DH Benelux Journal

Volume 4:
The Humanities in a Digital World

Summer 2022

General Editors
Wout Dillen, Margherita Fantoli, Marijn Koolen, and Marieke van Erp

Guest Editors
Lorella Viola, Jelena Prokic, Antske Fokkens and Tommaso Caselli

digital humanities Benelux Journal

DH Benelux Journal 4. The Humanities in a Digital World.

ISSN: 2666-6952

2022

This journal, including all its contents, is licensed under a Creative Commons Attribution 4.0 International license, and made available in Open Access at https://journal.dhbenelux.org. The authors of the individual contributions, who are identified as such, retain the copyright over their original work.

For more information on the CC BY 4.0 license, please refer to:
https://creativecommons.org/licenses/by/4.0/deed.en.

This journal was typeset in LaTeX by Marijn Koolen using *varianTeX* — a reusable template for journals in the Humanities, developed by Wout Dillen. *varianTeX* is open source, available on GitHub, and deposited in the Zenodo Open Science Repository. DOI: 10.5281/zenodo.3484651.

Contents

Editors' Preface i

Introduction: DH Benelux Online iii

Essays

A Game of Persistence, Self-doubt, and Curiosity: Surveying Code Literacy in Digital Humanities
Elli Bleeker, Marijn Koolen, Kaspar Beelen, Liliana Melgar, Joris van Zundert, and Sally Chambers 1

Introducing the DHARPA Project: An Interdisciplinary Lab to Enable Critical DH Practice
Angela R. Cunningham, Helena Jaskov, Sean Takats, and Lorella Viola; with contributions from Markus Binsteiner and Mariella de Crouy-Chanel 29

Examining a Multi Layered Approach for Classification of OCR Quality without Ground Truth
Mirjam Cuper 43

Modeling Ontologies for Individual Artists: A Case Study of a Dutch Ceramic Glass Sculptor
Victor de Boer, Daan Raven, Erik Esmeijer, and Johan Oomen 61

Judging a Book by its Criticism: A Digital Analysis of the Professional and Community Driven Literary Criticism of the Ingeborg-Bachmann-Preis
Lore De Greve and Gunther Martens 79

When No News is Bad News. News-Based Change Detection during COVID-19
Kristoffer L. Nielbo, Frida Hæstrup, Kenneth C. Enevoldsen, Peter B. Vahlstrup, Rebekah B. Baglini, and Andreas Roepstorff 107

Combining Tools with Linked Data: a Social History Example
Ivo Zandhuis 121

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

Editors' Preface

As the general editors of the DH Benelux Journal, we are proud to present our trilateral Digital Humanities research community with the fourth volume of this Open Access journal. As in previous years, authors of accepted conference abstracts were invited to submit full versions of their papers, which were then subjected to a stringent peer reviewing process. The resulting volume includes research presented at the eighth DH Benelux conference. The theme of the conference was 'The Humanities in a Digital World,' and took place on 2-4 June 2021. In this volume's Introduction by our guest editors Lorella Viola, Jelena Prokic, Antske Fokkens and Tommaso Caselli, this theme and the individual contributions in the volume are described in more detail.

As our guest editor also explain in their Introduction, COVID-19 obliged DH Benelux 2021 to become our second DH Benelux conference to take place fully online – despite the organisers' significant efforts. We know that organizing a conference like DH Benelux is a great and challenging undertaking in itself, and doubly so when unforeseen circumstances necessitate organizers to pivot quickly and change direction. That is why we would also like to give special thanks to the local organizers for all the hard work they did to make DH Benelux 2021 possible. Thankfully, they could count on an enthusiastic, versatile, and understanding community that took these changes in stride, determined to help make DH Benelux 2021 another great success. Last but not least, we would also like to thank our guest editors – delegates from the programme and organizing committees of DH Benelux 2021 – for all their help putting this volume together!

As we are currently receiving and processing submissions for the fifth volume of our Journal (based on the presentations of the 2022 conference in Luxembourg), and look forward to our next gathering in Brussels, we hope you enjoy reading this volume's fascinating contributions – and that they may inspire lively academic discussion in the DH Benelux community, and beyond!

July 15, 2022
Amsterdam, Borås, and Leuven

Wout Dillen
Margherita Fantoli
Marijn Koolen
Marieke van Erp

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

Introduction: The Humanities in a Digital World

Lorella Viola[1], Jelena Prokic[2], Antske Fokkens[3], and Tommaso Caselli[4]

[1]University of Luxembourg
[2]Leiden University
[3]Vrije Universiteit Amsterdam & Eindhoven University of Technology
[4]Rijksuniversiteit Groningen

DH Benelux 2021 marks the eighth edition of the annual conference of Digital Humanities (DH) in the Benelux community. The conference started with the hope of being a hybrid event, although with limited participation onsite. At the end of November 2020 the first authorization requests for the approval of vaccines for COVID-19 had been submitted, and massive vaccination campaigns started in the EU in January 2021. Unfortunately, it soon became rather clear that, by the time of the conference (early June), the number of potential participants who had received their first dose would not be sufficient. Moreover, with travel restrictions mostly still in place, hosting an onsite conference was simply not feasible. Learning from the experience of the first wave of COVID-19, we therefore had no choice but to move the conference fully online.

After only one year in the pandemic, making sure to have a smooth online conference was still challenging. In coordination with the local organizers, we discussed and evaluated different solutions concerning the presentations and the online environments. We finally decided to keep the traditional distinction between oral and poster presentations and to have social events too.

We are very pleased to say that the conference was a very successful event, hosting two panel sessions and a total of 42 abstracts. In the tradition of DH Benelux conferences, after the event we invited authors to submit extended versions of their abstracts to the Special Issue of the DH Benelux journal. We received nine submissions, each reviewed by experts in a single-blind setup. We accepted seven papers after minor or major revisions. The results of this process can be found in this issue.

As usual, the contributions of this issue reflect a mix of articles, some directly addressing the theme of DH Benelux 2021 *The Humanities in a Digital World*, others presenting case studies within the broader spectrum of Digital Humanities.

The paper by Bleeker and colleagues "A Game of Persistence, Self-doubt, and Curiosity: Surveying Code Literacy in Digital Humanities" targets the thematic strand of *DH Practices and Didactics* from a new perspective. They investigate what *code literacy* means in the humanities. The paper describes the outcome of a survey that asks

members of the DH community how they define code literacy and how important it is for them. Using 399 definitions from their survey, they show that code literacy is a complex concepts where different kind of skills (reading, interpreting, writing, using, publishing, reviewing code), as well as the research or social context play a role.

The *DH Practices and Didactics* theme is also represented by the article by Cunning and colleagues "Introducing the DHARPA Project: An interdisciplinary lab to enable critical DH practice". In this contribution the authors introduce a Virtual Research Environment (VRE) software under development by the Digital History Advanced Research Projects Accelerator (DHARPA), an interdisciplinary team of researchers and developers. The presented VRE comprises of a modular and data-centric backend and an interactive frontend capable of documenting scholarly processes. The main goal of the VRE is to "(re)build a critically-aware connection between scholar and sources in a digital world".

Connected to the theme of *Data for a Digital Age*, the contribution by Zandhuis "Combining Tools with Linked Data: a social history example" highlights the challenge of working with a variety of tool sets. The author explains how different tasks in a typical research setting each involve their own (preferred) tool. The author describes the process of combining references stored in Zotero, photos from Tropy and digitized text from Recogito using Linked Data. The result is a highly accessible introduction to Linked Data and how this may be used, illustrated through a use case around 19th century print laborers.

The article by De Boer and colleagues, "Modeling Ontologies for Individual Artists: A Case Study of a Dutch Ceramic Glass Sculptor" focuses on the theme *Creativity through a Digital Lens*. The paper discusses the challenges and advantages related to the use of ontologies as a solution to better structure and enrich data for Cultural Heritage databases facilitating "both experts' as well as the general public's access to information". They present a dedicated case study related to an ontology for a ceramic glass artist illustrating the methodology used to develop and evaluate it.

In the paper by Nielbo and colleagues "When no news is bad news - News-based change detection during COVID-19", the authors illustrate the transformations of news media outlets into "*Corona-news*" where "the same news was repeated over and over" decreasing the novelty of the information presented to the readers. By using a Bayesian approach, the authors investigates whether a "change in novelty can be used to detect change in news media" when societies are affected by major negative events.

De Greve and colleagues' article "Judging a Book by its Criticism: A Digital Analysis of the Professional and Community Driven Literary Criticism of the Ingeborg-Bachmann-Preis" offers a peek into a relationship between literary prizes and social media by examining the Ingeborg-Bachmann-Preis. The authors provide an in-depth qualitative and quantitative analysis of differences between the Twitter, Instagram, Goodreads and jury corpora and between the evaluative literary criteria of professional and layperson critics surrounding the German Ingeborg-Bachmann-Preis.

Lastly, Cuper's paper "Examining a multi layered approach for classification of OCR quality without Ground Truth" proposes and describes a multi-layered approach for the classification of OCR quality called QuPipe. The contribution of the method, which is tested and evaluated on a data-set of sentences from 17th Century Newspapers, lies in its potential to be used without Ground Truth for OCR quality classification.

The last words of this editorial are for the Leiden local organizers: Angus Mol, Sjef Barbiers, Alison Carter, Jelena Prokic, Laurents Sesink, and Erik Weber. They have all played a key role in making DH Benelux 2021 possible. We also would like to

thank the editors of this journal for their support in the process of making this volume possible. A special thank goes to the reviewers: they have provided thorough and constructive comments to the submissions both for the conference programme and the Special Issue, crucially contributing to knowledge advancement. Last but not least, we want to thank the whole DH Benelux community: your participation and engagement were vital to the success of DH Benelux 2021.

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

A Game of Persistence, Self-doubt, and Curiosity: Surveying Code Literacy in Digital Humanities

Elli Bleeker[1], Marijn Koolen[1], Kaspar Beelen[3], Liliana Melgar[4], Joris van Zundert[1], and Sally Chambers[2]

[1]Huygens Institute for Dutch History and Culture
[2]Ghent University
[3]Alan Turing Institute
[4]Utrecht University

The interpretation of "code" and its role in the digital humanities has been a topic of debate ever since the first experiments in "humanities computing" in the late 1950s. Even some years ago, the question whether humanities students need to know how to code was sincerely provocative. At present digital humanists seem to agree that knowing how to code is relevant, if not essential for digital humanities research. However, there is a lack of agreement on what the community means by "code literacy", and as a result the efforts to improve code literacy among students and researchers are dispersed. This paper presents the, to our knowledge, largest ever survey on code literacy and related questions in the field. We expound the results and analysis of the first two overarching research questions of the four that the survey aimed to tackle: 1) What are the definitions and interpretations of code literacy across humanities disciplines?; 2) How important is code literacy as part of digital humanities scholarship? We explain what according to the survey answers are the most important elements or dimensions of a definition of "code literacy". Among these are the different levels of competence and the varying importance of code literacy across disciplines. The paper concludes by discussing the implications of these findings, opening up questions for the debate on academic curricula regarding digital humanities.

Keywords: code literacy; pedagogy; survey; computational thinking; coding skills; humanities computing/programming; concept analysis

1 Introduction

"Should humanists learn to code?" Less than a decade ago this question would have ignited quite a controversy in the field of digital humanities (DH). Today, the consensus

is that a certain level of code literacy is preferred.[1] Instead of arguing whether code literacy deserves to be part of DH's skill set, the debate has moved on to discussing what it means, exactly, to be code literate. How we define code literacy appears a Rorschach test: our definition depends on "our values, our experiences, our skills, our visions of the future as it pertains to technology, to computers, to communications, to information" (Beverly Hunter, cited in Dobberstein 1993, 430).

The somewhat confused state of definitions and discourse is not much helped by a proliferation of adjacent, overlapping, and vaguely defined literacies such as "computer literacy" (Dobberstein 1993, Tafazoli et al. 2017), "digital literacy" (Spante et al., 2018), "media literacy" (Potter, 2010), and so on. After mapping the use of the term "computer literacy" between 1965 and 1985, Kenneth King concluded half-jokingly, that this type of literacy is "widely regarded as an impossible term to define; however, whatever it is, it is important for every student to have it" (King, 1985, 20). Even those opposing the promotion of computer literacy, such as Douglas Noble – "What has apparently convinced an entire population that something as vague and worthless as computer literacy is essential to their lives?" – acknowledge that the community has "unusual difficulty arriving at a suitable definition."(Noble, 1984, 602,607) It does not help that a definition is inevitably influenced by the technological developments of the time.

Over the years, various researchers engaged with the topic of code literacy, presenting valuable insights. Their observations remained, however, limited to their own experiences with teaching digital humanities courses or with working in a digital humanities context (e.g., Van Zundert, Antonijević, and Andrews 2020). During a recent round table at the DH Benelux 2019 conference in which we explored the perception of code literacy in the field, the participants expressed smart opinions without providing much empirical evidence (Melgar et al., 2019). The animated discussion was largely informed by anecdotal evidence, repeating phrases such as "in my experience, ...", "in my department, ...“). To establish an informed and evidence-based definition of code literacy in the humanities we decided to ask the community for their opinion by means of a survey.

In this paper, we report on the results of the questionnaire answers focusing, first, on the present-day definitions and interpretations of code literacy across humanities disciplines and, secondly, on the importance of code literacy as a part of DH scholarship. The survey used a questionnaire that was widely distributed and received a large number of responses from a demographically diverse audience. As a result, the results allow us to analyse how factors like background, education, and experience shape opinions about code literacy. Instead of providing a clear-cut definition of "code literacy" – which would suggest a unity of perspective that in reality does not exist – this paper highlights a number of aspects that are considered important across the DH community. Accordingly, the contribution of this paper lies in a community-inferred vocabulary and a practical framework for the discussion, evaluation, teaching, and promotion of code literacy in digital humanities.

[1] Callaway et al. 2020 offer an overview of the topics discussed within the DH community over the past ten years; specifically paragraphs 3, 6, and 13 on the topic of "code". A Google Scholar query for "code", "coding", and "literacy" shows that the discussion is ongoing and anything but limited to the humanities domain.

2 Background

The interpretation of "code" and its role in the digital humanities has been a topic of discussion ever since the first experiments in "humanities computing" in the late 1950s. Few studies have, however, reviewed the history of this notion in a systematic way. This is surprising given that each generation of scholars was required to learn a different set of computer skills, so that perceptions of code (literacy) changed over the years.[2] Since the discussion on code literacy (in the humanities) is distributed over a wide range of sources, reconstructing its history proved to be a difficult task. Academic publications contain various traces and clues. More recent debates tend to take place online, on platforms like Twitter, or blog posts and DH forums.[3] In order to contextualise the findings of our survey, this section synthesises the historical discourse on code literacy on the basis of four shared themes:

1. The perceived role and potentially divisional nature of programming;
2. The importance of context for (teaching) code literacy.
3. The influence of code literacy on humanist thinking;
4. A distinction between multiple levels of literacy;

The literature used in this section is selected following a combination of purposive (or judgement) sampling and snowballing (or citation tracking) (Wildemuth, 2009). Starting from several seminal works on the topic of DH and code literacy, such as Gold (2012a) and Vanhoutte, Nyhan, and Terras (2013), we pursued relevant citations in their bibliographies. And finally, we carried out several search queries via Google Scholar and the JSTOR archive with various combinations of the keywords "code", "coding", "computer", "programming", "digital humanities", "humanities", "electronic", "literacy", and "literate". Although we will not claim that our literature review is exhaustive, we trust that it is representative. A final note on terminology: historical sources often refer to "computer literacy" rather than "code literacy'. For consistency we will use the term "code literacy" throughout.

2.1 In or Out: the Programming Divide

A major theme in the discourse on code literacy is the question of programming. The Oxford English Dictionary (OED) defines "literacy" – rather narrowly – as "the ability to read and write". If applied to code literacy, then, being code literate can be taken to mean "the ability to read and write code". Incidentally, "code" is already an ambiguous term in the humanities, because it can refer to text encoding with XML

[2] For example, being code literate had a different meaning in the 1960s, when few scholars owned a (in the parlance of the time) microcomputer, compared to the early 1980s, when the development of personal computers with a graphical user interface (GUI) took flight, and in the 1990s during the rise of the world wide web. Interestingly, educators have recently reported that students in STEM disciplines are no longer familiar with the directory structure, saving thousands of files on their desktop instead. This development could be explained by the omnipresence of the search function, which no longer requires users to know where their files are located (Chin, 2021). Although we will touch upon it briefly, the educational aspects of code literacy are not within the scope of the present article.

[3] As Gold notes, "some conversations, especially those on Twitter – a platform used extensively by digital humanists – are hopelessly dispersed and sometimes even impossible to reconstitute only a few months after they have taken place" (Gold, 2012b).

or HTML as well as using programming languages to write computer programs.[4] The question whether programming skills are a prerequisite for digital humanists has divided the field from the 50s until today (see Dobberstein 1993; van Zundert and Andrews 2017, ii86; Ide 1987).[5]

Especially when postulated as essential, proficiency in programming languages risks becoming a harmful criterion for exclusion. In the words of Douglas Rushkoff, "program or be programmed" Rushkoff. Earhart et al., for instance, notes an "increasing tension" between DH practitioners and DH theorists. As she sees it, both sides look down on the other: code literate scholars spurn those who are not (Ramsay, 2012), while simultaneously "builders" are not considered real digital humanists because "they haven't theorized their work within the context of humanities and technology" (Earhart et al., 2016). Annette Vee, too, points to the intricate tensions that hide behind the deceptively obvious compound formed by code and literacy. Coming from the sociological oriented New Literacy approach, she compares the historical development of textual literacy to the ideological push for "mass programming movements" (Vee, 2017, 152) and thus explains how a particular set of skills or capabilities can become a much coveted "literacy". Depending on one's perspective such a literacy may present itself as an individually empowering advantage or rather as a power game that excludes specific groups of people.

In 2015, O'Sullivan, Jakacki, and Galvin conducted a survey to find out whether programming is indeed perceived as a barrier to join the DH community. The trigger for the survey, according to O'Sullivan et al., was the essay "Who's in and Who's Out?" (Ramsay, 2013b), in which Ramsay famously declared that digital humanists should learn to code if they were to be part of the field. Although this essay caused quite a backlash and Ramsay later nuanced his statement, the sentiment appears to be persistent. The survey, which was completed by 96 respondents in the field, set out to study the "relevant attitudes in relation to [software] development within Digital Humanities projects" (O'Sullivan et al., 2015, 142). Notably—and against the authors' expectations—it was found that young scholars adhere less importance to programming than older generations: 40% of the respondents in the 25-35 age group considered programming as essential to DH work, in contrast to 60% in the category of 50 years or older. Senior scholars usually claimed to do the programming themselves, whereas younger scholars either collaborated with skilled developers or delegated coding to colleagues. Younger respondents also considered themselves less "technically proficient" (O'Sullivan et al., 2015, 144).

More recently, Callaway, Turner, Stone, and Halstrom (2020) examined "the gate keeping around coding and technical skills more broadly" by topic-modelling a corpus of 334 (English) definitions of DH found in readers and companions on the topic. In a way, their approach can be considered an indirect survey of the field. The authors found "a range of different views towards coding" (Callaway et al., 2020, §13). Few contributions took up a "hard stance" (e.g., humanists *should* learn how to code); most discussed whether programming skills formed a threshold for participation in DH and

[4] In a survey among 96 digital humanists, O'Sullivan, Jakacki, and Galvin found considerable disagreement on the definition of programming: "many respondents mention text encoding, particularly XML and HTML, as opposed to more sophisticated dynamic programming languages" (O'Sullivan et al., 2015, 145).

[5] It is worth pointing out that the term code literacy (or any of its related terms) rarely if ever appears in the computer science literature from this period. The explanation for this could be quite simply that "computer scientists have no more reason to refer to each other as 'computer literate' than say, physicists have in referring to their peers as 'science literate' " (Dobberstein, 1993, 430).

if so, how to overcome related hurdles. Interestingly, the authors noted that—being early-career academics themselves—interpreting the topics proved more challenging (and rewarding) than understanding the technical aspects (of topic modelling). This led them to conclude that "less emphasis should be placed on digital competencies and more emphasis on the step of interpretation that comes after the use of the digital tool" (Callaway et al., 2020, 28).

2.2 The Computer in Context

Indeed, most curricula concentrating on humanities and computing have taken code literacy to be more a broad than a narrow concept which usually comprises more than just programming skills. Early on, Zemsky contended that "we [historians] must ultimately invent a methodology—including computer programs—of our own, a methodology designed to cope with the peculiar kinds of evidence with which we deal" (Zemsky, 1969, 39). During a three-day workshop *Teaching Computers and the Humanities Course* (see also Hockey 1986 and Tannenbaum 1987), all participants—educators in humanities computing—agreed that when teaching computer skills "the example data must be from humanities disciplines" (Hockey, 1986, 228). And, Ide adds, they generally agreed that the computational methodologies need to be contextualised by traditional methodologies from the relevant humanities disciplines (Ide, 1987, 212).

In "Technology and the Historian" Crymble discusses the changing role of technology in the history classroom. Taking a longitudinal perspective—broadly ranging from 1980s to the 2010s—he describes the different ways digital technologies have been embedded in the historical curriculum: the initial emphasis on coding and counting (i.e. statistical analysis) in 1980s, was replaced with a focus on using digital, web-based technologies for public history, making historical narratives and sources accessible to a wider audience. Together with the rise of digital humanities in the 2000s, historians increasingly perceived digital methods as "tools". Historians were less concerned with teaching students how to code than to apply digital tools to historical data. The interest for digital tools was often applied and instrumental, ignoring the mathematics that underpinned many of methods and technologies used in digital history. Moreover, Crymble (2021) argues that "There was no overarching plan for a digital transformation of the historical classroom. Instead, it was a reactive intellectual space led by a few passionate individuals and largely ignored by the rest of the profession. These passionate few were not pushing a coherent agenda."

Reflecting on fifteen years of teaching computing to humanities students, Oakman found that mismatching "technique and subject" can result in outright "boring" courses for humanities students (Oakman, 1987, 232). To prevent this, he argues, humanists need to learn coding in combination with subjects of their interest. Furthermore, Oakman states that being code literate should include a knowledge of "the positive and negative effects of the Computer Revolution on the modern world (computers and society issues: privacy and government datafiles, automation and unemployment, robotics, technophobia, etc.)" (Oakman, 1987, 229). This definition foreshadows the current increase in interest for the sociological, political, and cultural effects of code on society, such as the emergence of the field of critical code studies (Marino, 2006), or the most recent volume of *Debates in the Digital Humanities* (Gold and Klein, 2019).

