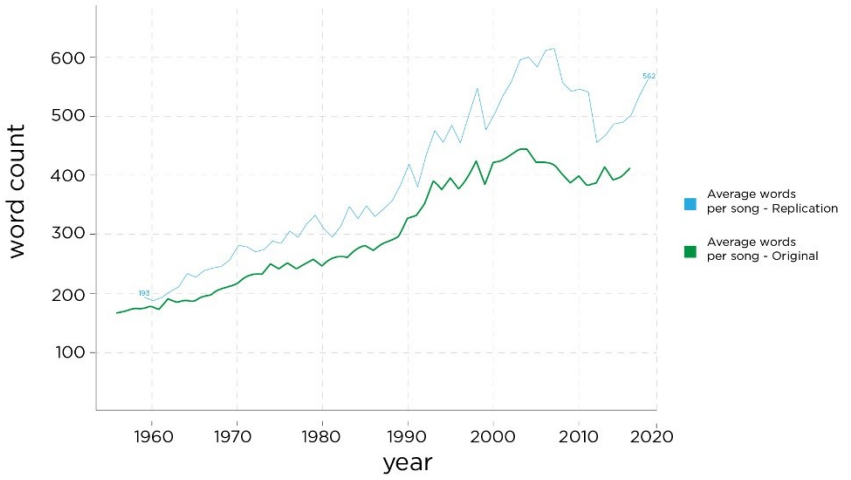
Calculate k means for the lyrics word counts, plot the results and save cluster labels to csv:

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# Create pandas dataframe with word counts
df = pd.read_csv('wordCountsNLTK.csv', usecols = ['Year','Chart', "Word count", "Unique Word
Count"])
df1 = df[['Word count', 'Unique Word Count']]

# Calculate k means
kmeans = KMeans(n_clusters=5)
clusters = kmeans.fit(df1)
df['Cluster'] = clusters.labels_.tolist()
df.to_csv("kCluster.csv")