More recent sources also emphasise the importance of situating code literacy in its historical and social context (Vee, 2013, 43), and "to provide practical technical skills within a humanistic framework" (Clement, 2012, 24). Andrew Piper, for instance, in a recent thread on Twitter, states that to be successful as a DH scholar one needs to be able

to design and employ computational methods, but one also needs to be able to frame these in a theoretical, disciplinary context.[6] From a methodological perspective, and also relating computational work to quantified approaches, Piotrowski and Fafinski (2020) draw similar conclusions. And Ramsay, too, finds that programming offers a "methodology by which the artefacts of the human record are understood, critiqued, problematized and placed into dialogue with one another" (Ramsay, 2012, §6).

2.3 New Ways of Thinking

Acquiring practical computer skills is often associated with learning new or different ways of thinking. This pertains to both reasoning–more formal, more explicit–and to attitudes towards the role of code in DH research. In 1972, Oakman claimed that "learning to program computers [is] excellent training in rigourous thinking and logical reasoning" (quoted in Oakman 1987, 227). And, when evaluating the graduate course *Computing for Humanities* taught at the University of Aberdeen in 1987, Holland and Burgess concluded that the official course objectives were "relatively modest" – learning the terminology, key concepts, components, and customs of computing, including some relevant packages and "very simple programming" – while unofficially, they hoped to bring about a change in the mind set of their students:

> we wanted to encourage a confident, independent, exploratory attitude amongst students which we hoped would eventually give them the confidence and assurance to work out for themselves how to use new machines, new software and how to apply them to new humanities problems. (Holland and Burgess, 1992, 268-269)

Similarly, Ramsay found that learning basic programming skills alters one's way of thinking: "what is gained when humanities students learn to think in the context of sophisticated computational tools is not only computational thinking, but also 'humanistic thinking'", because they learn to engage differently with traditional objects of humanistic study (Ramsay, 2012). Nick Montfort, too, sees programming "as a way of inquiring about and constructively thinking about important issues" (Montfort, 2015, 98) and argues that learning computational methods can augment, diversify, and improve humanistic methods. As iterated above, the sociological study of code literacy and programming present similar findings (Vee 2017, 105; Burgess and Hamming 2011).

2.4 Levels of Literacy

Over the years, several studies have proposed a taxonomy of code literacy in an attempt to analyse the concept in a systematic manner. The general impetus for creating a taxonomy of code literacy is the acknowledgement that a single formal definition of code literacy is too restrictive since individuals may have "differing requirements" (Webster and Webster, 1985, 2). In his analysis of the code literacy in the 1980s, Dobberstein discusses different classifications, each identifying a number of levels, domains, or components of expertise (Blau 1985, Webster and Webster 1985, Konar, Kraut, and Wong 1986 among others). Notably, all taxonomies discussed by Dobberstein are hierarchical and define code literacy along an ascending scale of expertise. A disadvantage of such a scale is that it places "all users on the same continuum of expertise" (Dobberstein, 1993, 431). Consequently, the taxonomy still

[6] See the discussion on Twitter: `https://twitter.com/_akpiper/status/1430265428711034881`, August 24, 2021.

presents a rather narrow concept of code literacy. In line with the studies discussed in 2.2, Dobberstein argues for a "context-sensitive" approach to code literacy, which accords with domain-specific requirements and skills.

2.5 Summarising the Status Quo on Code Literacy

The discourse on code literacy in the humanities can be characterised by several recurring themes. This is not to say that scholars agree on the themes: overall, we noted a general disagreement—and even confusion—in the literature about the (methodological) role and status of code and code literacy in the humanities. Another topic of discussion is the origins, motivations, and ramifications of the use of "code literacy" as a concept in the humanities. Several contributions mentioned the positive effect of (acquiring) computational skills on humanist research methods: scholars learn to be more formal and more explicit in their argumentation, and to have a confident, exploratory attitude toward computational methods. At the same time, scholars have emphasised the value of a humanist perspective on the sociological effects of code.

There is no lack of opinions from (digital) humanist scholars and educators, but there is little actual research from humanists into the functioning of code and its (methodological) effects in the humanities domain or in our society as a whole. Without a vocabulary or a clearly defined framework, the debate risks getting stuck in confusion and misunderstanding, making it hard to find consensus on the topic. In the past, a number of studies have proposed a taxonomy of code literacy that distinguishes various levels, components, types of knowledge and skills. However, the abstract, one-size-fits-all taxonomies have proven problematic, as they fail to capture the disciplinary diversity in the humanities. A context-sensitive definition of code literacy, one which is firmly embedded into a humanities discipline, is suggested as more appropriate. The contextual aspect of code literacy is emphasised by several scholars: in this definition, being code literate would include understanding how computational technologies can be applied to humanist subjects and methodologies. Opinions differ on whether "applying" means using of existing tools, developing new software, or conceptualising of computational methods.

Indeed, a large part of the discussion tends to focus on whether being code literate also implies being able to read and write code. Having (a lack of) programming skills is often felt to create a gap between DH scholars and practitioners. Some DH scholars consider programming an essential component of DH research, while others prefer theorising over practice. Both sides tend to perceive the other as exclusionary. In fact, a simple dichotomy between "programming skills" and "thinking skills" does not seem tenable in the case of code literacy in the digital humanities. As Burgess and Hamming (2011) and Vee (2017, 105) argue, the distinction between programming as mere material labour and academic scrutiny as pure intellectual endeavour is a gross underestimation of the particular type of interpretation and understanding of information that code literacy affords, and, which cannot be executed without that particular literacy.

3 Methods

The studies discussed in the previous section reveal the complexity of the topic of code literacy: there is no lack of valuable definitions and approaches, but all of these are based on theoretical considerations only or on limited empirical observation.

Furthermore, what counted as code literacy in the 1970s is often outdated in the 2020s due to technological advances. We therefore decided to question individuals currently working in DH by means of a survey in order to create a more stable dataset on the topic of code literacy in digital humanities.

As an instrument for collecting this data we decided upon an online questionnaire, to be distributed widely to anyone working in DH, including those working in the fields of GLAM (galleries, libraries, archives and museums) and LIS (library and information science), regardless of job type or background. Designing the questionnaire took seven months, from March until October 2020, and about 38 hours of discussion spread over 27 Skype calls. We were fortunate to have a diverse team that includes members with a background in information science and the social sciences, and took care to consult several experts in survey design.[7]

3.1 Survey design

By mapping the different forms of code literacy across DH disciplines, identifying problem areas and learning about existing approaches, we aimed to understand better the challenges involved with furthering code literacy in DH. We designed the survey around four research questions:

1. What are the definitions and interpretations of code literacy across humanities disciplines?
2. How important is code literacy as part of digital humanities scholarship?
3. How can we effectively approach the teaching and training of code literacy?
4. How can scholars (be supported to) incorporate code literacy in their research practice and methods?

After providing personal information (background, age group, career stage, etc.), participants were asked four sets of questions, each set designed to address one of the research questions. We followed a mixed-method approach, using both qualitative (open-ended) and quantitative (multiple-choice) questions, which is appropriate for assessing complex topics (Timans et al., 2019). For example, participants had to indicate whether they were satisfied with their own level of code literacy on a Likert scale from 1 (dissatisfied) to 5 (highly satisfied).[8] The respondents who indicated a number lower than 3 were asked to elaborate upon the reason why, in an open text field. All respondents were then asked whether they—using a yes/no question—would like to expand their level of code literacy. If they answered "no", then they were asked in a multiple choice question to give at least one reason why (options ranging from "I don't know where to start" to "I have no time".) This flexibility allows us to gain a more complete understanding of the respondent's situation.

As survey designers, we are all working in the field of DH and each of us has a different experience with learning to become code literate. Being vigilant of our own biases playing a role in the design of the survey, we opted for the "post-positivist" approach as the best way to design the survey (Ryan, 2006). A post-positivist approach

[7] The authors would like to express their gratitude to their peers, with a special mention of Merisa Martinez, PhD candidate at the School of Library and Information Science in Borås, Sweden, for her insights and advice into effective survey design and analysis.

[8] All Likert scale questions in the survey also included the option "Don't know".

recognises that the biases of the parties involved will have some effect on the targeted audience, on the survey's structure, and on the phrasing of the questions. For example, respondents may not share our perceived value of code literacy in research practices, so we should be vigilant about this and phrase questions as neutrally as possible. We also tried to be transparent about our intentions when we distributed the survey.

A post-positivist method requires us to regularly reflect on what exactly we wanted to know and how we could best articulate a question. For example, we needed to provide a broad working definition of code literacy which a large number of respondents would recognise as useful. At the same time, we realised that not all respondents would agree with how we defined it. We concluded that any definition we would provide was open for debate. Therefore, we asked respondents to provide their own definitions as well. They were subsequently asked to indicate whether they would be using our definition or their own when answering the remaining survey questions. We realised that this setup could complicate the analysis of the survey results, as for each question we would need to keep in mind that a respondent may have a diverging interpretation of code literacy. However, it also enabled us to gain insight into the various understandings of coding and code literacy in the DH community. We could explore the relationships between a respondent's definition and their answers to other survey questions, and thus step outside the framework of our own experiences.

As said, the intended audience of the survey was the wider DH community, including but not limited to scholars, teachers, students, librarians, and developers. We distributed the questionnaire via contacts at international research institutes and universities, various international (digital) Humanities email lists—e.g., the Humanist Discussion Group, the mailing list of the Text Encoding Initiative (TEI), DM-L (Digital Medievalist)—several DH slack channels (e.g. DHtech), and social media (Twitter). We identified the intention and objectives of the survey in an accompanying email as well as on the survey's home page.

With 399 responses the questionnaire has reached a large audience, but we are mindful of a self-selection effect among the respondents. That is, a digital humanist might be more inclined to answer a survey on code literacy if they consider themselves (somewhat) code literate. Another (potential) pitfall is that a qualitative survey is by nature a self-report measure. It aims to capture how a respondent feels about, and relates to, the object of study, but it is entirely depended on what the respondent chooses to share. Self-report measures have therefore been said to be subjective and "less robust" (O'Brien and McCay-Peet, 2017, 27). We prove the validity of our findings by triangulating the responses with, first, the findings from related studies (methodological triangulation) in section 2 and secondly, by distinguishing types of respondents based on background, career stage, etc.

3.2 Testing the Survey

We first collaboratively drafted the questionnaire using Google Docs. An extended period followed in which we commented and made suggestions on how the questions were articulated, on the order in which they appeared, and whether they sufficiently addressed the underlying research questions. We then structured the questionnaire using the open source software LimeSurvey[9] and tested it ourselves, discussing over the course of several meetings what did, and what did not, work. Once we agreed

[9] See https://www.limesurvey.org/. The software was running on a secure server from the Humanities Cluster of the Royal Netherlands Academy for the Arts and Sciences.

on the instrument as a team, we piloted the questionnaire by asking eight peers from a range of backgrounds (PhD students and postdocs in the Humanities, Cultural Heritage experts, and an Information Science professional) to test it and provide feedback. Specifically, we asked the testers to check the different "paths" through the questionnaire (depending on their answers to a certain question, a respondent would get a different follow-up question), to see if the questions were clearly phrased, and to inform us about the average time it took them to complete the questionnaire. Based on their feedback we adapted the instrument into its final form and distributed it over the aforementioned channels.

3.3 Preparing Response Data for Analysis

In the first round of analysis (on which we report in this article) we focus on the responses to questions related to RQ1 and RQ2 – the definition and relevance of code literacy – and the personal information provided by the respondents regarding their background, career level, etc. We analysed the results following an inductive method, as we had no hypothesis as point of departure, but a set of research questions.[10] The qualitative (open-ended, free text) questions were coded using open coding (Corbin and Strauss, 1990) that was triangulated by several members of our team in order to strengthen the validity of our coding. To give but one example: research question 7 (RQ7) asked all participants to define code literacy in a free text box. Three members of our team each coded the responses to this research question and subsequently categorised their coding. Examples of such code categories are "Conceptual aspects of code", "Aspects of literacy", or "practice". We then merged our codes into one Axial list, and recorded the responses to RQ7 using this axial coding. Next, the other members of our team used the Axial list to code the same data. In between, we held regular meetings to discuss the codes and consolidate our views on tricky categories. We further analysed the data using Jupyter notebooks. The results of this analysis are discussed in section 4.

3.4 Privacy

The landing page of the survey informed the respondents of the privacy aspects regarding the handling of their responses, which also included the names, affiliations and email addresses of the survey creators. We also informed respondents that their responses will not be shared in raw form beyond the creators of the survey. The questionnaire was anonymous, meaning that we do not store any identifying information about the responses, such as IP address. The only exception to this was the option for respondents to leave their email address to be contacted for a follow-up interview. The questionnaire included a number of questions on personal information, such as job type, academic background, gender, career stage, and country. We follow the data management guidelines of the KNAW[11] and the European Union[12] and store the data for a period of five years, to allow our analysis to be reproducible. The survey has been reviewed according to the GDPR and the Utrecht University guidelines for research

[10] Azungah 2018. See also https://conjointly.com/kb/deduction-and-induction/.

[11] https://www.knaw.nl/en/news/publications/big-data-in-wetenschappelijk-onderzoek-met-gegevens-over-personen/@@download/pdf_file/20180515-advies-Big-Data.pdf

[12] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf.

involving human subjects.[13]

4 Results

In this section we discuss the responses related to the two research questions addressed in this article: the definition and relevance of code literacy.

4.1 Demographics of Respondents

The questionnaire started with a series of demographic questions to get an understanding of who completed the questionnaire and therefore also of potential over and under-representation of certain groups within the DH community.

We made the question about the respondent's gender an open text field to allow them to express their gender in the way they saw fit. As the vast majority filled either "female" or "male" , we coded the responses using four groups: female ($n = 166$ or 42%), male ($n = 208$, or 52%), other ($n = 7$, or 2%) and unspecified ($n = 25$ or 5%). The latter corresponds to respondents who left the field empty. From this we observe that female and male sexes seem to be equally represented in the sense that the respondents are not predominantly male or predominantly female. For gender, as for other demographic aspects, we cannot make any claims as to how representative this distribution is for the DH community, as it is hard to determine the boundaries of this community.

The majority of respondents work at universities and research institutes. We asked for the country of their current affiliation (in the case of multiple affiliations, we asked them to select one), using a controlled list of country names, see Figure 1. Just under half of all respondents are from Europe ($n = 190$ or 48%), with another large group ($n = 106$ or 27%) being from North America. Not all respondents filled in an affiliation and corresponding country ($n = 85$ or 21%). This puts some limitations on the representativeness of this survey, as Oceania, South America, and especially Africa and Asia are underrepresented (or not at all as in the case of Africa). Any claims we make should therefore be considered to apply more to Europe and North America than to the DH community more broadly.

We also asked respondents about their current role(s) in DH scholarship, and provided a list of roles (e.g. *academic*, *developer*, *librarian*, *administrative personnel*, *student*) as well as a free-text box. Respondents could select and/or enter any number of roles. The free form roles overlapped partially with the list of provided roles, so we coded them into five main groups: *researcher*, *student*, *IT specialist* (including developer, software engineer and technical support staff), *information specialist* (which includes archivists, librarians, curators, documentalists and data specialists) and *other* (including administrative personnel, editors, publishers and designers), see left side of Figure 2.

[13] For the GDPR, see: https://gdpr-info.eu For the Utrecht University guidelines, see: https://www.uu.nl/en/research/research-data-management/guides/handling-personal-data#anonymise

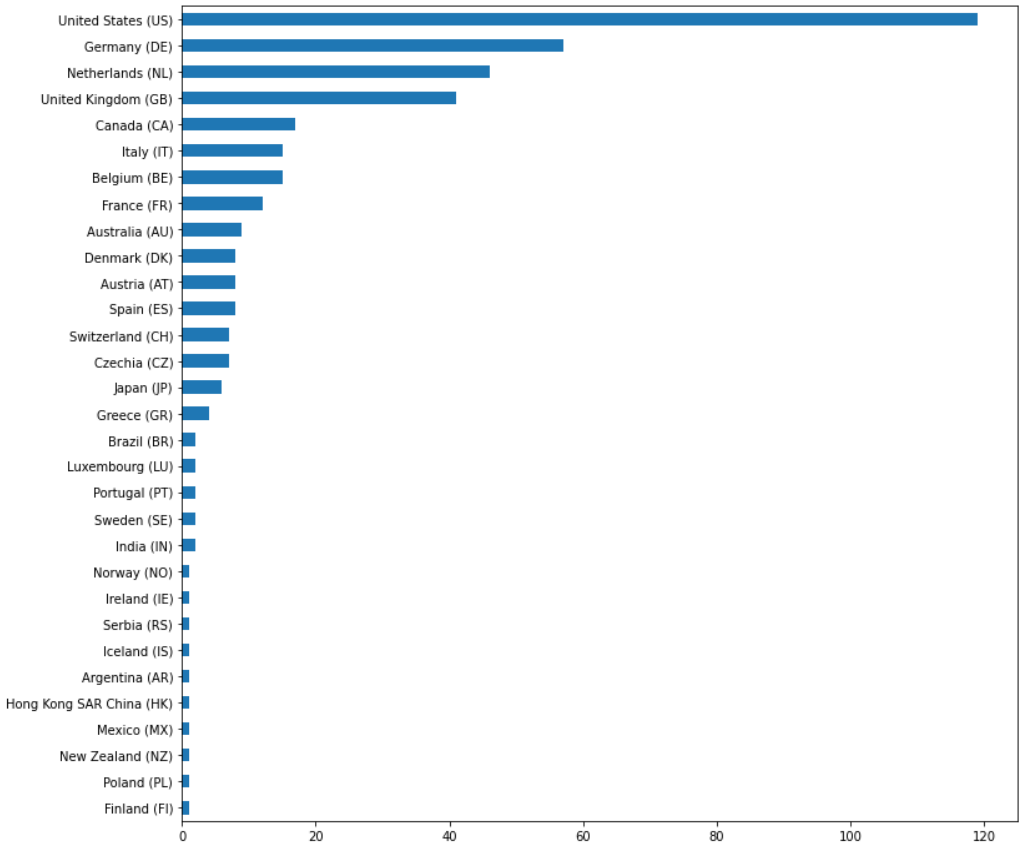


Figure 1: Distribution of respondents across countries.

The majority of respondents consider themselves *researchers* ($n = 301$ or 75%). The other two main groups, not surprisingly, are *IT specialists* ($n = 76$ or 19%) and *information specialists* ($n = 77$ or 19%). Most groups overlap substantially with the *researcher* group. For instance, 57% of *IT specialists* ($n = 44$) are also *researchers*. The other groups are more distinct, with only 19% of *information specialists* also having a role as *IT specialist* (vice versa is 20%). Altogether this suggests we managed to reach people with a diverse set of roles in the DH community.

Next, we asked respondents about the number of years of experience they have working with the field of Digital Humanities (see right side of Figure 2). Note that this does not necessarily correspond to their number of years of experience within their current roles, as some may have been a researcher or developer for decades but only started working in DH less than a year ago.

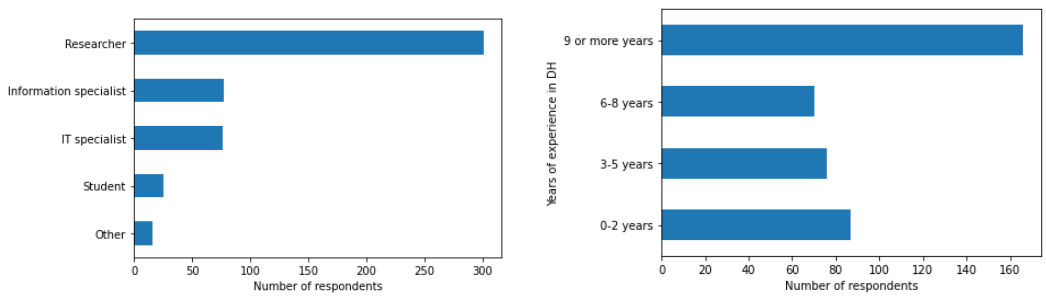


Figure 2: Distribution of respondents across roles and career phases in DH.

Almost half of all respondents has 9 or more years of experience in DH and less than a quarter, fewer than 2 years. It is difficult to establish how representative this is of the entire DH community. Our guess is that the more experienced contingent of the community is somewhat over-represented, as they are more likely to have come across our survey via the channels we used.

Respondents also have a wide range of academic backgrounds (see Figure 3), with large groups of respondents having a humanities background, specifically in *History*, *Language and Literary Studies*, or *Textual Scholarships*. (The latter includes Book History, Paleography, Scholarly Editing and Textual Criticism.) There are also many respondents with backgrounds in *Computer Science* and *Library and Information Science*.

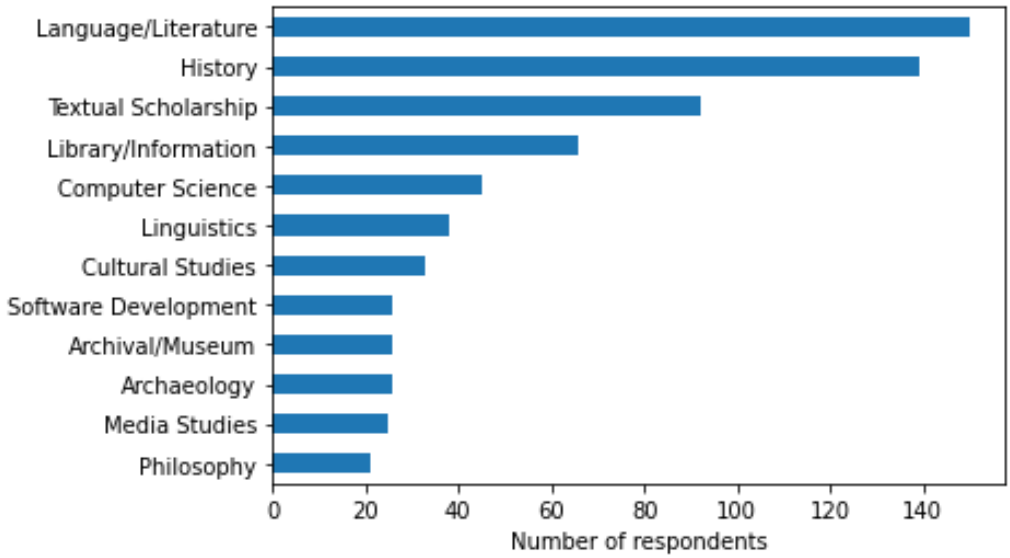


Figure 3: Academic backgrounds of respondents, where they could check multiple backgrounds from the ones we listed as well as add additional backgrounds in a free text field. The table lists only the backgrounds checked by at least 5% of the respondents.

4.1.1 Overlapping categories and significance testing

Before moving on to discussing how respondents define code literacy and answered further questions, we discuss ways in which we can meaningfully compare sub-groups of respondents along different demographic dimensions. To test the observed differences for statistical significance, we need to consider different tests for the variables that were coded with mutually exclusive categories - such as gender and career phase - and those that were coded with overlapping categories, like role or disciplinary background, where respondents could select multiple answers.

The former type of variables were tested using a one way ANOVA with Tukey Honest Significant Difference (HSD) post-hoc tests (Sachs, 2012, Ch. 7). For the latter type of variables, we consider two options. One is to make as many groups as there are combinations of answers. E.g. all respondents who indicated to have both a role as *researcher* and as *information specialist* are in a separate group from those who selected only *researcher* or only *information specialist*. In this way, no respondents belong to multiple groups, but we can still compare the groups for differences in how they define code literacy or answer other questions. There are two drawbacks to this approach. One is that it creates many groups, each with only a small number of respondents, resulting in many comparisons of pairs of groups, which are often too

small to establish statistical significance. The other issue is that even though this step establishes distinct groups, they can be hard to interpret. How should we interpret the difference between the group of respondents who are *researcher*, *information specialist* and *IT specialist* against the group who are *researcher* and *information specialist*, the group who are *researcher* and *IT specialist*, and those who only selected *IT specialist*, *information specialist* or *researcher*?

The other option we consider is to compare all respondents who selected a particular role or background, against who did not, using a χ^2 test. This has the disadvantage that we do not directly compare groups, but the advantage of more clearly interpretable test hypotheses and outcomes. E.g. are respondents with a role as *IT specialist* more or less likely to include *writing code* as part of the definition of code literacy than those who do not have this role? In the analysis below, we use the latter option. That is, we use the χ^2 test on the group with a particular attribute or response and the complement of that group.

4.2 Definitions and Interpretations

After the demographic questions, we asked respondents to provide their own definition of code literacy. As this was a required question, all 399 respondents provided a definition.

Below are a few examples of typical definitions given by the respondents:

> *"The ability to understand and write code and to use it to achieve some research goal."*
> *"The ability to read, write and use code."*
> *"knowledge and experience of solving problems through the use of programming skills"*

But there are also many definition that are more elaborate and rich:

> *"I'd say there's informal and formal code literacy. Formal literacy is what you find with colleagues in the sciences and engineering. There's a strong emphasis on style, standards, and efficiency. Tests and quality control are required. Then there are the programming historians, the self-taught, and highly pragmatic types. It's just amazing that it works at all. The code is an odd pastiche of cut and pastes from Stack Overflow. It's a game of persistence, self-doubt, and curiosity. I mostly work with informal code literacy to give people the practical skills of reading error messages and documentation, finding helpful solutions to problems, and making something that works but is by no means pretty or professional."*

To be able to analyse and compare these definitions, we coded them using open, axial, and selective coding methods (Corbin and Strauss, 1990). The open coding step was done individually by three of the authors. In the axial coding step, each of the annotators had already created a hierarchy for their own codes, so we compiled a list of all the hierarchical codes and derived a single hierarchy of codes. The overall hierarchy consists of five basic aspects, each with several sub-aspects. The basic aspects are *Communication*, *Code Type*, *Contextual Understanding*, *Level of Competence* and *Other*, see Table 1.[14] Apart from the five basic aspects, we added a code for responses that do not resemble a definition but, according to us, seem like a response to a different question. An example of this "I have none experience whatsoever, but I would like to change that." This is the case for 23 responses. One respondent only filled in a hyphen, possibly to be able to move on to the next question without providing any definition.

Further in the survey, we included a working definition of code literacy, and asked respondents if they wanted to answer the remaining questions using our definition or

[14] The full hierarchy of aspects, including the scope notes of each, is provided in Appendix A.

their own. We included this in case respondents were uncertain or not satisfied with their own definition. Our working definition was:

> Code literacy has to do with different levels of ability to recognise, interpret, and use code; not necessarily being able to create it yourself. In this context, code can refer to encoding (e.g. XML or MPEG) as well as program code (e.g. Python or R) to operationalise a process in concrete steps and actions.

Code	*Description*	*Scope note*
COM	Communication	Knowing how to communicate about code with others, either as a coder yourself, or with someone who codes for you, including communication about purpose, workings, role and implications, as well as teaching code literacy.
CT	Code Type	Whether the code explicitly refers to a specific type of code, either 1) code as encoding of documents or 2) performative code for e.g. processing of data.
CU	Contextual Understanding	Understanding that code sits in a context and how it relates to 1) research and operationalising research questions, 2) it's possibilities, limits and biases, 3) the culture and attitudes toward code, 4) the ecosystem around code of ethics, privacy, security, maintenance, documentation, versioning, licensing, practices, platforms and software, and 5) other aspects of context.
LOC	Level of Competence	Literacy is divided into 7 different levels of competence, from 1) understanding the basics of what code is and does, to being able to 2) read code, 3) write basic code, 4) modify existing code, 5) review code, 6) create and package code from scratch, and 7) understanding of the theory of computation and coding paradigms. Separate from those levels there is a sub-category for definitions that mention there are multiple levels of competence.
OTHER	Other aspects	Any code literacy aspects not covered by the above four aspects.
MIS	Misinterpretations	Responses that are not definitions but answers to a different question.

Table 1: The top level aspects of our axial coding of the respondent's definitions of code literacy.

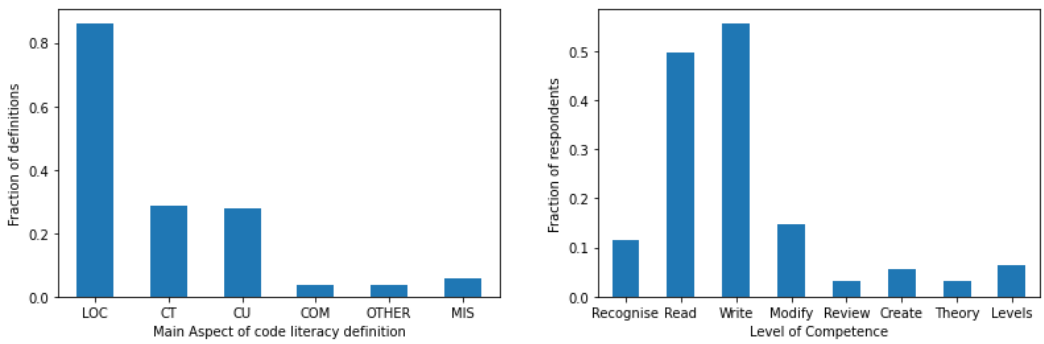


Figure 4: Fraction of definitions that include a main definition aspect (left) and that include sub-aspects of Level of Competence for all 399 respondents.

The distribution of aspects given in respondents' definitions is shown on the left side of Figure 4. The vast majority of respondents mention at least one of the levels of competence (*LOC*) as part of the definition of code literacy, such as the following two definitions:

> *"the ability to recognize, understand and write codes of any kind, e.g. markup languages, programming languages, annotations or any other kind of symbolic, rule-based/formalized means of communication"*
>
> *"the ability to read code to comprehend the purpose of functions, their syntax, and what output results; the ability to interact with code to improve its functionality, ether in terms of efficiency/clarity or in terms of research purpose; the ability to write (invididually or collaboratively) new code"*

Only around 28% mention code type (*CT*) or contextual understanding (*CU*). Communication (*COM*) is rarely mentioned (less than 4%). We note that our working definition did not include any aspects of contextual understanding (although it hints at operationalising research processes), but aspects of it were mentioned by at least 20% of respondents, in each of the 12 disciplines, with more than 20 respondents. There are significant differences between disciplines in how many respondents mention any aspects of *Contextual Understanding*.

Looking at the number of coded aspects in each definition, we found that most respondents mention one, two or three aspects (30%, 34% and 20% respectively), with the remaining 16% of definitions mentioning between four and seven aspects.

> *"being able to write and deploy code when you encounter problems that code could help you deal with; being able to read code of others; being able to read pseudo-code and series of formulae in publications; knowing how to use tools that support code development and implementation (version management tools, Pypi, etc.); being familiar with terminology used in discussing code and coding issues."*

We found that *IT specialists* mention more aspects in their definitions than non-IT specialists (2.8 versus 2.2 aspects, χ^2 with $p = 0.002$), but career phase has no significant impact.

4.2.1 Level of Competence

Level of Competence (LoC) is by far the most mentioned aspect in the code literacy definitions. Across gender categories, career phases, roles and disciplinary backgrounds, we find no significant differences in how often LoC is mentioned. This suggests that at least some sub-aspects of LoC are core elements of code literacy on which there is a broad consensus. Moreover, it reflects that discussions on the role of programming are still central to defining code literacy (Dobberstein, 1993) even though respondents differ when it comes to prioritising competences. Zooming in on the sub-aspects of *Level of Competence*, shown on the right side of Figure 4, reveals that the main aspects that respondents agree on are the ability to read ($N = 198$ or 50%) and write code ($N = 222$ or 56%). Digging further into the *read* and *write* aspects of definitions, we noticed that there is a substantial group ($n = 77$ or 19%) who mention *reading* code as a core competency of code literacy, but not *writing* code. This group of definitions is not associated with any specific disciplines. Many of these definitions emphasise that at a minimum, those involved in DH research should be able to read and get the gist of what a piece of code does, without necessarily being able to write such code. Examples include:

> *"Being able to read and understand, not necessarily write, code"*
>
> *"Understanding of how code works. Not necessarily how to code something specific yourself, but you*

are able to understand how code is build and how to read it".

On the other hand, 101 respondents (25%) mention *writing* code but not *reading* code, indicating that for another substantial group, the best or most logical approach to learning to code is by doing. Examples are:

"Ability to write/correct code that runs successfully, eventually"
"The ability to write practically usable code in some computer programming language"
"be able to write code to solve problems"

Overall, these results suggest that the DH community is still largely split on the question if code literacy implies possessing programming skills (Ide, 1987; Ramsay, 2013b; Rushkoff, 2010). While a majority supports the contention that literacy means writing of code, a still substantial group of respondents would disagree, foregrounding other ways of engaging with programming languages.

This prompted us to think about the different ways in which code literacy can be taught or how different types of learning materials can be offered. For some groups of students and scholars, or within some curricula, it might make sense to focus first and foremost on how code relates to aspects of their discipline, how researchers translate between research questions and following the logic of code, thereby emphasising how code fits in the research *process*, while for others, the best way might be to start with writing straightaway. For instance, people who already have some experience writing basic code, might improve their skills (e.g. with additional elements or a different language or paradigm) more efficiently through writing. Alternatively, the nature of research in some disciplines is more directly translatable to computational thinking and breaking down the process into procedures with explicit steps, making it more relevant to learn writing small bits of code from the start. While for other disciplines, the connection between their research process and code might be more complex, in which case, learning about the role and possibilities of code might require focusing on its relevance and potential, and engaging with examples of code that have been successfully applied in their discipline.

The other aspects in the *levels of competence* are mentioned by only few respondents, and represent more advanced perspectives on code literacy.

4.2.2 Contextual Understanding and Code Type

Next, we zoom in on the two other main aspects mentioned regularly in definitions, *Code Type* and *Contextual Understanding*. First, since only 28-29% of respondents mention these aspects, we want to know if these respondents differ from the others in terms of demographics, background, literacy or career phase.

Among the groups of roles we created, IT specialists more often also think about other aspects (code type and contextual understanding). We assume that IT specialists are most involved in working with code, and as noted above, they have more elaborate definitions than most others.

For example, this definition is one of the most "dense" in terms of elements (given by an academic researcher):

"I distinguish:
- a basic understanding of the principle of coding and the affordances and limitations for one's discipline, necessary to collaborate with a computer scientist or software developer
- the capacity of executing predefined rules in a learning environment
- mastering a computer language and knowing how to write code to execute particular actions

- knowing several computer languages that are relevant to a specific domain and the specificities of each one of them
- being all round with computer languages that cover the whole eco-system of digital humanities , so that you can create an infrastructure".

As well as this one provided by an IT specialist:

"The ability to conceive and implement a software solution to solve a specific problem in an adequate and efficient manner. The ability to understand and augment solutions implemented by others. The ability to use common infrastructure tools needed in software development (text editors or IDEs, version control)."

Only 28% of respondents made explicit reference to type(s) of coding in their definitions (i.e., encoding or processing). When they did, some mentioned both, some mentioned only one. Textual Scholars are more likely to mention *encoding* than non-Textual Scholars (16% versus 6%, χ^2 with $p = 0.009$) and respondents who have a background in software Development are more likely to mention *processing* than respondents who do not (58% versus 24%, χ^2 with $p < 0.001$).

Within *Contextual Understanding*, the *possibilities, limitations* and *biases* are more likely mentioned by respondents with a background in *Archival & Museum Studies* (27% versus 8%, χ^2 with $p = 0.003$), *Library and Information Science* (17% versus 8%, χ^2 with $p = 0.003$) or *Linguistics* (21% versus 8%, χ^2 with $p = 0.002$).

Given the emphasis on contextual understanding in the (academic) literature (Clement, 2012; Dobberstein, 1993), the relatively low number of mentions of this aspect (mostly coming from IT specialists and *not* researchers) was somewhat of a surprise. It suggests that the community could benefit from expanding its notion of code literacy beyond discussing competencies, or programming languages, and instead perceive literacy as the understanding of how code interacts with/relates to the humanistic framework in which it operates(Clement, 2012).

4.2.3 Code Literacy Level of Respondents

As people with several years of coding experience may have different definitions than people who have never directly interacted with code before, we asked respondents to score themselves on a five-point scale as to how code literate they consider themselves to be, given their own definition of code literacy (see left side of Figure 5).

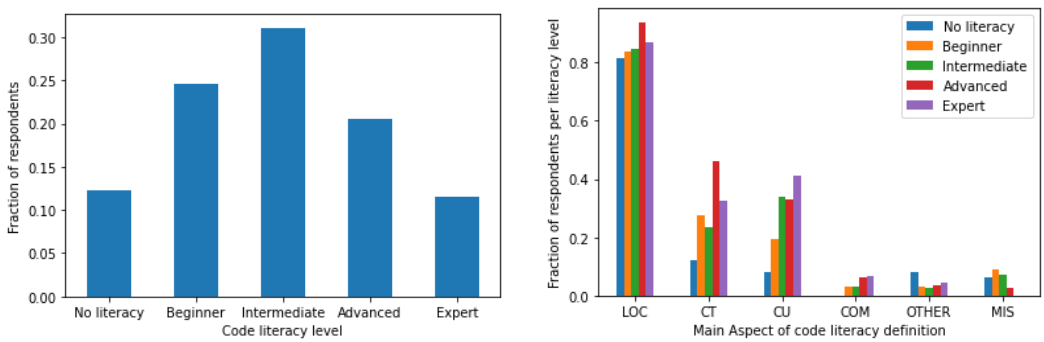


Figure 5: Distribution of self-reported code literacy levels among all 399 respondents (left), and of main aspects of definitions per code literacy leve.s

The five-point scale has the following labelled levels: *No literacy*, *Beginner*, *Intermediate*, *Advanced* and *Expert* (ConcordiaUniversity, 2011).

The distribution of the responses is close to a normal distribution, with the most frequent level being *Intermediate*, and frequencies tapering off towards the extremes. The *Experts* are the smallest group with $n = 46$ respondents.

There is an association between gender and literacy level. Women are more likely to consider themselves at *beginner* level than men (32% versus 18%, Tukey HSD with $p = 0.008$), while men are more likely to consider themselves at *advanced* (11% versus 29%, $p = 0.001$) and *expert* level (4% versus 17%, $p = 0.001$).

There is also a clear association between role and literacy level. *Information specialists* are less likely than non-information specialists to consider themselves *advanced* (10% versus 23%, $p = 0.02$), while, as is to be expected, IT-specialists less frequently have *no literacy* (0% versus 15%, χ^2 with $p = 0.001$) or be at *beginner* (11% versus 28%, $p = 0.003$) or *intermediate* (21% versus 33%, $p = 0.05$) level than non-IT specialists, but more likely to be at *advanced* (37% versus 17%, $p < 0.001$) or *expert* level (32% versus 7%), $p < 0.001$).

There are also significant relationships between disciplinary background and code literacy level, which we argue has implications for how code literacy is best taught within different disciplines. Respondents with a background in History, are more often than non-Historians at the level of *no literacy* (20% versus 8%, χ^2 with $p = 0.001$) or *beginner* (32% versus 21%, $p = 0.02$), and those with a background in Language and Literature are less likely to be at *expert* level than those without (7% versus 14%, $p = 0.03$), while those with a background in Linguistics are less likely to be at *beginner* level than those without (5% versus 27%, $p = 0.007$). These differences are visualised in Figure 6.

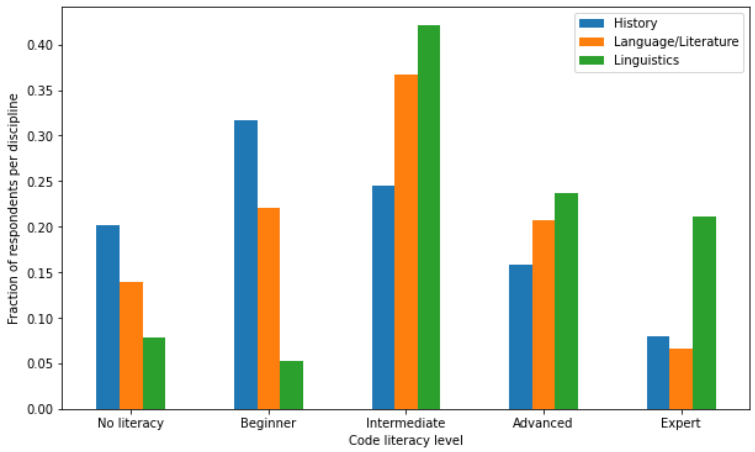


Figure 6: Code literacy level of respondents with a background in History, Language and Literature, and Linguistics.

One possible explanation for the differences between these humanities disciplines is the different research methods and data they use. Linguistics has a long history of computational approaches to analysing speech and textual utterances using quantitative methods (Hajic, 2004), so perhaps the step of translating methods and techniques to code is relatively small and requires no significant change in mode of thinking. Whereas for historians, the established methods of archival research and making associative connections through close reading of heterogeneous documents are perhaps less easily translated into quantifiable and computational steps. It may therefore be more beneficial for historians to see examples of how computational processes have been translated into a recognisable part of historical research. This can help both the

understanding of the potential and limitations of code for their discipline, as well as introduce "thinking in code" or "thinking through code" as a relevant mode of thinking.

Interestingly, although historians and language and literature scholars score themselves lower on code literacy than many others, they form the two largest disciplinary groups in this survey, each represented by over 140 respondents. This could potentially be related to representation and self-selection bias (see Section 4.1). Again, as expected, respondents with a background in Software Development or Computer Science are rarely at the level of *no literacy*, *beginner* or *intermediate*, and are more often at the level of *advanced* or *expert*.[15]

The level of code literacy of respondents is clearly associated with the definitions (right side of Figure 5). The majority of respondents at all levels agree that *Level of Competence* is part of the definition, but *Code Type* and *Contextual Understanding* are mentioned more frequently by respondents with higher levels of code literacy. Respondents at Advanced level are significantly more likely to mention *Code Type* (41%), specifically *Processing*, than those at No Literacy (8%, Tukey HSD with $p = 0.001$), Beginner (19%, $p = 0.003$) and Intermediate (34%, $p = 0.04$) levels. The same goes for *Contextual Understanding*, which respondents at No Literacy level are significantly less likely to mention (12%) than respondents at Intermediate (23%, Tukey HSD with $p = 0.005$), Advanced (46%, $p = 0.02$) and Expert (33%, $p = 0.003$) levels, and those at Beginner level are significantly less likely to mention it (28%, $p = 0.04$) than those at Expert level.

From this we speculate that more direct experience with code provides a richer vocabulary to talk about code, a wider perspective on how code relates to the research questions and processes, and that writing and interacting with code brings people into contact with the larger ecosystem of coding languages, interpreters, data formats, versioning and management of code, aspects of ethics and privacy and other elements.

This does not mean that *Contextual Understanding* is a more advanced concept that should be taught at a later stage in the curriculum. We argue that more experienced coders are more aware of the importance of the context in which code is created and used. This also ties in with the earlier observation that foregrounding Contextual Understanding would help establishing a more nuanced and broader perception of code literacy in the DH community.

To conclude our analysis of the definitions of code literacy, there are three important dimensions to consider when talking about code literacy: *Level of Competence* is the main dimension that respondents across all career phases, roles, disciplinary backgrounds and levels of code literacy mention, while *Code Type* and *Contextual Understanding* are more recognisable to those with some experience in using or creating code. The different perspectives of what code literacy is, also suggest that there is no one-size-fits-all approach to teaching it, but that it makes sense to differentiate between different disciplines and roles. We will address the question of how to teach and incorporate code literacy in curricula in a follow-up article.

4.3 Importance of Code Literacy

We turn now to the second research question. How important is code literacy as part of DH scholarship? After giving their own definition of code literacy, we asked respondents to consider how important they think code literacy is for digital humanities

[15] We leave out percentage and P-values for readability.

scholarship, provided them with five different levels of importance as well as the option *Don't know*. The five different levels are: *Not important at all*, *Somewhat important*, *Important*, *Very important* and *Crucial*. The distribution of responses is shown on the left in Figure 7. Only 5 respondents indicated that they did not know (1%), and only 21 (5%) answered *not important at all*, which means 373 respondents (93%) consider code literacy to be at least *somewhat important*.

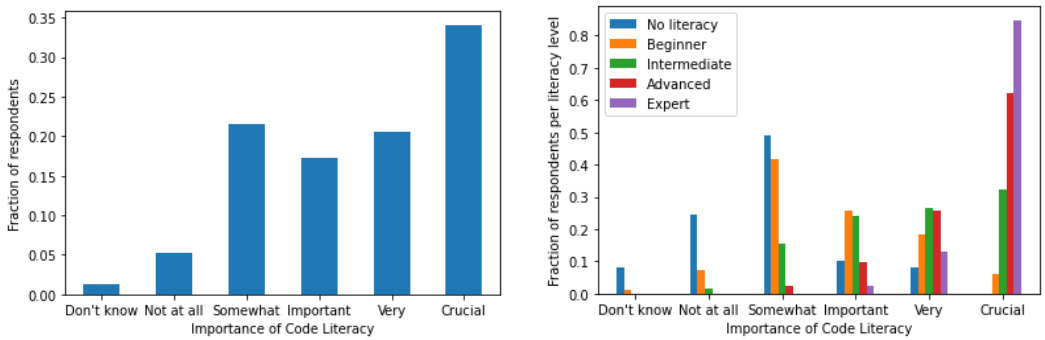


Figure 7: Distribution of the level of importance of code literacy for DH scholarship, among all 399 respondents (left) and across the five code literacy levels (right).