# Visualize k clusters
plt.scatter(df["Word count"], df["Unique Word Count"], s=1, c=kmeans.labels_)
plt.show()
```

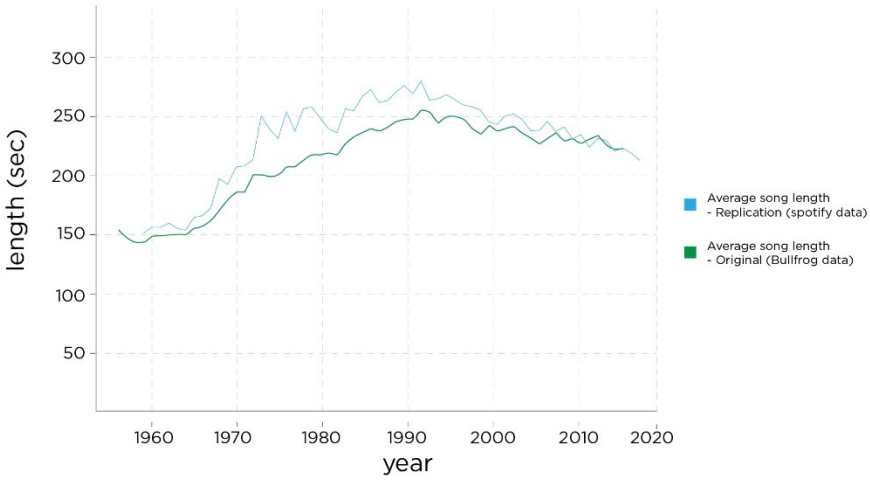
Appendix B

This appendix contains graphs that were not included in the results section of the replication of Meindertsma's study. To produce the images below, the graphs generated with the Python scripts described in Section 4.1 were saved in a vector format (.eps). This format allows the graph to be opened and its contents individually manipulated in Adobe Illustrator. Due to the mathematical nature of the .eps format, the lines and elements on each graph can be scaled and manipulated without any loss of information (as opposed to bitmap files such as .jpg). These graphs were then overlaid on top of the equivalent graphs produced by Meinderstma. Using the X and Y scales as reference, the scales were aligned and matched (e.g., The Y value of 100 was placed directly on top of the original study's Y value of 100; X values representing the years were also aligned and matched). This manual process leads to the graph lines of both the original and replication results being placed roughly in the exact position they would be had they been both generated mathematically, but the manual nature of the process cannot guarantee a pixel-perfect result. As the images below show, though, the lines can be used to visually compare the results and check if the results follow similar trends or not. Research in this space, or a literature review on Data Visualization, that was outside of the scope of this project, can help inform how much of this process can be used for replication efforts where the replication researcher does not have access to the necessary data and code for producing graphs to compare both results.

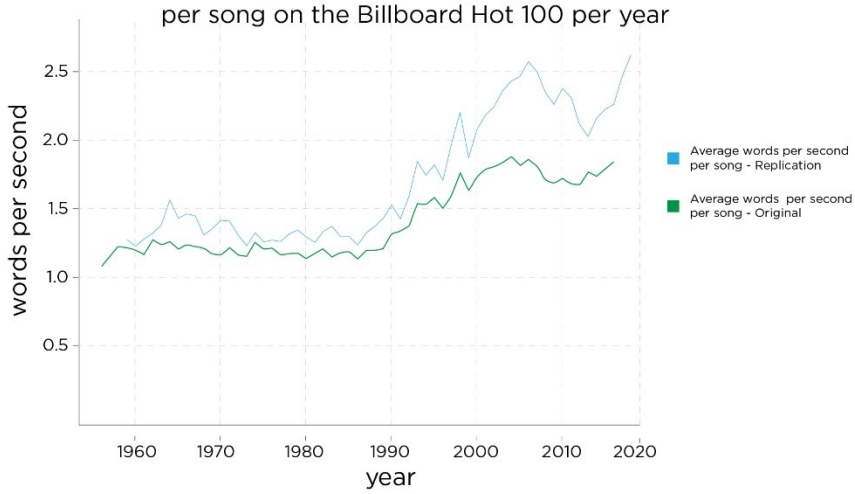
Average number of words per song on the Billboard Hot 100 per year



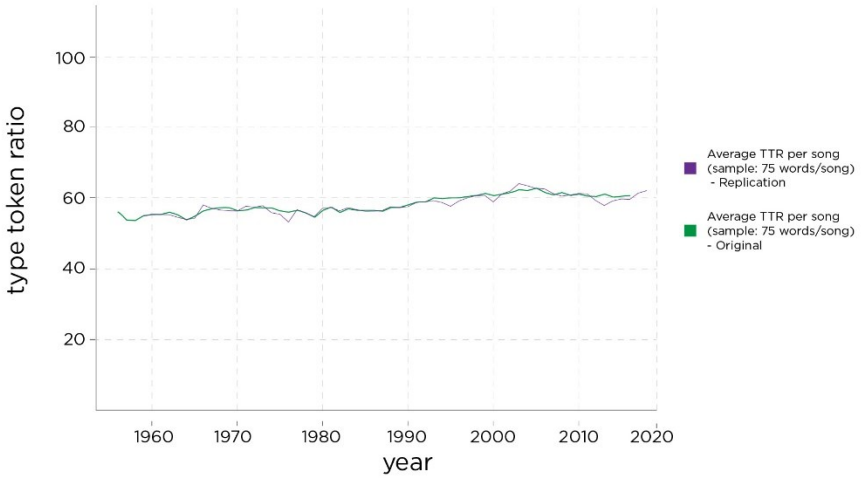
Average length in seconds of songs on the Billboard Hot 100 per year



Average number of words per second per song on the Billboard Hot 100 per year



Average type token ratio per song on the Billboard Hot 100 per year



Average valence scores (ANEW) of songs on the Billboard Hot 100 per year