There are differences across background disciplines. The disciplines with the highest percentage of respondents who consider code literacy not important at all are Media Studies (12%), History (10%) and Cultural Studies (9%). So across all disciplines in our survey, the vast majority agree code literacy has a place in DH scholarship.

Historians are significantly more likely to consider code literacy as *not important at all* compared with non-historians (10% versus 3%, χ^2 with $p = 0.004$) and significantly less likely to consider it *very important* (12% versus 25%, $p = 0.002$). Respondents with backgrounds in Archival and Museum studies are significantly less likely to consider code literacy *crucial* than those without these backgrounds (7% versus 36%, $p = 0.006$). On the other hand, linguists are significantly more likely to consider it *crucial* than non-linguists (55% versus 32%, $p = 0.007$). Again we see the strong contrast between historians and linguists. Their difference in perceived importance might be related to their difference in levels of code literacy.

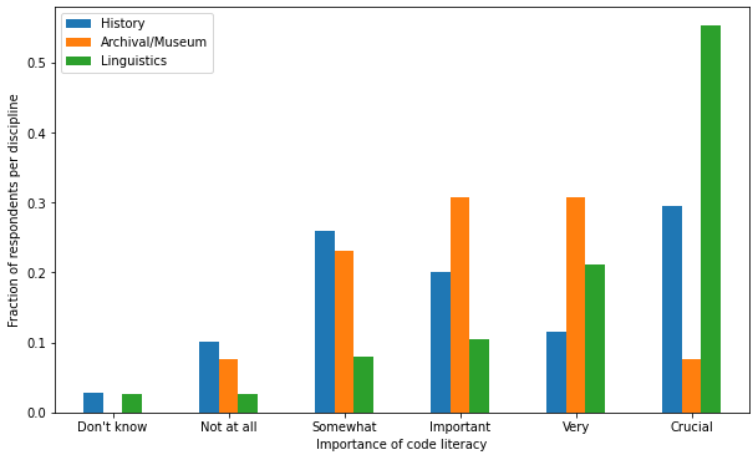


Figure 8: Distribution of respondents on importance levels per discipline for historians, linguists, and specialists in archival and medium studies.

Apart from these three humanities disciplines (represented in Figure 8), the only other disciplines with statistically significant differences are computer science and software development, which, not surprisingly, score lower on *somewhat important* (7%, χ^2 with $p = 0.02$ and 0% with p$p = 0.01$ versus 23% respectively), but higher on *very* and *crucial* (respectively 60% and 77% versus 31%, $p < 0.001$). Keeping in mind the possible self-selection bias mentioned in Section 4.1, the results suggest there is a broad consensus across a wide range of disciplines, within the humanities and beyond, that code literacy is important to many DH scholars.

5 Discussion

This paper analysed the results of a questionnaire on the definition and importance of code literacy in digital humanities. Our coding and analysis of the 399 definitions of code literacy reveals a complex, multi-layered and multi-faceted perspective on code literacy; from the basic skills of reading, interpreting, writing and using code, to publishing and reviewing code, to the different contexts in which code is created and used – the research context, the ecosystem of hardware, software, and communities and conventions – as well as the societal context relating to ethics, privacy and bias.

We found that these different aspects are related to the experience levels of respondents, with more code literate respondents providing more elaborate and nuanced definitions. This suggests that code literacy is not a single level that one reaches at a certain point, but is a set of skills that people continuously improve and extend within a particular context.

The order in which these skills are best learnt and developed is not necessarily related to the level of literacy of the respondents that mention them. The context of research in which one translates between research questions and methods on the one hand, and how that can be expressed, modelled or performed via code on the other hand, is mentioned most by respondents who identify as more code literate. However, we argue that this contextual understanding should be learnt early on, as it is the most directly relevant aspect for integrating code into DH scholarship. Many coding practitioners in the (digital) humanities, humanities researchers, and sociologists of code literacy seem to have gravitated to a similar attitude, although they differ of opinion on how to define this "contextual understanding".

A large part of the respondents think that code literacy is important for DH scholarship. However, the particular distributions we see regarding who thinks code literacy is important should also inspire questions about the reasons *why* we consider code literacy to be important. It seems safe to assume though, that the perception that code literacy is important in DH is on the rise. And yet, many respondents are dissatisfied with their own level of code literacy. This prompts the question as to whether there is currently a gap in academic curricula regarding digital humanities? And if so, would enhancing code literacy fill this gap? And what should such teaching look like according to our respondents? Questions such as these will be addressed in follow-up publications in which we will analyse and discuss the remaining questions and responses of the questionnaire.

References

Theophilus Azungah. Qualitative Research: Deductive and Inductive Approaches to Data Analysis. *Qualitative Research Journal*, 2018.

BenAmi Blau. Computer Literacy: What and Who? *Personnel*, 62(12):7–8, 1985.

Helen J. Burgess and Jeanne Hamming. New Media in Academy: Labor and the Production of Knowledge in Scholarly Multimedia. *DHQ: Digital Humanities Quarterly*, 5(3), 2011. URL http://digitalhumanities.org/dhq/vol/5/3/000102/000102.html.

Elizabeth Callaway, Jeffrey Turner, Heather Stone, and Adam Halstrom. The Push and Pull of Digital Humanities: Topic Modeling the "What is digital humanities?" Genre. *Digital Humanities Quarterly*, 14(1), 2020. URL http://www.digitalhumanities.org/dhq/vol/14/1/000450/000450.html.

Monica Chin. Kids Who Grew Up With Search Engines Could Change STEM Education Forever, 2021. URL https://www.theverge.com/22684730/students-file-folder-directory-structure-education-gen-z.

Tanya Clement. Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind. In Brett D. Hirsch, editor, *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge, UK: Open Book Publishers, 2012. ISBN 9782821854031. URL http://books.openedition.org/obp/1656.

ConcordiaUniversity. Computer skills: Levels of proficiency, 2011. URL https://www.concordia.ca/content/dam/concordia/services/hr/docs/employment/guides/proficiency-computer-skills.pdf.

Juliet M. Corbin and Anselm Strauss. Grounded Theory Research: Procedures, Canons, and Evaluative Criteria. *Qualitative Sociology*, 13(1):3–21, 1990.

Adam Crymble. *Technology and the historian: transformations in the digital age*, volume 1. University of Illinois Press, 2021.

Michael Dobberstein. Computer Literacy for the Rest of Us. *Computers and the Humanities*, 27(5-6):429–433, September 1993. ISSN 0010-4817, 1572-8412. doi: 10.1007/BF01829393. URL http://link.springer.com/10.1007/BF01829393.

Amy Earhart, Matthew Kirschenbaum, Bethany Nowvisky, Katherine Harris, Vika Zafrin, Patrick John Murray, and Daniel Allington. Doing DH vs. Theorizing DH, 2016. URL https://dhanswers.ach.org/topic/doing-dh-v-theorizing-dh/#post-436.

Matthew K. Gold, editor. *Debates in the Digital Humanities*. 1. University of Minnesota Press, 2012a. doi: https://doi.org/10.5749/9781452963754.

Matthew K. Gold. Introduction: the Digital Humanities Moment. In Matthew K. Gold, editor, *Debates in the Digital Humanities*, pages 9–16. Minneapolis, MN: University of Minnesota Press, 2012b.

Matthew K. Gold and Lauren Klein, editors. *Debates in the Digital Humanities*. 5. University of Minnesota Press, 2019. doi: https://doi.org/10.5749/9781452963785.

Jan Hajic. Linguistics meets exact sciences. *A companion to digital humanities*, 42:79, 2004.

Susan Hockey. Workshop on Teaching Computers and the Humanities Courses. *Literary and Linguistic Computing*, 1(4):228–229, 01 1986. ISSN 0268-1145. doi: 10.1093/llc/1.4.228. URL https://doi.org/10.1093/llc/1.4.228.

Simon Holland and Gordon Burgess. Beauty and the Beast: New Approaches to Teaching Computing for Humanities Students at the University of Aberdeen. *Computers and the Humanities*, 26(4):267–274, August 1992. ISSN 0010-4817, 1572-8412. doi: 10.1007/BF00054272. URL http://link.springer.com/10.1007/BF00054272.

Nancy M. Ide. Computers and the Humanities Courses: Philosophical Bases and Approach. *Computers and the Humanities*, 21(4):209–215, 1987. ISSN 0010-4817, 1572-8412. doi: 10.1007/BF00517809.

Kenneth M. King. Evolution of the Concept of Computer Literacy. *Educom Bulletin*, 20 (3):18–21, 1985. URL https://eric.ed.gov/?id=EJ331749.

Ellen Konar, Allen I. Kraut, and Wilson Wong. Computer Literacy: With Ask You Shall Receive. *Personnel Journal*, 65(7):83–86, 1986.

M.C. Marino. Critical Code Studies, 2006. URL http://www.electronicbookreview.com/thread/electropoetics/codology.

Liliana Melgar, Mari Wigham, and Marijn Koolen. Programming Humanists – What Is the Role of Coding Literacy in Digital Humanities and Why Does It Matter? In *DH Benelux Conference 2019*. University of Liège, 2019. URL https://2019.dhbenelux.org/.

Nick Montfort. Exploratory Programming in Digital Humanities Pedagogy and Research. In *A New Companion to Digital Humanities*, pages 98–109. John Wiley & Sons, Ltd, 2015. ISBN 978-1-118-68060-5. doi: 10.1002/9781118680605.ch7. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118680605.ch7.

Douglas Noble. Computer Literacy and Ideology. *Teachers College Record*, 85(4): 602–614, 1984.

Robert L. Oakman. Perspectives on Teaching Computing in the Humanities. *Computers and the Humanities*, 21(4):227–233, 1987. ISSN 0010-4817, 1572-8412. doi: 10.1007/BF00517811. URL http://link.springer.com/10.1007/BF00517811.

Heather L. O'Brien and Lori McCay-Peet. Asking "Good" Questions: Questionnaire Design and Analysis in Interactive Information Retrieval Research. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 27–36, 2017.

James O'Sullivan, Diane Jakacki, and Mary Galvin. Programming in the Digital Humanities. *Digital Scholarship in the Humanities*, pages 142–147, October 2015. ISSN 2055-7671, 2055-768X. doi: 10.1093/llc/fqv042. URL https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqv042.

Michael Piotrowski and Mateusz Fafinski. Nothing New Under the Sun? Computational Humanities and the Methodology of History. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, volume 2723, pages 171–181, Amsterdam, November 2020. CEUR Workshop Proceedings. doi: urn:nbn:de:0074-2723-3. URL http://ceur-ws.org/Vol-2723/short16.pdf.

W. James Potter. The State of Media Literacy. *Journal of Broadcasting & Electronic Media*, 54(4):675–696, November 2010. ISSN 0883-8151. doi: 10.1080/08838151.2011.521462. Publisher: Routledge.

Stephen Ramsay. Programming with Humanists: Reflections on Raising an Army of Hacker-Scholars in the Digital Humanities. In Brett D. Hirsch, editor, *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge: Open Book Publishers, 2012. ISBN 9782821854031. URL http://books.openedition.org/obp/1642.

Stephen Ramsay. On Building. In Edward Vanhoutte, Julianne Nyhan, and Melissa Terras, editors, *Defining Digital Humanities: A Reader*, pages 243–246. Ashgate, 2013a.

Stephen Ramsay. Who's In and Who's Out. In Edward Vanhoutte, Julianne Nyhan, and Melissa Terras, editors, *Defining Digital Humanities: A Reader*, pages 239–242. Ashgate, Surrey, UK, 2013b. URL https://web.archive.org/web/20121015012254/http://stephenramsay.us/text/2011/01/08/whos-in-and-whos-out.html.

Douglas Rushkoff. *Program Or Be Programmed: Ten Commands for a Digital Age*. OR Books, 2010.

Anne B. Ryan. Post-positivist Approaches to Research. *Researching and Writing your Thesis: a Guide for Postgraduate Students*, pages 12–26, 2006.

Lothar Sachs. *Applied statistics: a handbook of techniques*. Springer Science & Business Media, 2012.

Maria Spante, Sylvana Sofkova Hashemi, Mona Lundin, and Anne Algers. Digital competence and digital literacy in higher education research: Systematic review of concept use. *Cogent Education*, 5(1):1519143, January 2018. doi: 10.1080/2331186X.2018.1519143. Publisher: Cogent OA.

Dara Tafazoli, Ma Elena Gomez Parra, and Cristina A Huertas Abril. Computer literacy: Sine qua non for digital age of language learning & teaching. *Theory and Practice in Language Studies*, 7(9):716, 2017. ISSN 1799-2591. doi: 10.17507/tpls.0709.02. Publisher: Academy Publication Co., Ltd.

Robert S. Tannenbaum. How Should We Teach Computing to Humanists? *Computers and the Humanities*, 21(4):217–225, 1987. ISSN 00104817. URL http://www.jstor.org/stable/30207392.

Rob Timans, Paul Wouter, and Johan Heilbron. Mixed Methods Research: What It Is and What It Could Be. *Theory and Society*, 48:193–216, 2019.

Joris J. van Zundert and Tara L. Andrews. Qu'est-ce qu'un texte numérique? a New Rationale for the Digital Representation of Text. *Digital Scholarship in the Humanities*, 32(suppl_2):ii78–ii88, 08 2017. ISSN 2055-7671. doi: 10.1093/llc/fqx039. URL https://doi.org/10.1093/llc/fqx039.

Joris J. Van Zundert, Smiljana Antonijević, and Tara L. Andrews. 'blackboxes' and true colour — a rhetoric of scholarly code. In Jennifer Edmond, editor, *Digital Technology and the Practices of Humanities Research*, page 123–162. Cambridge, UK: Open Book Publishers, 2020. URL https://www.openbookpublishers.com/product/1108.

Edward Vanhoutte, Julianne Nyhan, and Melissa Terras, editors. *Defining Digital Humanities: a Reader*. Ashgate Publishing, Ltd., 2013.

Annette Vee. Understanding Computer Programming as a Literacy. *Literacy in Composition Studies*, pages 42–64, 2013.

Annette Vee. *Coding Literacy: How Computer Programming Is Changing Writing*. Software Studies. The MIT Press, Cambridge, July 2017. ISBN 978-0-262-03624-5.

Staten W. Webster and Linda S. Webster. Computer Literacy or Competency? *Teacher Education Quarterly*, pages 1–7, 1985. URL https://www.jstor.org/stable/23474573.

Barbara M. Wildemuth. 14. sampling for extensive studies. In Barbara M. Wildemuth, editor, *Applications of Social Research Methods to Questions in Information and Library Science*, pages 116–128. ABC-CLIO, 2009.

Robert M. Zemsky. Numbers and history: The dilemma of measurement. *Computers and the Humanities*, 4(1):31–40, 1969. ISSN 00104817. URL http://www.jstor.org/stable/30199320.

6 Appendix A - Code Literacy Definition Aspects

Code	*Description*	*Scope note*
LOC	Level of Competence	Different levels of competence of interacting with code, from recognising, reading and basic writing, to increasingly complex aspects of these. Competencies thereby have different levels.
LOC-1	Recognize	The ability to recognize code as code, having an understanding of the general purpose of code, e.g. that code is used to tell a processor what actions to perform (processing) or to structure the content of a document (encoding)
LOC-2	Read and apply	The ability to understand the syntax of a coding language and to read a specific bit of code and figuring out what it does or what it conveys. This can include understanding data structures (like arrays, hashes, ...), databases, api's, and control-flow aspects like loops and conditionals. This also includes knowing how to apply it, e.g. by changing a few parameters or variables.
LOC-3	Write	The ability to write syntactically correct code in a specific language (processing or encoding). Correct means that it can be executed without error, but says nothing about its quality or organisation. Use this when a definition only speaks of 'writing code' without specification. We assume this level is also applied in definitions that say something like 'knowing at least one programming language'.
LOC-4	Repurpose (copy-paste/libraries), edit/modify	The ability to identify a relevant piece of existing code or code libraries and incorporating it in ones own code, or adjusting it to ones own purpose and context (beyond changing a few parameters or variables and running the repurposed code as a tool).
LOC-5	Review/evaluate	The ability to review code to decide if it corresponds to its creator's intended use and purpose, the ability to evaluate the quality of code
LOC-6	Create, test, improve and deploy	The ability to create code from scratch to solve a concrete task (either to process data or to encoding documents)
LOC-7	Paradigms and theoretical aspects of computation	The understanding of various programming paradigms and how they differ in terms of modelling and extending core programming concepts. This refers to aspects like the differences between object-oriented and functional programming, or declarative versus procedural. Theory of computation includes references to computability, P vs. NP complete, ...
LOC-M	Different levels of literacy	Use this for answers that explicitly refer to different levels of code literacy. For example: "basic literacy is being able to read and understand code, intermediate literacy is being able to write code, advanced literacy is being to write high quality code."

Table 2: Code literacy definition aspects related to Level of Competence.

Code	*Description*	*Scope note*
CU	Contextual Understanding	Understanding that code is not created or used in a vacuum, but is used within a wider context of the research it is part of, cultural attitudes and conventions and other technological aspects like environments and platforms.
CU-1	Transforming research problems to computation	Understanding how a research question or problem can be divided into increasingly smaller questions or problems, which can be more directly addressed through code or software. People mentioning this aspect might focus more on code or on the research question.
CU-2	Potential, limits, biases	Understanding how specific technologies and tools are relevant to a research problem, what they can and cannot do, when it's appropriate to use them, and how they can create, propagate or exacerbate biases in the data. This is not about social/cultural aspects of code, but about the context of the coder and their intention.
CU-3	Attitude	"The attitude towards learning to code and using and creating code, as well as to its value for and role in research. This includes things like being open-minded and confident in accepting that you maybe don't know what you're doing or that you make a lot of mistakes. This also includes attitudes around autonomy and ownership as well as attitudes towards finding help (e.g. don't be afraid to ask for help or copy from others)."
CU-4	Code ecosystem (ethics, security, privacy, maintenance, documentation, versioning, licensing, good practices)	The understanding and ability to handle aspects of ethics and privacy, security, to maintain, document and version of code, understanding licensing and open-source aspects. Knowing about the software ecosystem around code (code editors, repositories) and knowing how the web works (servers, domains, sites, etc). Knowing how to find help (in documentation but also on forums and via colleagues).
CU-5	Other references to context	Any aspect of code context that is not covered by the first four aspects
CT	Code Type	Whether the code explicitly refers to a specific type of code, either code as encoding of documents or as in processing of data.
CT-1	Encoding	A form of coding that segments and structures the content of a document (be it text, image, audiovisual or data) and identifies elements of interest
CT-2	Processing	A form of coding that is performative, in the form of processing instructions that can be run to perform some task
COM	Communication	Being able to communicate about code and collaborate with others through code (includes a.o. code sharing, pair programming, co-designing specifications, co-designing a methodology), this could also include teaching and explaining of code to others.
Other	Other aspects	Anything not related to code literacy
MIS	Misinterpreted	Misinterpretation of the question. For answers where it's clear they answered a different question or least give no definition.
NA	Not Applicable	For answers that are not captured by any of the codes

Table 3: Code literacy definition aspects related to Contextual Understanding, Code Type, Communication and aspects not related to code literacy.

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

Introducing the DHARPA Project: An Interdisciplinary Lab to Enable Critical DH Practice

Angela R. Cunningham[1], Helena Jaskov[1], Sean Takats[1], Lorella Viola[1], and with contributions from Markus Binsteiner and Mariella de Crouy-Chanel[1]

[1]Luxembourg Centre for Contemporary and Digital History (C^2DH), University of Luxembourg

In this article, we introduce software under development by the Digital History Advanced Research Projects Accelerator (DHARPA), an interdisciplinary team of researchers and developers working to enable best practices in the humanities through technology. We argue that the strength and appeal of historical inquiry lies largely in the relationship between scholars and their sources, a connection in which the former engage with the latter through critical assessment, contextualisation and documentation. However, we also contend that these symbiotic processes themselves need to be evaluated and recorded. While digital tools and techniques have been accused of alienating historians from their materials, critically informed digital methodologies can also fortify and extend this relationship. To that end, we are building software that will enable users to not only apply best practices to their sources but also to allow them to write the history of those interactions, thus providing material for self-reflection and critique. In this article, having laid out our goal, we describe how the modular and data-centric design of our software's backend and the interactive documentary capabilities of its frontend operationalize a critical epistemology centered on the scholar-source relationship. We continue with a discussion of our team's internal dynamics in creating this software, and conclude with an invitation to readers to contribute to the process through commentary and testing.

Keywords: digital critical literacy, encoding criticism, digital transparency, digital traceability; reproducibility

1 Introduction

At the level of research outputs, the Digital History Advanced Research Projects Accelerator (DHARPA) responds to the ways that technology has already reshaped

historians' work by developing tools and methods. But more broadly speaking, the project's conception and execution draw on and contribute to long-running debates surrounding the digital humanities (DH). *The Chronicle of Higher Education* offers an insight into the violent nature of such debates by characterizing the dispute between the proponents and opponents of the DH as a "war" (Da, 2019). While introducing Nan Z. Da's scathing assessment of the application of computational methods to literary interpretation, *The Chronicle* also resurfaced past articles across nearly a decade that catalogued the promises and the pitfalls of DH (Brennan, 2017; Conrad, 2014; Fitzpatrick, 2011). The critiques raised in these articles and others are now, sadly, quite familiar.[1] It has been argued that in the flood of digital data and techniques, humanists have lost their way, blinded by the spectacle of technology and the glittering grant money attached to the latest academic trend (Allington et al., 2016). Even DH proponents have suggested that a surfeit of data and push-button tools have led to a false sense of mastery (Tenen, 2016; Weingart, 2011),[2] the unproblematized (mis)application of methodologies that originated in other disciplines (Drucker, 2011), and stolen or abandoned scholarly agency (Grimshaw, 2018; Liu, 2012). Poetic metaphors like "black boxes" and "mechanical turks"[3] evoke a naïve and starry-eyed reliance on poorly understood computer-enabled techniques (Diakopoulos, 2014; Noble, 2018; Pasquale, 2016): at best, critics charge, digital humanists have allowed themselves to become dependent on technologies of which they cannot or will not examine the inner workings, or, worse, they have let themselves believe in the specious autonomy of the machine (Tenen, 2016). In short, digital intermediation has broken the essential connection between scholars and their sources, a loss that can feel profoundly dehumanizing.

We all know the critiques, and many of us have been unsettled by the rupture. But how do we as humanists take back ownership of and responsibility for the work that we do increasingly – and some would say unavoidably – with digital data, digital tools and digital methodologies? How do we shine a light on the interior workings of technological and intellectual black boxes? This article introduces the software currently under development by the Digital History Advanced Research Projects Accelerator (DHARPA), an interdisciplinary team working to address these issues.[4] While DHARPA is multifaceted, with an interest in collaboration, pedagogy, and infrastructure, here we will focus on the software at the foundation of our project, and particularly on aspects of that software which speak to the need to (re)build a critically-aware connection between scholar and sources in a digital world. In the following section, we will lay out our ideas about DH best practice. Then, in describing our "Virtual Research Environment" we will outline how we are operationalizing

[1] In this article we will focus largely on epistemological issues. For more discussions on the economic and political issues that are associated with the digital humanities as an institutional entity, see the various contributions to the C21 conference under the telling header "The Dark Side of Digital Humanities" (Chun, 2013; Jagoda, 2013; Raley, 2013).

[2] Weingart (2011) critiques the push-button mentality with popular network analysis software, while Tenen (2016) observes that "some tools encourage intellectual laziness by obscuring methodology". Tenen, however, also points out that "more often, it is not the tool but rather a mode of lazy thinking that is at fault".

[3] The Mechanical Turk was a clever illusion, devised in the late 18th century and exhibited well into the 19th, which was composed of a mannequin dressed in traditional Ottoman costume and ersatz machinery in full view and, hidden from the audience, the human chess master who was in truth responsible for winning the majority of games the apparatus played against sundry famous opponents (Wikipedia contributors, 2022).

[4] The DHARPA project has been generously funded by the Luxembourg National Research Fund, grant 13307816. https://www.fnr.lu/projects/digital-history-advanced-research-projects-accelerator/.

these ideas through a modular and data-centric backend and an interactive frontend capable of documenting scholarly processes. Through this work, we hope to turn the digital into an opportunity not only to regain but also deepen our relationship with sources through technology.

2 Enabling and encouraging best (digital) humanities and history practice

History as a discipline is already home to a set of best practices. No historian worthy of the name would think of *not* critiquing their sources in light of their content and the context in which they developed, nor would any good historian fail to take copious notes or keep track of citations. Yet, the actual enactment of this best practice often goes unrecorded. Historians rarely document their research methods and are often reluctant to share their experience on how they identified and collected their sources in the first place (Beals, 2013; Putnam, 2016). Among many historians, there is often little or no documentation of the scholarly engagement that turned their sources into data nor of how they used these data to produce analyses and obtain insights (Viola and Fiscarelli, 2021). A similar charge may be leveled against practitioners of other humanities (and indeed of scientific and technological disciplines), where the influence of the researcher on the research is rarely acknowledged (Marcus, 2021). On the other side of the digital scholarship/traditional scholarship divide, some advocates for computer-supported methodologies laud technology's ability to make such scholarly engagement with sources more dynamic, with interactive on-the-fly functionality allowing the scholar more direct and exploratory involvement with their sources (Travis, 2015). Yet the outputs of this sort of playful exploration are likewise often unmoored from any accounting of the processes that created them. In both instances, the traditional and the digital, there is rarely any narrative of the researcher's interventions, any material for self-reflection, nor much evidence to counter the depiction of the humanities as esoteric. Arguably, both digital and traditional ways of doing history and humanities research may be accused of being like the Mechanical Turk, with the decisions and actions made by the researcher hidden from view and only the well-oiled and seemingly autonomous product – the article, the book, the conference presentation – on display.

In the unfortunate tendencies to simply "push the execute button" or to present scholarly work as the product of similarly automatic and obvious processes lie other dangers. Disassociating the humanities scholar, scientist or programmer from the technologies that they build and employ downplays the need to understand the cultural, philosophical and historical contexts in which digital tools and operations originated (Dobson, 2019; Hitchcock, 2013): this is akin to studying history without regard to historiography. In portraying computational and computer-aided methods as objective, unassailable, and observer- or user-independent – an attitude that Daston (1992) has described as a belief in "instrumental objectivity" – we risk perpetuating "the god trick" (Haraway, 1988; Sheppard, 2001), and thus supporting objectivist and positivist understandings that are 'fundamentally at odds with approaches to humanities scholarship premised on constructivist principles' (Drucker, 2011). As we find ourselves in an increasingly digital world, it is all the more important to incorporate critical theory into the digital (Berry and Fagerjord, 2017).

At the heart of DHARPA is the desire to encourage and enable good historical and humanities practice – or indeed the good practice of other disciplines that a scholar may wish to call upon – but also to extend that best practice through technology. We

argue that we need to document and critique our sources, but also our engagement with them. For example, how to best record a researcher's thought process? What literature were they thinking about when they chose to alter or analyse the data in a particular way? We argue that digital humanists should be able to keep track of when they set processes in motion, because scholarship is never finished, always evolving, and knowing when we thought about something and in what order can help us and others re-trace and re-evaluate our work. DHARPA aims to encourage self-reflective appreciation of how the application of expertise can work in tandem with technology to produce knowledge, and through documenting academic labour, to make its value evident, to make it transparent, and to keep it honest and accountable. Digital humanists, we contend, should be able not only to look inside the box to see its inner workings, but to place themselves as the expert inside the machine.

3 DHARPA- practicing critical digital humanities through interdisciplinary collaboration

Writing in an era of boundary-drawing, Liu (2012) called on digital humanists 'to start reflecting on the wider net of relations of digital research'. A decade later, there is still a need not only to self-reflexively address the tools we use, the sources we apply them to, and - as we have argued here - how we work in relationship with them, but also how we work in relationship with each other. To that end, we now turn to describing our interdisciplinary team and how we have collaborated to bring the best practices of our own fields into application within the Virtual Research Environment (VRE).

Our project builds on the lessons of past digital humanities projects by operating on a longer time frame and with a team that mixes not only disciplinary expertise but also encourages collaboration across a wide range of skill sets and experience. DHARPA is supported by the Luxembourg National Research Fund under a scheme that is intended to promote high-risk, high-reward research, providing an unusually long period of performance of five years, plus ongoing institutional support beyond the initial five years. Embedded within the Luxembourg Centre for Contemporary and Digital History (C^2DH), currently the largest digital history center with 120 full-time equivalent staff, the DHARPA team is composed of four developers, six academics (including PhD students, postdoctoral researchers, and faculty), and an administrative assistant. The composition of the development team is explicitly modeled on the teams that developed Zotero, Tropy, and PressForward, all long-running and co-designed projects. Each of us brings our own expertise and opinions to the table, drawn from years of experience in computer programming, data engineering, data visualization, linguistics, geography, and various strains of history.

An important starting point for our collaboration was the collection of "user stories" from researchers. These plain-text descriptions of common operations or expectations when dealing with data, like recording and saving metadata, tracing data provenance, or validating input, allowed our developer team to infer the technical requirements that the backend was expected to provide. Throughout the process of building the VRE, we have had many discussions about the very nature of sources, how they are transformed into computer-legible data, and the ways that we in our varied positionalities interact with these data through technology (Cunningham et al., 2021) [5]. For instance, we

[5] The citation here is for a recording made of one of these discussions for presentation at an academic conference, using this polyvocal format to decenter one of us scholars as "the expert" and

have noted how problematic the term "dirty data" is, a term implying that there is a precisely recorded (and recordable) reality that we can distill if only given the right tools. Yet we have also needed to acknowledge that sources that have not been standardized and formatted in some way are useless to a block of code. Recalling the main argument of this article, we find that the scholar must intervene to make sources usable, whether as digitized data or as handwritten notes from an archival visit, and that it is best to keep a record of this manipulation. Such a realization has informed the functionality of all the elements of the DHARPA software.

As the software itself is envisioned to promote transparency and documentation in digital research, so is its development and the decision-making process meticulously documented. A small part of this is already publicly available via the project's GitHub account,[6] with more detailed publications in preparation. Browsing through this material, we can revisit our previous decisions, reflect on the work that we have accomplished, and make adjustments as we encounter challenges. For instance, an early debate that stands out in this documentation material is the question of what type of software we were going to build. Would it be a desktop app or a web application? Are we building a workflow execution tool or a workflow assembly tool? As the project continues to evolve, we have retained some flexibility on these issues by focusing first on the specification and development of a data orchestration backend. While our initial target remains a desktop application, we have also created the conditions for producing a future web app as well.

4 The Virtual Research Environment

To attain these goals, we at DHARPA are building a virtual research environment. Our software will be free and open-source, and to further encourage widespread and confident adoption the VRE will be available remotely and for downloadable local use; it will rely on a generalizable hosting infrastructure that ensures privacy and portability; and it will be supported by a long-term sustainability plan and training opportunities. The VRE will have modular design, allowing users to build and rebuild, and to run and rerun their own workflows or those they have adopted and adapted. The VRE's modules will include functionality for data ingestion, data standardization, textual analysis, network analysis, and geographical analysis, all brought within a seamless environment where work can flow between tasks from experimentation and modelling to presentation and dissemination. We anticipate that in the future this repertoire of core modules will be expanded by the community of DH scholars who would be able to create their own modules tailored to their specific research needs. The VRE will also be equipped with essential features for documenting the user's research process and tracing all the transformations which their data undergo, allowing the historian to write a self-reflexive meta-history of their relationship with their sources through the software. To describe in more detail how our VRE functions, we turn first to *kiara*, our backend, and then to *Lumy*, comprised of our frontend and the code necessary for the two pieces to communicate with each other. Figure 1 provides a simple schematic of how *kiara* and *Lumy* work together.

instead highlight the collaborative nature of our work.

[6] https://github.com/DHARPA-Project

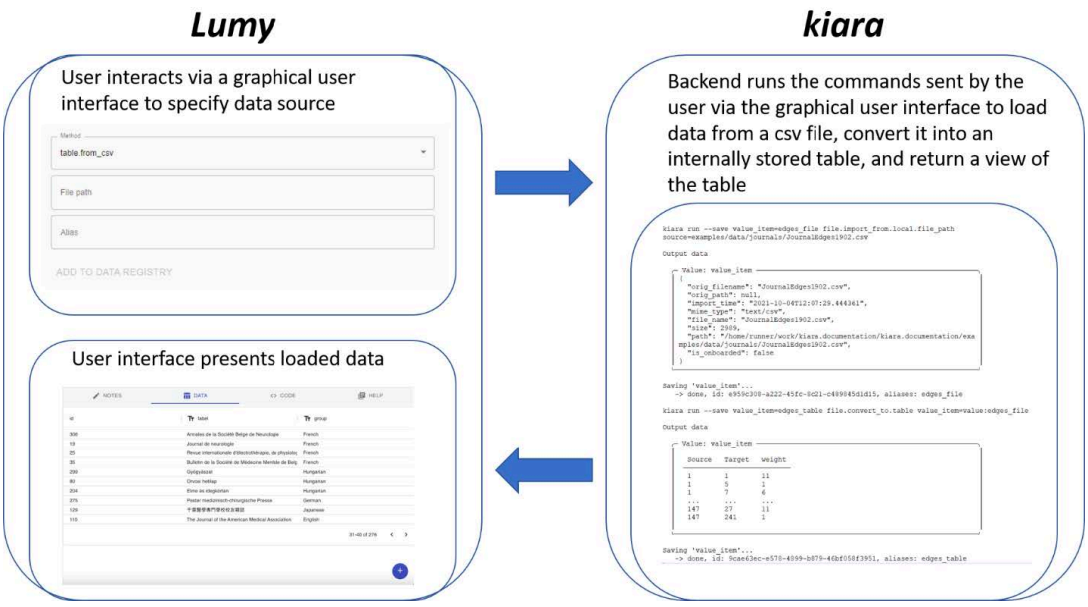


Figure 1: A schematic of how *Lumy* and *kiara* will work together with the scholar/user to complete an important task: bringing sources into the system and transforming them into computer-legible data.

4.1 kiara: backend modularity and a data-centric approach to encourage engaged exploration

Our VRE is built on top a backend, *kiara*, comprised of modules developed on the basis of collaboration between researchers and developers. *kiara* has been inspired by a variety of data pipeline and orchestration tools like airflow (`https://airflow.apache.org/`), prefect (`https://www.prefect.io/`), dagster (`https://dagster.io/`), snakemake (`https://snakemake.readthedocs.io/en/stable/`) and kedro (`https://kedro.readthedocs.io/en/stable/`).[7] What sets *kiara* apart from these tools is its strong focus on metadata management and interactive workflow execution. While most tools feature batch processing, *kiara* offers an interactive approach in addition to batch processing. This allows users to interact with their data at crucial steps within the workflow and to take control of these decisions in an interactive and user-friendly environment. *kiara* is a Python-based tool that can be navigated via the command line. Anyone with basic familiarity with Python can adapt existing Python code to *kiara* modules which can then be shared with other researchers. As shown in Figure 2, a *kiara* user obtains a list of available modules for a specific method by using the `<kiara operation list>` command and adding a keyword. In the following example, *kiara* displays the modules that belong to the network analysis group.

[7] For a survey of this software, see Ruf et al. (2021) as well as Matskin et al. (2021). Other comparable tools specializing in scalable workflow execution and data management are Ray (`https://docs.ray.io`) and Dask (`https://dask.org/`).

All operations

Id	Type(s)	Description
network_graph.add_centrality_calculations		*Add centrality properties to a graph.*
network_graph.add_nodes		*Add nodes to an existing graph.*
network_graph.analyze		*First iteration of a network analysis pipeline. This doesn't do much yet.*
network_graph.augment		*Augment an existing graph with node attributes.*
network_graph.components		*Extract component information from a graph.*
network_graph.extract_metadata.network_graph	extract_metadata	*Extract 'network_graph' for value of type 'network_graph'.*
network_graph.extract_metadata.python_class	extract_metadata	*Extract 'python_class' for value of type 'network_graph'.*
network_graph.find_shortest_path		*Find the shortest path between two nodes in a network graph.*
network_graph.from_edges_table		*Create a network graph object from table data.*

Figure 2: An excerpt from *kiara*'s collection of network analysis modules and their short descriptions. This output can be generated by executing the `<kiara list operations network>` command. A module-creating user can modify the description field and create categories within the "type" field to facilitate the search for desired module.

A *kiara* module consists of a metadata block and three core sections: input, output, and process. The input section specifies any required input, providing a short description of the expected properties in order to help users make informed decisions about the use of the module. The `<add_nodes>` module shown in the list above, for example, requires three input items: a graph object (the graph you wish to augment), a table (the nodes table from which you wish to extract the new nodes), and a string (the name of the column that holds the index of the nodes). The output section defines the expected output (in our example: the new augmented graph object), and the process section holds the Python code that is required to execute the module. Because *kiara* compartmentalizes into separate modules the workflow steps that would usually appear in a single Python script, they can be rearranged and chained together to create a range of workflow pipelines according to the researcher's needs.

Apart from the advantage of offering a flexible solution with a potential for further customization, this data-centric, modular approach also ensures a full recording and documentation of research data's creation and usage. In order to be able to keep track of the various transformations that the data undergoes in such a setting, *kiara* provides the ability to trace data ancestry. In the following example from the network analysis module, the user has requested the lineage of `<journals_graph>` (via the command `<kiara data explain --lineage journals_graph>`), and has received a record of the modules and inputs used to generate that graph. As can be inferred from Figure 3, the graph has been created from two different data sources. It has first been generated from an edges table and afterwards augmented with additional nodes and node attributes from a separate nodes table via the `<add_nodes>` module.

```
Value: journals_graph

id          e5c9af08-edd2-4fcf-b0a8-df3f79a1725e
type        network_graph
desc        The network graph
is set      yes
metadata    network_graph
              graph_type  directed
              number_of_nodes  276
              number_of_edges  321
              density  0.0042292490118577076
            python_class
              class_name  DiGraph
              module_name  networkx.classes.digraph
              full_name  networkx.classes.digraph.DiGraph


lineage     network_graph.add_nodes
            ├── input: graph (network_graph)
            │   └── network_graph.from_edges_table
            │       ├── input: edges_table (table)
            │       │   └── table.convert
            │       │       └── input: value_item (file)
            │       │           └── file.import
            │       │               ├── input: aliases (list)
            │       │               ├── input: save (boolean)
            │       │               └── input: source (file_path)
            │       ├── input: graph_type (string)
            │       ├── input: source_column (string)
            │       ├── input: target_column (string)
            │       └── input: weight_column (string)
            ├── input: index_column_name (string)
            └── input: node_attributes (table)
                └── table.convert
                    └── input: value_item (file)
                        └── file.import
                            ├── input: aliases (list)
                            ├── input: save (boolean)
                            └── input: source (file_path)
```

Figure 3: An example of *kiara*'s ability to record and recall the ancestry of a data value. In this example from the network analysis module, the user has requested the lineage of `<journals_graph>`, a directed graph of the citation connections among scientific journals, and has received a record of the modules and inputs used to calculate that graph.

kiara can also be used in combination with the open-source tool Streamlit (https://streamlit.io/) which opens up new possibilities of rapid app development and has the benefit of making *kiara* accessible within a community based software environment. Streamlit offers several advantages: on the one hand, it accelerates workflow prototyping, and on the other hand it can function as a handy tool to assist *kiara* module developers and the broader potential community of *kiara* contributors to create their own modules and pipelines. Streamlit enables contributors, who are familiar or beginners with Python, and who are not familiar with JavaScript, to very easily create user interface (UI) elements for *kiara* workflow steps. Streamlit's ease-of-use and flexibility provide module contributors with sufficient freedom on the arrangement of UI interactive and narrative components thereby enabling researchers to embed criticism into the UI. Moreover, Streamlit also makes it easy to create auto-generated UI elements linked to *kiara* modules, that can act as helpers for *kiara* module contributors, to facilitate developing and testing data science back-end processes. Such a development process becomes consequently more widely accessible to the broader community.

4.2 Lumy: frontend interactivity and documentation to promote holistic humanities practice

Much as *kiara* offers the ability to engage with data iteratively while helping to keep track of the history of those interactions, so too will *Lumy* encourage users to adopt

recursive, self-reflexive and reproducible work practices. Figure 4 presents a wireframe mock-up of what *Lumy*'s interface will look like for one part of the network analysis module. On the left side of the panel, the user will be able to interact with tabular data and interconnected visualizations simultaneously. On the right side of the panel, *Lumy* will pair the ability to record what the software is doing with the ability to record what the scholar is doing, cultivating the holistic practice of interweaving methodology, data, code, and metadata with a narrative of researcher's choices and actions. The user will be able to both snapshot a list of inputs and current parameters (all of which might be changed in a future iteration), and take their own notes, through a panel that will also be able to accommodate links, images, and references to relevant secondary literature.

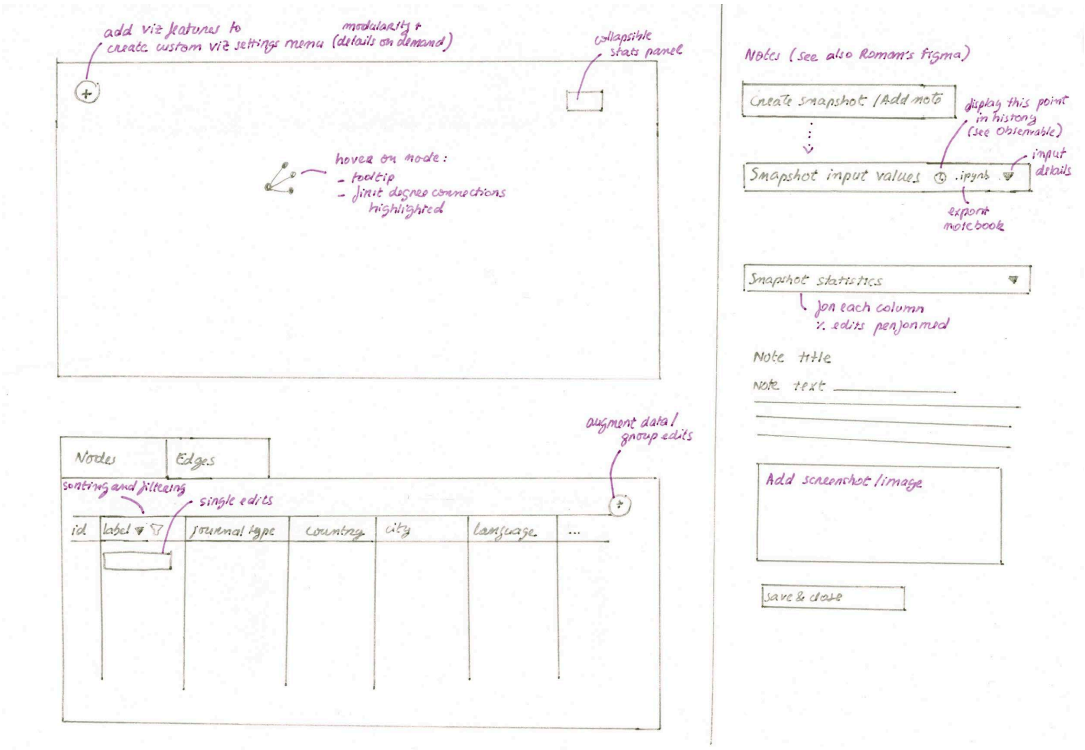


Figure 4: A wireframe outlining *Lumy*'s interface for network analysis. A visualization panel appears at top left and a data panel with tabular data (here separated into a node and an edge table) at bottom left, panels which will be connected to each other and interactive. To the right, the wireframe depicts buttons for taking a snapshot of the current parameters, taking narrative-style notes, and allowing the export of code to an html "codeview" document or a runnable Python notebook. The note panel will also allow the user to save screenshots and relevant citations. Buttons on other screens will also allow the user to view other citations and references that informed the implementation of the underlying methods.

For an overview of how different modules and submodules have been built into the larger scholarly process that the user has enacted, we will provide two functionalities. First, it will be possible to download and export a record of how the data in their current state were produced, as an immutable html document and as an editable and independently runnable Jupyter Notebook. Both the html "codeview" and Jupyter Notebook formats will document the interaction of scholar and source, as code plus notes, as machine process (via lineage tracing, metadata management, code view)

plus thought process (via note-taking function and annotation features), and allow this interaction to be reproduced to create the same output. Second, a more complex and iterative accounting of scholarly engagement will be captured and accessible in a work history pane in *Lumy*. This work history will allow the user to see the paths they have been on – including both those that have been abandoned and those that have been further pursued – and other paths that they could try. All the pieces of the relationship among scholar, source and software will again be preserved *in situ*, in a form that one will be able to review, change or revert to later as their thinking evolves. It will allow a researcher – whether the original user or another – to review former choices and decisions in their entire complexity.

To aid the user in making informed decisions, tools available in *Lumy* will be presented with informational popups describing how the tools were developed and tips for their usage with reference to established literature and techniques. This will help fulfill DHARPA's pedagogical intent to encourage best application of data engineering and textual, spatial or network analytical methodologies even among those who might be unfamiliar with them. It is also just one concrete example of how the VRE is drawing on the diversity of the project team.

5 Onwards

In the spirit of viewing scholarship as an iterative and never ending process rather than a product, we end our article not with a conclusion but a view towards the future of our project. We are looking forward to the moment when we will be able to run workshops to introduce students and our internal and external colleagues to our philosophy of technology-enabled and enhanced critical digital humanities scholarship. As proposed in the days of the original grant application, we aim to position our team into a rapid response lab and extend *Lumy* and *kiara*'s capabilities through new applications proposed by other scholars in the humanities and social sciences employing both qualitative and quantitative methodologies.

In the coming months, DHARPA will release a beta version of our initial modules. Although our frontend, backend and the connections between them are still works in progress, we alert the reader that portions of our code and documentation are already accessible via our GitHub account `https://github.com/DHARPA-Project`. The modularity of our software will allow external developers to build their own modules and workflows to extend *kiara* and *Lumy*. Just as good scholarship relies on peer review and constructive criticism, so too will our open-source VRE, therefore we encourage you to get in touch, share your ideas, and help us advance this project of shining a light on the too-often hidden cogs and gears of humanities research.

References

Daniel Allington, Sarah Brouillette, and David Golumbia. Neoliberal Tools (and Archives): A Political History of Digital Humanities. *Los Angeles Review of Books*, May 2016. URL `https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/`.

M H Beals. Record How You Search, Not Just What You Find: Thoughtfully Constructed Search Terms Greatly Enhance the Reliability of Digital Research,

June 2013. URL https://blogs.lse.ac.uk/impactofsocialsciences/2013/06/10/record-how-you-search-not-just-what-you-find/.

David M. Berry and Anders Fagerjord. *Digital Humanities: Knowledge and Critique in a Digital Age*. Polity, Cambridge, England, June 2017. ISBN 978-0-7456-9766-6.

Timothy Brennan. The Digital-Humanities Bust. *The Chronicle of Higher Education*, October 2017. ISSN 0009-5982. URL http://www.chronicle.com/article/The-Digital-Humanities-Bust/241424.

Wendy Hui Kyong Chun. The Dark Side of the Digital Humanities – Part 1, January 2013. URL https://www.c21uwm.com/2013/01/09/the-dark-side-of-the-digital-humanities-part-1/.

Kathryn Conrad. What the Digital Humanities Can't Do. *The Chronicle of Higher Education*, September 2014. URL http://www.chronicle.com/article/what-the-digital-humanities-cant-do/.

Angela R Cunningham, Markus Binsteiner, Helena Jaskov, and Lorella Viola. "Handmaiden to history"? A conversation on geography in a digital history center. In *American Association of Geographers Annual Conference*, virtual, 2021.

Nan Z. Da. The Digital Humanities Debacle. *The Chronicle of Higher Education*, 65(29), April 2019. URL https://www.chronicle.com/article/the-digital-humanities-debacle/.

Lorraine Daston. Objectivity and the escape from perspective. *Social Studies of Science*, 22(4):597–618, 1992. ISSN 0306-3127. URL https://www.jstor.org/stable/285456.

Nicholas Diakopoulos. Algorithmic accountability reporting: On the investigation of black boxes. Technical report, Tow Center for Digital Journalism, 2014. URL https://doi.org/10.7916/D8ZK5TW2.

James E. Dobson. *Critical Digital Humanities: The Search for a Methodology*. Topics in the Digital Humanities. University of Illinois Press, Urbana, 2019. ISBN 978-0-252-05111-1.

Johanna Drucker. Humanities approaches to graphical display. *Digital Humanities Quarterly*, 005(1), March 2011. ISSN 1938-4122.

Kathleen Fitzpatrick. The Humanities, Done Digitally. *The Chronicle of Higher Education*, May 2011. URL http://www.chronicle.com/article/the-humanities-done-digitally/.

Mike Grimshaw. Towards a manifesto for a critical digital humanities: Critiquing the extractive capitalism of digital society. *Palgrave Communications*, 4(1):1–8, February 2018. ISSN 2055-1045. doi: 10.1057/s41599-018-0075-y. URL http://www.nature.com/articles/s41599-018-0075-y.

Donna Haraway. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575–599, 1988. ISSN 0046-3663. doi: 10.2307/3178066. URL http://www.jstor.org/stable/3178066.

Tim Hitchcock. Confronting the Digital. *Cultural and Social History*, 10(1):9–23, March 2013. ISSN 1478-0038. doi: 10.2752/147800413X13515292098070. URL https://doi.org/10.2752/147800413X13515292098070.

Patrick Jagoda. The Dark Side of the Digital Humanities – Part 3, January 2013. URL https://www.c21uwm.com/2013/01/09/the-dark-side-of-the-digital-humanities-part-3/.

Alan Liu. Where Is Cultural Criticism in the Digital Humanities? In Matthew K. Gold, editor, *Debates in the Digital Humanities*, pages 490–509. University of Minnesota Press, Minneapolis, 2012. URL https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/896742e7-5218-42c5-89b0-0c3c75682a2f.

Miranda Marcus. The God Trick in Data and Design, July 2021. URL https://towardsdatascience.com/the-god-trick-in-data-and-design-4ec71e19811.

Mihhail Matskin, Shirin Tahmasebi, Amirhossein Layegh, Amir H Payberah, Aleena Thomas, Nikolay Nikolov, and Dumitru Roman. A Survey of Big Data Pipeline Orchestration Tools from the Perspective of the DataCloud Project. In *International Conference on Data Analytics and Management in Data-Intensive Domains*, page 16, 2021.

Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York, 2018. ISBN 978-1-4798-4994-9 978-1-4798-3724-3.

Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, Cambridge, Massachusetts London, England, first harvard university press paperback edition edition, 2016. ISBN 978-0-674-97084-7 978-0-674-36827-9.

Lara Putnam. The transnational and the text-searchable: Digitized sources and the shadows they cast. *The American Historical Review*, 121(2):377–402, April 2016. ISSN 0002-8762. doi: 10.1093/ahr/121.2.377. URL https://doi.org/10.1093/ahr/121.2.377.

Rita Raley. The Dark Side of the Digital Humanities – Part 4, January 2013. URL https://www.c21uwm.com/2013/01/09/the-dark-side-of-the-digital-humanities-part-4/.

Philipp Ruf, Manav Madan, Christoph Reich, and Djaffar Ould-Abdeslam. Demystifying MLOps and Presenting a Recipe for the Selection of Open-Source Tools. *Applied Sciences*, 11(19):8861, 2021. doi: 10.3390/app11198861.

Eric Sheppard. Quantitative Geography: Representations, Practices, and Possibilities. *Environment and Planning D: Society and Space*, 19(5):535–554, October 2001. ISSN 0263-7758. doi: 10.1068/d307. URL https://doi.org/10.1068/d307.

Dennis Tenen. Blunt Instrumentalism: On Tools and Methods. In Matthew K Gold and Lauren F Klein, editors, *Debates in the Digital Humanities 2016*. University of Minnesota Press, Minneapolis, 2016. URL https://dhdebates.gc.cuny.edu/read/untitled/section/09605ba7-ca68-473d-b5a4-c58528f42619#ch09.

Charles B. Travis. *Abstract Machine: Humanities GIS*. Esri Press, Redlands, 2015. ISBN 978-1-58948-368-2.

Lorella Viola and Antonio Fiscarelli. From digitised sources to digital data: Behind the scenes of (critically) enriching a digital heritage collection. In A Webber, M Heerlien, E.G. Miracle, and K Wolstencroft, editors, *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, 2021. URL http://ceur-ws.org/Vol-2810/paper5.pdf.

Scott Weingart. Demystifying networks, parts I & II. *Journal of Digital Humanities*, 1(1), 2011. URL http://journalofdigitalhumanities.org/1-1/demystifying-networks-by-scott-weingart/.

Wikipedia contributors. Mechanical Turk. *Wikipedia*, May 2022. URL https://en.wikipedia.org/w/index.php?title=Mechanical_Turk&oldid=1087252650.

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

Examining a Multi Layered Approach for Classification of OCR Quality without Ground Truth

Mirjam Cuper

KB, national library of the Netherlands

While the digital availability of heritage text collections is increasing, there is a lack of reliable methods to assess the OCR quality of these texts. There are several possible measures, but each has its disadvantages. We examined if, instead of a single measure, a *combination* of measures would give a more accurate indication of OCR quality. We therefore built a first version of a multi-layered approach for the classification of OCR quality, named QuPipe. We tested QuPipe on a set of sentences from 17th century Dutch newspapers and found that using QuPipe led to an increase in correctly classifying the quality. However, although these results are positive, QuPipe needs to be developed and tested further to become feasible as an instrument to classify the OCR quality.

Keywords: Optical Character Recognition (OCR); OCR quality; digital heritage; digital humanities; digitized texts

1 Introduction

In the past few decades, more and more heritage institutions have made their collections digitally available. At the same time, the accessibility and amount of computer driven research tasks is increasing. With this combination, large data sets can be analysed in a fairly short time, something which would not be possible by hand. However, there is a pitfall; the Optical Character Recognition (OCR) quality of digitised text is not always high enough. This leads to several possible problems which can cause bias in the research results, both on the information retrieval and the analysis level (Nguyen, 2020). Although most researchers are aware of the presence of OCR errors, often they are not able to quantify these errors or the impact on their research. This leads to uncertainty about whether results can be published or not (Traub et al., 2015).

A measure for the (relative) quality of OCR would be very beneficial for the field of Digital Humanities. Furthermore, digital heritage institutions can use such a measure to improve their digitised collections. To get an indication of the OCR quality of a collection, two methods are most prevalent. The first method consist of extracting a sample from a batch of digitised text and manually inspecting the quality of this

sample. The results of this quality control are then extrapolated to the rest of the batch. The second method is based on the existence of a 'Ground Truth'-set. A Ground Truth set consist of digitised texts that are manually corrected by humans. Therefore, these digitised texts are of high quality and contain very few or no errors. They can be used to determine the quality of the corresponding OCR output. The results can then be extrapolated to the rest of the collection. However, the creation of a Ground Truth set is time consuming and expensive (Holley, 2009), which results in only a small number of available Ground Truth sets. The extrapolations of these methods can only be considered as rough estimates and are not very reliable for a variety of reasons, such as: the variety in quality of the original material, the used font types, the text direction, and how the original material is stored and bound. This can even differ between various issues of, for example, the same newspaper.

Due to the unreliability of above methods and the scarcity of Ground Truth sets, various other methods have been developed to determine the quality of OCR-ed text without the existence of a Ground Truth. Previous research has mentioned methods such as a dictionary lookup (Strange et al., 2014; Van Strien. et al., 2020), garbage detection (Kulp and Kontostathis, 2007; Taghva et al., 2001), and confidence values from the OCR engine (Holley, 2009; Springmann et al., 2016). Some of these measures are more accurate than others, but they all have problems with their accuracy due to the nature of language and problems in digitisation. In addition, some measures are not always available, such as a historical dictionary or confidence values.

In an attempt to overcome these problems, we developed a first version of a multi-layered approach that combines measures: a quality pipeline named QuPipe. We included statistical measures for word length and sentence length. Furthermore, we added more complex measures: language detection, garbage detection, trigrams and two different approaches of dictionary lookup. Since every language and time period has its own characteristics, we created a set with reference values which we evaluated the measures against.

We tested our first version of QuPipe on a controlled set of sentences from 17th century newspaper articles. For this set, both the original OCR and the Ground Truth are available. The Ground Truth was used to calculate the Character Error Rate (CER) for every sentence. This CER was used to calculate the precision and recall of our experiments. Historical dictionaries are not always available for historical languages, therefore, we tested our approach with and without a dictionary. We also examined the minimum amount of Ground Truth needed to create useful reference values. With our experiments, we test the following hypothesis:

"A combination of measures is more accurate for the prediction of OCR quality than just a single measure."

This paper is structured as follows. In section 2, we examine related work in which various quality measures are described in more detail. Section 3 explains our method, including the used data, implemented measures, creation of the reference sets, and the QuPipe scores calculations. Section 4 shows the results of our experiments. Section 5 contains our conclusion, in which we reflect on our hypothesis and the advantages and disadvantages of our study. This section also contains our plans for future work.

2 Related work

A commonly used method for measuring OCR quality is a dictionary lookup, where every word in a text is matched against a dictionary (Strange et al., 2014; Van Strien.

et al., 2020). Then, the total amount of found words is divided by the total amount of words to get an indication of the quality of the OCR-ed text. The higher the result, the better the quality. Various research has been conducted to determine what can be considered as the minimum dictionary lookup value to indicate a high enough OCR quality. Strange et al. (2014) performed a sample task of finding words. Their findings were that correcting the accuracy of OCR from 80% to a higher accuracy is not essential for such a task. Van Strien. et al. (2020) performed an analysis on several Natural Language Processing tasks and concluded that a dictionary lookup score of at least 80% is preferable for these tasks. We conducted a small experiment and also concluded that a dictionary lookup is overall quite reliable, see Figure 1. However, we detected that several problems can occur, such as false positive prediction due to incomplete OCR or garbage in texts (Cuper, 2021). Strange et al. (2014) mention that problems can be caused by correct words which are not present in a dictionary, and with incorrect words in the digitised text which are correct words in real life. An example of such a real world error is the word 'grass' that has been interpreted by the OCR software as 'glass'. Van Strien. et al. (2020) mention that how languages change over time can possibly lead to difficulties with historical texts. Also, especially for older languages, a historical dictionary is not always available.

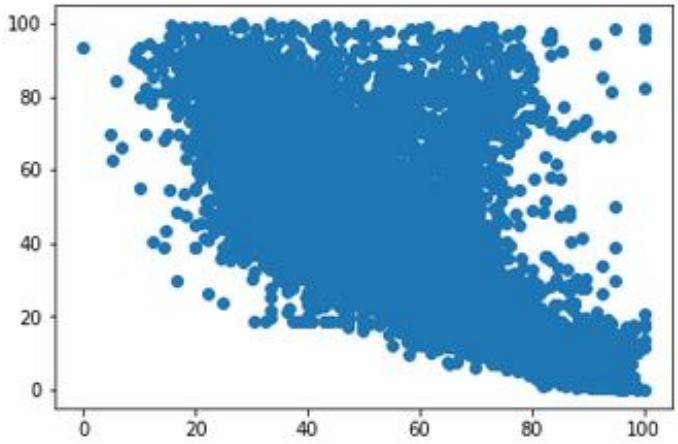


Figure 1: Correlation between CER and dictionary lookup.

A Python tool called 'Language Identifier' (Lui, 2011) was used by Baumann (2015) to measure OCR quality. He used the tool line by line and used the provided confidence score of the tool as OCR quality measure.

Holley (2009) proposed a method for the development of an accuracy algorithm based on the correlation between manually determined quality and the confidence values from the OCR engine itself. However, they did not test the algorithm themselves. Springmann et al. (2016) used character confidence values as one of the measures for OCR accuracy. A downside of the use of confidence scores is that they are not always available to researchers.

Various studies focused on the detection of errors instead of measuring OCR quality. However, if one can detect errors, this information can be used to determine the quality of a text.

Taghva et al. (2001) came across the problem of garbage strings in OCR-ed texts that were complicated for information systems. They developed a rule-based approach

to detect such garbage in texts, consisting of six rules. Kulp and Kontostathis (2007) adapted this idea, altered the rules and added two new rules. Their method is more strict, likely leading to more words that are classified as garbage. Wudtke et al. (2011) used the idea of garbage detection and implemented this idea as a support vector machine, leading to better results than the rule-based approach.

In multiple studies, different types of ngrams were used for the detection and correction of errors in texts. All these studies suggest that the use of ngrams is effective (Ahmed et al., 2009; Atawy and ElGhany, 2018; Robertson and Willett, 1998; Wu et al., 2013). Although we could not find any studies about the use of ngrams for measuring OCR quality, the use of ngrams for error detection and corrections suggests that ngrams can be a useful addition for the determination of OCR quality.

3 Method

Based on the related work and practical considerations, we created a list of five measures for our first implementation of QuPipe: language detection, garbage detection, trigrams and two different approaches of a dictionary lookup. Since every language and time period had its own characteristics, we also included the statistical measures average word length and sentence length. A detailed description of every measure is provided in section 3.2. We used part of the available Ground Truth data to create a set of reference values. Except for the language detection measure, every measure uses a reference value to determine whether the measure is in the expected range or not. Section 3.3 provides a description of the creation of these reference values.

We tested QuPipe on a dataset of 17th century newspapers. We used a controlled dataset with pairs of sentences. Each pair consisted of a Ground Truth sentence and the corresponding OCR sentence. A benefit of such a controlled dataset is that you can closely monitor the effect of changes in approach. Section 3.1 describes the preparation of the data.

Since we had the Ground Truth available, we were able to calculate the Character Error Rate (CER) of every OCR sentence. This was done with the ocrevalUAtion tool (IMPACT Centre of Competence, 2019). Holley (2009) described an accuracy of below 90% as low. Based on this, we classified our CER output as 'good' and 'not good', with a CER equal to or lower than 10 classified as 'good', and a CER higher than 10 classified as 'not good'.

The primary output of QuPipe is a binary value per measure, but for our experiments we needed a single output from QuPipe. We tested three different ways to calculate this single output: the normal calculation, the 'smart quality' calculation and the 'smart quantity' calculation. Section 3.4 describes these calculations in more detail. After calculation, QuPipe has classified every OCR sentence as either 'good' or 'not good'.

To test the performance of QuPipe, we compare the outcome of QuPipe with the calculated CER. This comparison leads to a precision and a recall for every calculation type. The precision indicates how much of the sentences that were classified as 'good' indeed had a CER score of 10 or higher, whereas the recall indicates how many of the sentences with a CER score of 10 or higher were correctly classified as 'good'.

Since historical dictionaries are not always available, we performed experiments with and without a dictionary. For the experiments with a dictionary, we compared the performances of a modern dictionary, a historical dictionary, and a combined dictionary. Furthermore, we compared QuPipe outcomes with a 'token' dictionary

lookup with a cutoff point of 80%, based on the minimum required OCR quality according to the literature (Strange et al., 2014; Van Strien. et al., 2020).

3.1 Collecting and preparing data

For our experiments, we used a dataset containing 34.808 articles from 6.425 newspapers from the 17th century (Colavizza and Cuper, 2021). For this dataset, both the original OCR and the manually corrected Ground Truth were available. The OCR of these articles has a high variance in quality. Sometimes parts of articles are missing, which can lead to misleading results if we compare the CER with QuPipe outcomes. To decrease the chance of misleading results due to unmatched texts, we decided to create a controlled set based on matched sentences.

Per article, we automatically extracted sentences from the Ground Truth and matched these with sentences from the OCR. To split the text in sentences, we arbitrarily used the indicators '.', '?' and '!'. We used only sentences that contained at least 7 words. The Python package SequenceMatcher (Python Software Foundation, 2022) was used to match the sentences from the Ground Truth with the corresponding sentences from the OCR. The basic idea of SequenceMatcher is to find the longest contiguous matching subsequence. Sentences were selected as a pair when the match was equal to, or higher than 75%. The cut-off point of 75% was based on a random sample draw, in which we determined the lower limit to prevent false positive matches.

After matching, 146.216 sentences were returned. A disadvantage of SequenceMatcher is that it can match sentences with a difference in length. This means, that it can match a complete sentence to a partial sentence, which can lead to misleading results in the experiments (see table 1). We therefore decided to remove all sentences with a larger than 10% difference in word count. This led to a final dataset of 94.471 matched sentences.

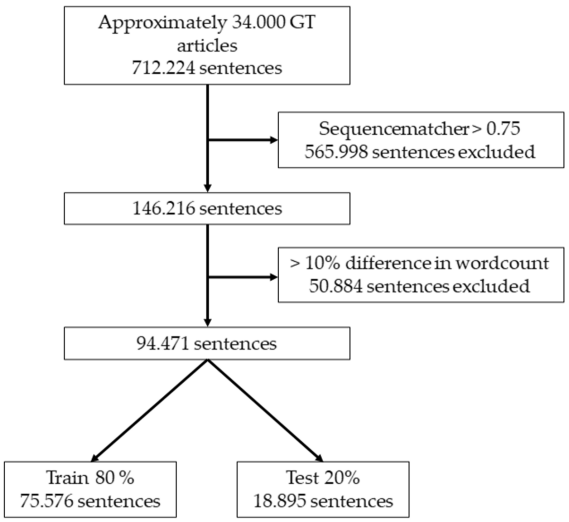


Figure 2: Schematic overview of the data collection process .

Table 1: An example of noise in sentence matching

Type	Sentence
GT	een wijdt-loopighe memorie heeft overgelevert raeckende eenighe geestelijcke goederen
OCR	mog een wijdt-loopighe memorie heeft overgelev raeckende eenighe geeftelijcke goede ridders van malt ha ondert staet leggendec

We divided the dataset in a training and test set with a ratio of 80/20, leading to a training set of 75.576 sentences and a test set of 18.895 sentences. The training set was used for the creation of the trigram model and the reference values. The test set was used for evaluating QuPipe. Figure 2 shows a schematic overview of the data pre-processing. Figure 3 shows the distribution of the test set among the various newspaper publishers.

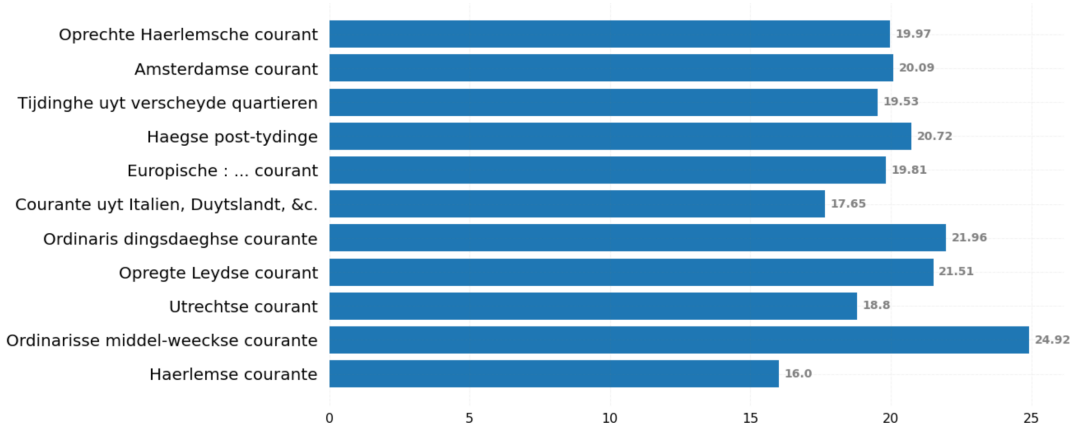


Figure 3: Distribution (%) of the test set across the newspaper publishers.

We calculated the Character Error Rate (CER) for the test set using the ocrevalUAtion tool (IMPACT Centre of Competence, 2019). The results were then split into 'good' (CER equal to or less than 10) or 'not good' (CER above 10). The distribution between these two categories in the test set is shown in Table 2.

Table 2: Distribution of sentences between CER ≤ 10 and CER > 10

CER	absolute	percentage of total
CER ≤ 10	10330	54.7%
CER > 10	8565	45.3%
Total	18895	100%

3.2 Implementation of the measures

3.2.1 Dictionary lookup 'token'

We performed a 'token' dictionary lookup with a modern dictionary, a historical dictionary, and a combination of both. The modern dictionary consists of a word list from Open Taal (Open Taal, 2020) and the historical dictionary consists of two word lists from INT (Instituut voor de Nederlandse taal, 2012). The combined dictionary is a combination of all these word lists. For the 'token' dictionary lookup, all words were used, regardless of the number of time they appeared in the text.

For all dictionary lookups, we followed these preparation steps:

- All punctuation marks, apart from the hyphen, are removed from the text;
- The text is lemmatised with the Python package SpaCy;
- All characters are set to lower case.

After these steps, every word of the text is checked against the selected dictionary. This results in a number of words that are found in the dictionary. This number is divided by the total amount of words in the text to get a percentage of correct words. The value of the dictionary lookup 'token' is 1 if the percentage is equal or above a the reference value (see section 3.3) and 0 if otherwise.

3.2.2 Dictionary lookup 'type'

As with the dictionary lookup 'token', the dictionary look up 'type' was performed on a modern, a historical and a combined dictionary. Furthermore, the preparation steps are the same. However, where the dictionary lookup 'token' is performed on all words, the dictionary lookup 'type' is performed on the set of words. This means that words that occur multiple times in a text, only count once for the dictionary lookup 'type'. When a text has a high number of repeated words, a 'type' lookup can give a more accurate representation of the quality for those words that are important for the essence of a text. An example of the differences between a 'token' and a 'type' representation of a sentence is shown in Table 3.

Table 3: 'Token' versus 'type' sentence representation

Variant	Sentence representation
Token	(de, alexander, is, er, waarschijnlijk, de, merlijn, is, er, zeker, geschreven)
Type	(de, alexander, is, er, waarschijnlijk, merlijn, zeker, geschreven)

Every word in the set is checked against the selected dictionary. This results in a number of words that are found in the dictionary. This number is divided by the total amount of words in the set to get a percentage of correct words. The value of the dictionary lookup 'type' is 1 if the percentage is equal or above the reference value (see section 3.3) and 0 if otherwise.

3.2.3 Average word length

For the average word length, the length of every word is calculated first. Then, we calculated both the mean and the median word length of all words in a sentence. Both

the mean and median are compared with a reference set (see section 3.3). If both the mean and median are in the range of the reference value, the average word length value is set to 1. If only one or neither of the values are in this range, the average word length value is set to 0.

3.2.4 Sentence length

The sentence length is calculated by counting the amount of words in the sentence. This value is compared with the reference set for sentence length (see section 3.3). The sentence length value is set to 1 if the sentence length is in the range of the reference value and 0 if otherwise.

3.2.5 Language detection

Lui (2011) developed a Python package which can identify the language of a given text. The output of the tool is the identified language and the non-normalized probability estimate for the language. We ran some sample tests to test how this tool performed on historical Dutch texts and on texts with OCR errors. We found that the tool was capable of recognizing historical Dutch as the a Dutch language. However, when there are a lot of OCR-errors in the text, it predicts another language.

Since we are testing on Dutch historical news articles, the value is set to 1 if the identified language is 'nl' (Dutch) and 0 if otherwise. For this first version of QuPipe, we did not take the probability estimates into account.

3.2.6 Garbage detection

For our garbage detection measure, we used some of the rules introduced by Taghva et al. (2001) and Kulp and Kontostathis (2007). For this first version of QuPipe, we implemented the following selection of rules:

- A word with more than 20 characters is considered as garbage.
- A word with more than 3 identical characters in a row is considered as garbage.
- If a word has only vowels and consonants, and if the number of vowels is 8 times greater than the number of consonants or the other way around, the word is considered as garbage.
- If we strip the first and last letter of a word, and there are more than two punctuation characters in the word, it is considered as garbage.
- If a word start and ends with a lower-case letter, and one of the remaining characters is upper-case, it is considered as garbage.

Furthermore, we added our own rules, based on Dutch language characteristics:

- If a word contains letters or punctuation that are not part of the Dutch languages, it is considered as garbage.
- If a word starts or ends with more than 4 consonants, it is considered as garbage.

For every rule, the percentage of words in a sentence that meets this rule is calculated, leading to a set of percentages. The garbage value is set to 0 ('garbage') if the percentage of at least one of these rules is equal to or above reference value (see section 3.3) and 1 if otherwise.

3.2.7 Trigram

A trigram is a contiguous sequence of three items. These items are, for example, words or characters. Trigrams can be extracted from a corpus of text. A trigram model is a probabilistic model that can be used to predict if a certain sequence of three items is expected or not. We use the perplexity score for this prediction. The perplexity is the inverse probability, normalized by the number of words. It can be interpreted at how 'perplex' the model is to see a certain sequence. With a perplexity score of zero, the model is not perplexed at all. This indicates a high chance of a existing sequence. With a high perplexity score, the model is 'perplex' about the score, indicating that it is less likely to be a correct sequence. An example of perplexity scores on word level is shown in Table 4.

Table 4: Example of perplexity

Word	Perplexity	Existing word
fish	0	yes
flurp	105	no

To be able to use the trigram measure, we needed a trigram model with probabilities per trigram. We created this model based on our training set. Since we have a historical data set that may contain words that occur very seldom, we decided to use a character level trigram instead of a word level trigram. This means that our trigram sequences consist of a sequence of three contiguous characters. We used our trigram model to calculate the perplexity per word. If the perplexity was equal to or lower than the reference value (see section 3.3), the trigram value was set to 1. If otherwise, the value was set to 0.

3.3 Creation of reference sets

Every language and time period has its own characteristics for written texts. Therefore, it is important to have reference values specifically geared towards these characteristics. These reference values can be used as cutoff or boundaries to determine whether the outcomes of the QuPipe measures are within the expected range.

Since the Ground Truth consists of nearly error-less texts, we used the Ground Truth sentences from the training set to obtain our reference values. First, we calculated the outcome values of every measure for every sentence of this set. The resulting values were ordered and the cutoff and boundaries were chosen to include 90% of the data, as detailed below.

For the **dictionary lookup**, both 'token' and 'type' measures, we used the values based on the combined dictionary. The reference value was selected as the value corresponding to the cutoff point at 10% of the data. This means that all values above this value are considered as desired. The reference values for the words statistics **mean** and **median** were selected by taking the values corresponding to 0.05% and 0.95% of the distribution. The values in between are considered as the desired values. Since the **sentence length** was artificial set to a minimum of 7 words, this is considered as the lowest reference value. The highest value was then selected as value for the cutoff point at 90% of the data. The values in between are considered as the desired values. For both the **garbage** and **trigram**, the reference value was selected at the cutoff point at 90% of the data. This means that all data beneath this value is considered as desired.

3.4 Calculation of the total score of QuPipe

There are various ways in which QuPipe can be used to determine the total score. Which method to choose, depends on the needs of the user. In this section, we describe the three approaches we used for our study: the total calculation, the 'smart quality' calculation and the 'smart quantity' calculation. As this is a first exploratory study on the impact of various methods for determine the OCR quality, the proposed methods are arbitrarily chosen and based on an inspection of a small manual sample. Both smart calculations need to be pre-trained on Ground Truth data. For this study, we focused our calculations on the division between 'good' OCR quality and 'not good' OCR quality.

The normal calculation is the most basic calculation and uses all of the current implemented measures. As described in section 3.2, every measure results in a score of 0 or 1. A score of 1 means the measured value was in the expected range, whereas a value of zero indicates it was not in the expected range. For the normal calculation, these scores are added up to one final score. Only sentences from which the final score is equal to the maximum achievable score is considered as 'good', whereas the rest is considered as 'not good'.

We introduce two other methods to explore if modification of the selected measures changes the output of the calculation. We introduce the 'smart quality' measure that focuses on precision, and the 'smart quantity' measure with a focus an recall. Due to the variety of dictionary measures and the fact that dictionaries are not always available, we exclude these measures during this calculation.

For the 'smart quality' calculation, the most restrictive approach was used to classify a sentence as 'good' or 'not good', meaning that sentences were only classified as 'good' if the final score is equal to the maximum achievable score. This maximum achievable score is based on the number of used measures with every step. This method will lead to the highest achievable precision. To calculate the optimal combination of the 'smart quality' calculation, we started with calculating the precision for all isolated measures. We select the measure with the highest precision. Then, we calculate the precision of this measure with every single other measure. From this, we selected the two measures with the highest precision, but only when the precision is equal to or higher than the current precision. We continued these steps until there was no further increase in precision. An illustration of this process is shown in Figure 4.

The calculation for the best combination of measures for the 'smart quantity' calculation differs slightly from the 'smart quality' measure. For this measure, an sentence was classified as 'good' when its final score was equal to half of the maximum achievable score. For odd numbers, the maximum achievable score was rounded down. With this method, there is more recall while still providing a high precision. This method also starts with the single measure, after which one by one other measures are added. Furthermore, instead of only looking at precision, we chose the measure with the highest recall having a precision that was equal to or above the current precision. We continued these steps until there was no further increase in recall without decreasing the precision. An illustration of this process is shown in Figure 5.

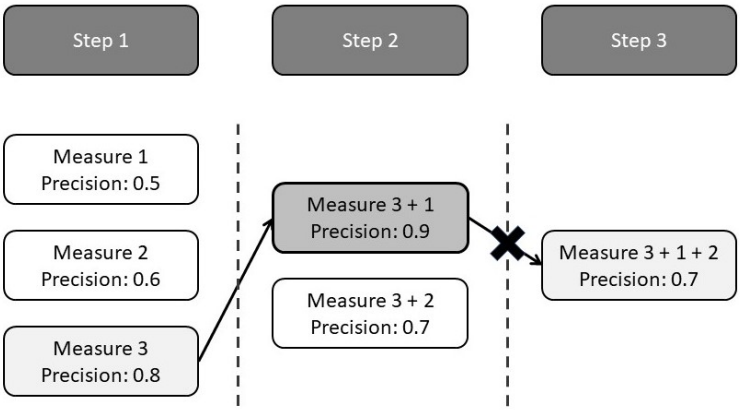


Figure 4: Illustration of the process for the 'smart quality' measures calculation: the optimal score per step (light grey) leads to the total optimal score (dark grey)

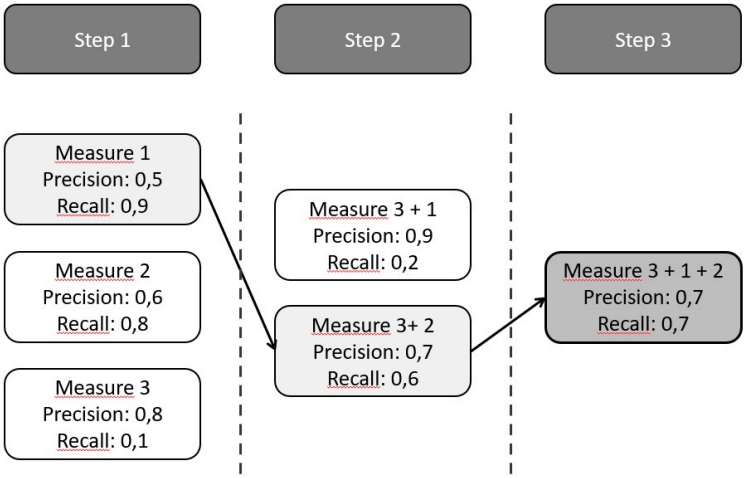


Figure 5: Illustration of the process for the 'smart quantity' measures calculation: the optimal score per step (light grey) leads to the total optimal score (dark grey)

4 Results

After preparing the data and implementing the measures in QuPipe, we performed a series of experiments to evaluate the performance of QuPipe. The first experiment compares the results of a 'token' dictionary lookup (which is the most common used dictionary lookup) with the results of the QuPipe calculations. We used a modern dictionary, a historical dictionary and a combined dictionary (see section 3.2). The second experiment focuses on classifying the OCR quality without a dictionary. Since historical dictionaries are not always available, it is valuable to know if usable predictions can be done without one. For this experiment, we compare two isolated measures with the results of QuPipe. The reference values that were used for the QuPipe measures are shown in Table 5.

Table 5: Reference values

Measure	Reference value
Dictionary lookup token	84.6%
Dictionary lookup type	83.3%
Word length median	between 3 and 6
word length mean	between 4.15 and 6.25
Sentence length	between 7 and 31
Garbage	0.08
Tri-gram	46

For the QuPipe total score, all measures were used. To obtain the best combination of measures for the 'smart quality' calculation, we calculated the combination of measures with the highest precision. For the 'smart quantity' calculation, we calculated the combination of measures that represents the best trade-off between recall and precision. Both were determined using a sample of the training set, as described in section 3.4. Based on these calculations, we excluded the sentence length measure from the smart calculations, as it had a negative influence on the precision. All other measures were kept. For section 4.1 we added the various dictionary measures to the smart calculations.

We calculated the precision and recall of the classification from QuPipe based on the CER classification (see section 3.1).The precision and recall were used to compare the various measures and calculations.

We conclude the results with a comparison of reference values with various amounts of Ground Truth data.

4.1 Results with dictionary lookup

Since historical dictionaries are not always available, we compare the use of a modern dictionary, a historical dictionary, and a 'combined' dictionary (with both historical and modern words). Furthermore, we compared the QuPipe outcomes with the outcomes of a 'token' dictionary lookup with the cutoff point at 80%, based on the suggested minimal accuracy in literature (Strange et al., 2014; Van Strien. et al., 2020). The results are shown in Table 6.

For all three dictionaries, the QuPipe 'smart quality' calculation returned the highest precision with a precision above 0.9, which means that more than 90% of the sentences were correctly classified as 'good'. However, for all three dictionary types, the recall is low. For the modern dictionary, the recall is only 0.008, meaning that only 0.8% of all 'good' sentences were classified as good. The highest recall is for the combined dictionary, where 14.2% of all 'good' sentences were classified as good. When looking at the recall for the modern dictionary, the QuPipe 'smart quantity' calculation gives the best result in terms of both precision and recall compared to the 'token' dictionary lookup. For both the historical and the combined dictionary, the 'token' dictionary lookup provides the highest recall in comparison to the QuPipe results.

When using a modern dictionary, there is only a small difference between the QuPipe total calculation and the 'smart quality' calculation. Both have a high precision, but a very low recall. If we compare this to the 'smart quantity' calculation, we see that while the precision decreases around 0.095, the recall increases with 0.255.

For the historical dictionary, the QuPipe 'smart quality' calculation has a 0.01 higher

precision and a 0.021 higher recall than the QuPipe total calculation. If the 'smart quality' calculation is compared with the 'smart quantity' calculation, the precision decreases with 0.074. However, the recall increases from 0.139 tot 0.414. This means that from all 'good' sentences, the correct classified sentences went from 13.9% to 41.4%. For the combined dictionary, the QuPipe 'smart quality' calculation has a higher precision and a higher recall than the total calculation. If the 'smart quality' calculation is compared with the 'smart quantity' calculation, the precision decreases with 0.075. However, the recall increases from 0.142 to 0.418.

If we compare the three 'token' dictionary lookup methods, the modern dictionary performs the best on precision. However, it performs the worst on recall, with a recall of only 0.050, which means that from all sentences classified as 'good', only 5% was predicted correct. When comparing the historical and the combined dictionary, the historical dictionary has a slightly higher precision (+0.003), whereas the combined dictionary has a slightly higher recall (+0.008).

Table 6: Dictionary lookup versus QuPipe

Type	Outcome	80%	QuPipe total	'smart quality'	'smart quantity'
Modern dictionary	Precision	0.837	0.933	**0.935**	0.840
	Recall	0.050	0.008	0.008	**0.263**
Historical dictionary	Precision	0.821	0.897	**0.907**	0.833
	Recall	**0.462**	0.118	0.139	0.414
Combined dictionary	Precision	0.818	0.896	**0.906**	0.831
	Recall	**0.470**	0.119	0.142	0.418

4.2 Results without dictionary lookup

Since dictionaries are not always available, especially for older texts, we also performed an experiment to see how the pipeline performs without dictionary lookup. We started by calculating the precision and recall for every isolated measure, as shown in Figure 6.

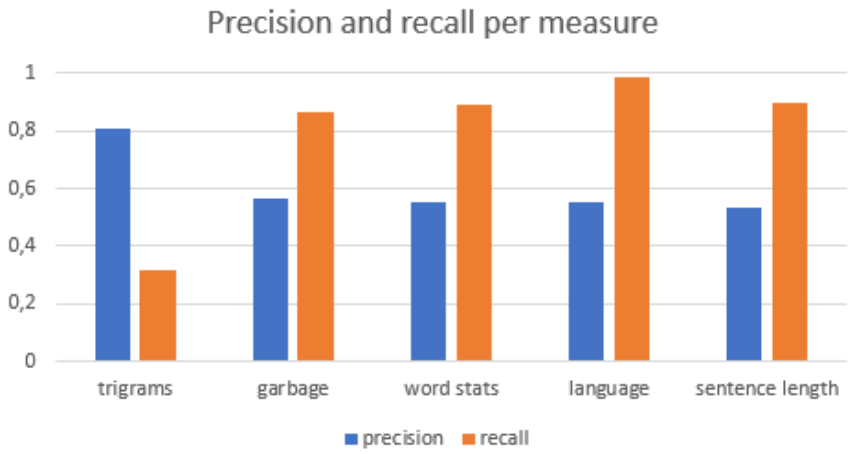


Figure 6: Precision and recall per measure

We first compare QuPipe's calculations with the measure with the highest precision, and then with the measure with the highest recall.

4.2.1 Quality before quantity

The measure with the highest precision is the trigram measure, with a precision of 0.810, which indicates that 81% of the data is correctly classified as 'good'. However, the recall is 0.317, which means that only 31.7% of all 'good' sentences were classified as good. Table 7 shows the precision and recall of the trigram measure, the QuPipe total calculation and the QuPipe 'smart quality' calculation.

With the QuPipe total calculation, we see that the precision increases from 0.810 to 0.823 (+0.013). This means that more data is correctly classified as 'good'. However, the recall decreases from 0.317 to 0.209 (-0.108), which means that less 'good' sentences were classified as good in total. Looking at the 'smart quality' calculation, we see that the precision increases further, leading to a difference of 0.029 in precision between the trigram measure and the 'smart quality' calculation. Also, there is a smaller decrease in the recall, with a difference of 0.071 compared to the trigram measure. When looking purely at quality, the QuPipe 'smart quality' calculation performs best.

Table 7: Highest precision measure versus QuPipe

Outcome	Trigram measure	QuPipe total	QuPipe 'smart quality'
Precision	0.810	0.823	**0.839**
Recall	0.317	0.209	0.246

4.2.2 Quantity before quality

If we look at quantity before quality, the measure with the highest recall is the language detection measure, with a recall of 0.982. This means, that 98.2% of all 'good' sentences were classified as good. However, this measure has a precision of only 0.552, which means that almost 50% of the data is incorrectly classified as good. Therefore, this measure is not much more effective than just using all the data without classifying. Table 8 shows the precision and recall of the language detection measure, the QuPipe total calculation and the QuPipe 'smart quantity' calculation.

When we use QuPipe total calculation, we see that the precision strongly increases from 0.552 to 0.823 (+0.271), however, the recall strongly decreases from 0.982 to 0.209 (-0.773). When we look at the QuPipe 'smart quantity' calculation in comparison to the language detection measure, the precision increases with 0.034 and the recall decreases with 0.149. Looking at as high as possible quality while also taking quantity into account, the QuPipe 'smart quantity' calculation performs slightly better than the language detection measure.

Table 8: Highest recall measure versus QuPipe

Outcome	Language detection measure	QuPipe total	QuPipe 'smart quantity'
Precision	0.552	0.823	0.586
Recall	**0.982**	0.209	0.833

4.3 References and lack of Ground Truth

Since Ground Truth data is sparse and expensive to obtain, we calculated and compared reference values based on various amounts of Ground Truth data. We started with

our 'base' reference values, which were calculated on the whole training set. Then, we calculated reference values with 25% of the Ground Truth and with 2.5% of the Ground Truth. Table 9 shows the difference between the values. The table shows that the reference values for the 'type' dictionary lookup, the word length median, the sentence length, and the garbage measure stay the same although the amount of Ground Truth is reduced. The reference values for dictionary lookup 'token', word length mean, and trigram show a small shift when they are based on a smaller amount of Ground Truth data.

Table 9: Reference values for different amounts of Ground Truth data

Reference	Full data 75576 sentences	25% of data 18894 sentences	2.5% of data 1890 sentences
Dictionary lookup 'token'	84.6%	84.6%	84.4%
Dictionary lookup 'type'	83.3%	83.3%	83.3%
Word length median	between 3 and 6	between 3 and 6	between 3 and 6
Word length mean	between 4.15 and 6.25	between 4.17 and 6.25	between 4.17 and 6.21
Sentence length	between 7 and 31	between 7 and 31	between 7 and 31
Garbage	0.08	0.08	0.08
Trigram	46	45.5	45.2

5 Conclusion and future work

Our study showed that when quality is the most important aspect, the QuPipe 'smart quality' calculation results in the highest precision when measuring OCR quality. This confirms our hypothesis that a combination of measures gives a more accurate prediction of OCR quality than a single measure, although sometimes the difference with the other methods is small.

Although the 'smart quality' measure has the highest precision, in all cases this is at the expense of recall. When only a high quality is important this is no issue. However, in most cases, quantity will also be taken into consideration as most analyses depend on having a big enough dataset. In such cases, one can opt to use the 'smart quantity' calculation, which has a lower precision but a higher recall. Even though it focuses on getting the highest quantity, in all cases the precision is still higher than when using a single measure. When compared to the historical and combined dictionary, the difference in precision and recall is small. However, when only a modern dictionary is available, QuPipe performs clearly better on both precision and recall. We also noticed that although the performance is not quite so good with the QuPipe total calculation, the precision of this calculation is still better compared to using a single measure.

A disadvantage of QuPipe is the need for a Ground Truth set to be able to create the reference sets. However, this only has to be done once for each language and/or time period. From then on, the created reference set can be used on all data with the same characteristics. Additionally our small experiment suggests that reliable reference sets can be obtained from only a small amount of Ground Truth, but further research is needed to confirm this. The same disadvantage is present for the QuPipe smart calculations. These also need a (small) Ground Truth set to optimise the combination of used measures. However, given the improved performance compared to the total calculation, and the fact that there is also Ground Truth needed for the reference sets, it seems to be worth the investment. Especially because some measures can unintentionally have a negative effect on the total result.

Since these first experiments with QuPipe seem positive and validate the idea that a combination of measures works better than using just one measure, we will continue improving QuPipe. Our next steps will be to expand and improve some of the measures, such as the garbage measure and the language detection. Then, we plan to examine new measures to be added to the pipeline, such as letter ratio and word embedding techniques. In addition, we want to add more variation in outcome instead of just 'good' or 'not good'. Furthermore, we want to investigate if reference values based on a small amount of Ground Truth are as reliable as those based on larger amount of Ground Truth. We also want to look at possible other smart calculation options and improve our current methods, thereby creating a more robust approach based on more extensive analyses and data. For both the references and calculation, we aim to explore if the type of material has an influence on the outcomes. Lastly, we want to expand our experiments to other data sets, with other types of text, different time periods and, if possible, other languages.

References

Farag Ahmed, Ernesto De Luca, and Andreas Nürnberger. Revised n-gram based automatic spelling correction tool to improve retrieval effectiveness. *Polibits*, 40: 39–48, 12 2009. doi: 10.17562/PB-40-6. http://dx.doi.org/10.17562/PB-40-6.

S. El Atawy and A. Abd ElGhany. Automatic spelling correction based on n-gram model. *International Journal of Computer Applications*, 182:5–9, 08 2018. 10.5120/ijca2018917724.

Ryan Baumann. Automatic evaluation of OCR quality. https://ryanfb.github.io/etc/2015/03/16/automatic_evaluation_of_ocr_quality.html, 2015.

Giovanni Colavizza and Mirjam Cuper. Is your OCR good enough? A comprehensive assessment of the impact of OCR quality on downstream tasks [Data set]. http://doi.org/10.5281/zenodo.4498186, 2021.

Mirjam Cuper. Examining a multi layered approach for classification of OCR quality without Ground Truth [presentation]. https://zenodo.org/record/4621253#.YWXrFBxcKUk, 2021.

Rose Holley. How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine: The Magazine of the Digital Library Forum*, 15, 01 2009.

IMPACT Centre of Competence. ocrevalUAtion. https://github.com/impactcentre/ocrevalUAtion, 2019.

Instituut voor de Nederlandse taal. INT Historische Woordenlijst. https://taalmaterialen.ivdnt.org/download/tstc-int-historische-woordenlijst/, 2012.

Scott Kulp and April Kontostathis. On retrieving legal files: Shortening documents and weeding out garbage. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, volume Special Publication 500-274. National Institute of Standards and Technology (NIST), 2007. URL http://trec.nist.gov/pubs/trec16/papers/ursinus.legal.final.pdf.

Marco Lui. Language Identifier. https://github.com/saffsd/langid.py, 2011.

Thi Tuyet Hai Nguyen. *Facilitating Access to Historical Documents by Improving Digitisation Results*. Theses, La Rochelle Université, April 2020. URL https://hal.archives-ouvertes.fr/tel-03058611.

Open Taal. Nederlandse woordenlijst. https://github.com/OpenTaal/opentaal-wordlist, 2020.

Python Software Foundation. difflib — Helpers for computing deltas. https://docs.python.org/3/library/difflib.html, 2022.

Alexander M. Robertson and Peter Willett. Applications of n-grams in textual information systems. *J. Documentation*, 54:48–67, 1998. https://doi.org/10.1108/EUM0000000007161.

U. Springmann, Florian Fink, and K. Schulz. Automatic quality evaluation and (semi-) automatic improvement of mixed models for ocr on historical documents. *CoRR*, abs/1606.05157, 06 2016.

Carolyn Strange, Daniel McNamara, Josh Wodak, and Ian Wood. Mining for the meanings of a murder: The impact of ocr quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8:1–17, 01 2014.

Kazem Taghva, Tom Nartker, Allen Condit, and Julie Borsack. Automatic removal of "garbage strings" in ocr text: An implementation. *The 5th World Multi-Conference on Systemics, Cybernetics and Informatics*, 01 2001.

Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman. Impact analysis of ocr quality on research tasks in digital archives. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Research and Advanced Technology for Digital Libraries*, pages 252–263, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24592-8. https://doi.org/10.1007/978-3-319-24592-8_19.

Daniel Van Strien., Kaspar Beelen., Mariona Ardanuy., Kasra Hosseini., Barbara McGillivray., and Giovanni Colavizza. Assessing the impact of ocr quality on downstream nlp tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH,*, pages 484–496. INSTICC, SciTePress, 2020. ISBN 978-989-758-395-7. doi: 10.5220/0009169004840496.

Jian-Cheng Wu, Jim Chang, and Jason J. S. Chang. Correcting serial grammatical errors based on n-grams and syntax. *Int. J. Comput. Linguistics Chin. Lang. Process.*, 18, 2013.

Richard Wudtke, Christoph Ringlstetter, and Klaus U. Schulz. Recognizing garbage in ocr output on historical documents. In *MOCR-AND '11: Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, MOCR_AND '11, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306850. doi: 10.1145/2034617.2034626. https://doi.org/10.1145/2034617.2034626.

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

Modeling Ontologies for Individual Artists: A Case Study of a Dutch Ceramic Glass Sculptor

Victor de Boer*[1], Daan Raven[1,2], Erik Esmeijer[3], and Johan Oomen[2]

[1]Vrije Universiteit Amsterdam, Amsterdam, the Netherlands.
[2]Netherlands Institute for Sound and Vision, Hilversum, the Netherlands.
[3]Independent researcher, Amsterdam, the Netherlands.

Most Cultural Heritage information currently available in the form of structured knowledge representations is limited to aggregated information extracted from general collection management systems and archives. These predominantly describe collections and artefacts at a general level. However, individual artists are able to provide information at a much higher level of variety and detail. This paper investigates the usefulness of ontologies to structuring and providing access to enriched information related to individual artists and their works. It reports on a case study developing and validating an ontology for a ceramic glass artist. The study follows the Methontology ontology engineering methodology. Ontology development is validated using domain-specific competency questions related to artworks and production processes. The case study shows how ontologies may be used to structure enriched information pertinent to works of individual artists. Ontologies provide access to such information, including details on activities related to creative and productive processes of artworks. Moreover, information may encompass materials processed, equipment used, locations, time spans, parties involved, and chronological and causal registration of how activities take place. Modeling information this way may further facilitate both experts' as well as the general public's access to information.

Keywords: ontology engineering; cultural heritage; methontology

* Corresponding author, v.de.boer@vu.nl

1 Introduction

There is a long tradition of using structured, machine interoperable knowledge using semantic technologies in the Cultural Heritage domain (Hyvönen, 2012), allowing experts to access and interpret Cultural Heritage information by means of intelligent reasoning and inferencing. As noted by Amin et al. (2008), Cultural Heritage experts have rather different information needs, ranging from handling information related to specific objects and (potential) collections, unconstrained intelligence for prospective publications to the construction of structured vocabularies for museums, archives, and libraries. Experts search for information in 'analog' literature and archives and use search engines to find information in collection management systems and digital sources.

Currently, most information on the (creation of) artworks that is available in the form of structured knowledge representations, such as Linked Data, is limited to the aggregated information stored in general collection management systems such as Europeana[1], the Getty Research Institute[2], RKDartists[3], and digital archives of museums. While several papers have been published related to the design of ontologies for Cultural Heritage institutions, there has been no detailed investigation of the use of ontologies to specify information related to the works of individual artists and the development of such ontologies. Artists dispose of detailed, pertinent, and significant information on (concepts related to) production materials, methods, and techniques as well as on creative and constructive processes. Such information cannot be specified in great detail using the current solutions.

While several papers are published on designing ontologies for Cultural Heritage institutions, there are as yet no detailed investigations into the use of ontologies to specify information relative to the work of individual artists and the systematic development of such ontologies. This paper investigates knowledge representations for modeling Cultural Heritage information at a high level of variety and detail as can be provided by individual artists. Special focus is on documenting artists' collections to better facilitate experts' search tasks (Munir and Anjum, 2018).

What follows describes the emergence of a formal ontology (Gruber, 1995) modeling specific information for individual artists. This includes developing and validating art-specific concepts and requirements. The paper presents a reusable method, building on the 'Methontology' method for ontology development (Fernández-López et al., 1997), describing the steps of specification, conceptualization, integration, implementation, and evaluation in a case study on a contemporary ceramic glass sculptor.

The aim is to create a machine-processable structure for information related to artworks and artist (Dijkshoorn et al., 2018). Documenting the artists' collection in this structure, in turn, enables improved query construction, thereby increasing the information retrieval capabilities of Cultural Heritage experts (Munir and Anjum, 2018).

Defining concepts related to individual artists and their work also opens up possible linkages of these concepts and a variety of other data sources. Currently, efforts are being made within the Cultural Heritage domain attempting to improve Linked Data practices describing art. An example of these efforts is the Linked Art initiative in which several institutions and museums cooperatively develop a shared data model with

1 https://www.europeana.eu/
2 https://www.getty.edu/research/
3 https://rkd.nl/nl/explore/artists

links to the terminology of Getty Vocabularies[4]. A second example is an initiative of the Dutch organization Netwerk Digitaal Erfgoed (NDE) in collaboration with CLARIAH, named Termennetwerk[5]. Termennetwerk provides access to various terminology sources such as the Getty Art and Architecture Thesaurus (AAT), Dutch heritage thesauri published by national aggregators (RCE, SVCN), and geographical thesauri such as GeoNames and WikiData.

Accessing and reusing shared vocabularies plays a vital role in creating semantic connections between (descriptions of) artworks and related concepts. Therefore, the focus is on linking concepts to the terminology of shared vocabularies too. There is a need for this new kind of representations as current solutions are limited to structuring general information only, while artists themselves are very well able to provide information at a much higher level of detail and variety.

1.1 Case Study

The case study involves modeling an ontology specifying information related to the work of the Dutch ceramic glass sculptor Barbara Nanning[6]. After graduating from the Gerrit Rietveld Academy forty years ago, she is now viewed as a well-respected artist by both national and international institutions (Eliëns, 2019). While she originally majored in ceramics, she has later shifted her focus to exclusively creating artwork made of glass.

Glass art can have particularly varying and complex creation processes. Production may require glass techniques such as 'casting', 'fusing', and 'slumping'. In the case of the artist, there also is a geographical division between activities related to conceiving, pattern building, and preparing in the Netherlands and creation and finishing of the work itself in the Czech Republic (Eliëns, 2019). Capturing and structuring this variation and complexity in an ontology is challenging. Moreover, as there are no structured data sources available for most of the information sought, deriving this information requires several intensive knowledge acquisition trajectories.

This paper describes the application of an existing ontology engineering methodology to develop a knowledge representation specifying and providing access to enriched information related to the works of the artist. This ontology is validated and evaluated to assess the extent to which the case study can be generalized to artists in general.

The remaining part of this paper is structured as follows: Section 2 provides an overview of related work. Section 3 describes the development of the ontology and its results. Section 4 discusses the evaluation of (the potential usage of) the ontology. Section 5 presents discussion, conclusion, and directions for future research.

2 Context

2.1 Cultural Heritage modelling

The engineering of ontologies is a core research topic within the semantic web community (Pattuelli, 2011). Over the last several decades, efforts have led to the development of ontologies within different domains (Iqbal et al., 2013), among which Cultural Heritage. Some domain ontologies are widely accepted and, though they make use of

4 https://linked.art/index.html; http://vocab.getty.edu/
5 http://demo.netwerkdigitaalerfgoed.nl/termennetwerk/en
6 http://www.barbarananning.nl/

different approaches, are generally able to describe artworks and collections (Dijkshoorn et al., 2018). One of these is the CIDOC-CRM ontology, developed for Cultural Heritage institutions to facilitate information integration (Doerr, 2003). CIDOC-CRM maintains an event-centric approach to describing artworks and the provenance of artworks. An example of an event would be the acquisition event or "E8:Acquisition" indicating a change in ownership. Events are connected to artworks by means of object properties[7]. CIDOC-CRM is developed specifically for use by institutions publishing collection metadata, including "museums, libraries and archives"[8], with its goal being the achieving of interoperability of collection metadata (Dijkshoorn et al., 2018).

Another widely adopted ontology is the Europeana Data Model (EDM), created to enable structured data delivery to Europeana by cultural heritage institutions (Doerr et al., 2010). EDM also maintains an object-centric approach, describing artworks using a collection of object properties[9]. Like CIDOC-CRM, EDM is designed both specifically for the museum sector at a high level of aggregation. It is designed as a more general interoperability layer on top of more specific models (Dijkshoorn et al., 2018).

Whilst these ontologies focus on structuring general information related to the collections of Cultural Heritage organizations on a general level, they are not designed to structure information related to works of individual artists at a high level of detail. This study investigates the bottom-up development of such a more specific ontology as well as to what extent the resulting ontology can be mapped to CIDOC and EDM.

2.2 Ontology Engineering Methodology

For the case study, an ontology engineering methodology is used, introduced by Fernández-López et al. (1997) named 'Methontology'. This is further described in Section 3.

Methontology appears in several publications presenting ontologies for different domains using the Methontology methodology. One of these is the ITEMAS ontology, as presented by Moreno-Conde et al. (2019), having its origins in the healthcare domain. The main domain concepts relate to public healthcare organizations and health technology innovations. In Kalbasi et al. (2014), a minimalist ontology is developed in the domain of geosciences to stimulate collaborative ontology development. Other ontologies include one presented by Hennig et al. used for annotating and comparing cytometric (cell measurement) data and one chemical ontology published as a best practice example of Methontology itself by Fernández et al. (Hennig et al., 2009; López et al., 1999). Several of the involved ontology engineers indicate having chosen Methontology because of the highly detailed documentation and examples available.

Some papers also refer to using an adapted version of Methontology or a mixture of several ontologies including Methontology. An ontology for Greek mythology defined by Syamili and Rekha makes use of three different ontology engineering approaches (Syamili and Rekha, 2018). Among these Methontology, from which the development stages are adopted, yet without, among other things, the evaluation stage. Finally, there is an ontology for specifying cultural heritage resources based on Methontology (Pattuelli, 2011). This ontology prototype is refined using a different methodology in the later stages of the project.

7 http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf
8 https://www.cidoc-crm.org/
9 https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/

2.3 Methontology

Over the last decades, several methodologies for ontology engineering have been defined, providing descriptions of techniques and activities that come with developing ontologies (Iqbal et al., 2013). Nevertheless, analysis by Iqbal et al. suggests a lack of mature methodologies (Iqbal et al., 2013). This is mainly caused by poor documentation related to techniques and activities being insufficiently detailed. Contrary to this, Methontology (Fernández-López et al., 1997) provides a methodology whose documentation contains a high level of details.

Methontology bases the development of ontologies on the concept of an 'evolving prototype' in which ontologies pass through well-documented stages of specification, conceptualization, integration, implementation, and evaluation (Fernández-López et al., 1997). The concept of an evolving prototype makes that definitions can be refined or complemented when necessary based on constructed user stories.

Given that the requirements of an ontology are not clear initially, the adoption of an evolving prototype approach is advisable as it can be used to develop application-independent ontologies (Iqbal et al., 2013). Other reasons motivating a choice for applying Methontology are supported for reusing existing ontologies in the implementation phase. This has resulted in the repeated use of Methontology among ontology engineers.

3 Development of the Ontology

This paper substantiates how Methontology can be adapted and applied to develop an ontology departing from unstructured information representing a domain. This section describes how the stages of the Methontology methodology are adapted and applied.

3.1 Specification

Purpose and requirements follow from a set of user stories, indicating potential use of the ontology, level of granularity required, and beneficiaries from its development. User stories and corresponding competency questions are helpful during specification to maintain alignment with the purpose of the ontology. User stories and competency questions are effective during the evaluation stage as well, as queries can be defined based on competency questions, thus also serving as a validation method for the whole development effort (Baker and Cheung, 2007). Typically, specification yields a partially complete picture of the domain that needs to be concise and consistent (Fernández-López et al., 1997).

3.1.1 User Stories

The purpose and requirements of an initial prototype follow from two user stories and related competency questions (CQ). User stories and competency questions are developed through analysis of domain literature and structured interviews with the artist[10] (Eliëns, 2019). We restrict ourselves to this information rather than interviewing

[10] An interesting challenge faced during development of the ontology follow from the restrictions placed on conducting face to face knowledge acquisition activities due to COVID-19. Rather than conducting interviews in person, this called for an alternative approach, especially during specification and conceptualization stages. The study therefore chose to work with digital forms, allowing the artist to

other stakeholders, as our aim is to produce an ontology that is a faithful representation of the world-view and internal information organization of the artist herself. Interviews clarify the kind of information the artist is able and willing to provide. An example of such a question is "Is there access to information related to craftsmen participating in the creation process?".

User stories address sub-collections of the artist: Verre Églomisé[11] and Coloured Shadows[12], shown in Figures 1 and 2.

Figure 1: Photographs showcasing the process and end result of the Verre Églomisé sub collection

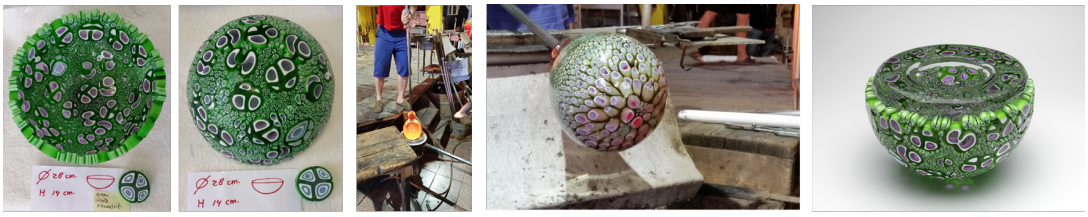


Figure 2: Photographs showcasing the process and end result of the Colored Shadows sub collection

Competency questions typically may then refer to:

- physical dimensions of artworks
- equipment required during creative and productive processes
- skills, crafts, and techniques necessary to achieve targets
- infrastructures, logistics, and locations
- collections, both private and public
- links to general collection management systems and archives

A top-level competency question was identified "As an artist I want to be able to describe artistic and production processes underlying the creation of Verre Églomisé / Coloured Shadows objects as well as to dispose of general information and digital representations of individual artworks, thus improving structuring and managing of documentation.". For Verre Églomisé (CQ-VE) / Coloured Shadows (CQ-CS), this led to several specific competency questions:

- CQ-VE-1: What are the top two artworks in the Verre Églomisé series sorted by height?

nevertheless purposefully describe artworks and related creative and productive processes.

11 https://www.barbarananning.nl/nl/project/verre-eglomise/

12 https://www.barbarananning.nl/project/coloured-shadows/

- CQ-VE-2: What equipment is used for sandblasting of the artworks within the Verre Églomisé artworks and how can this equipment be described?
- CQ-VE-3: Which artworks in the Verre Églomisé series are currently stored in her private collection?
- CQ-VE-4: What is the description of the technique that was last applied in the creation process of the 'Go with the Flow' artwork from the Verre Églomisé series and which parties were involved in this technique?
- CQ-VE-5: What are the links to the digital representations of the artworks in the Verre Églomisé series?
- CQ-CS-1: What are the top three artworks in the Coloured Shadows series sorted by descending year of completion?
- CQ-CS-2: What activities are part of the creation process of the 'Beneath the Water' artwork in the Coloured Shadows and where have these activities take place?
- CQ-CS-3: What are the materials processed in the creation process of the 'Coral Reef' artwork from the Coloured Shadows collection and during which activities were these materials processed?

Besides competency questions, specific Cultural Heritage data publishing requirements, as identified by Dijkshoorn et al. (2018), are also taken into account. Here, the top-level question is: "As an artist, I want to be able to publish the enriched data describing my artworks and the corresponding creation process to provide access to cultural heritage experts and the general public and thereby improve current fact-finding practices.". Specific requirements are:

- REQ1: assuring the interoperability of the data model while being able to specialize it
- REQ2: supporting both object and event-based approaches for describing artefacts
- REQ3: providing the ability to specify the changes made to artefacts over time
- REQ4: separating the meta-data and representations of artefacts
- REQ5: allowing the inclusion of multiple views of the same artefacts from different sources
- REQ6: providing the ability to add context to artefacts by including domain-specific information

3.1.2 Specification Document

The study involves several knowledge acquisition activities, calling for both structured and unstructured interviews with the artist. Topics addressed cover types of activities conducted during creation processes and materials used.

Domain literature is analyzed to gain a basic understanding of glassblowing, glass houses, and artworks of the artist (Eliëns, 2019).

Table 1: Ontology Requirements Specification Document

Ontology Requirements Specification Document	
Domain	*- Cultural Heritage / Arts*
Purpose	*- Ontology specifying enriched information describing artworks from the Verre Églomisé and Coloured Shadows series and modeling related creative and productive processes.* *- Providing Cultural Heritage experts access to this information.*
Level of Formality	*Semi-formal*
Scope	*- List of artworks: Coral Reef, Go with the Flow ...* *- List of Materials: Glass, Gold Leaf, Alexandrite Glass, Uranium Glass, AJETO Glass, Iron, Louvre Gold Leaf ...* *- Tools and Equipment: Furnace, Gilders Tip, Blowpipe, Plate System, Cooling Oven, Sharpening Stone ...* *- Location-related Concepts: Country, Glass Studio, Museum, City ...* *- People-related Concepts: Actor, Gilder, Glass Blower, Creator ...* *- Activities: Glassblowing, Gilding, Modeling, Sawing, Grinding ...* *- Concepts for the Requirements of Data Publication: Event, Thing, Entity, Digital Image ...* *- Properties: created-by, located-in, has-height, finishes-with, used-for, involves, part-of, has-duration, followed-by ...*
Sources	*- Domain Literature: Books by Eliëns and te Duits describing Barbara Nanning (Eliëns, 2019; te Duits, 2003).* *- Textual Descriptions of the artworks and the general creation process of glass artworks.* *- Interviews with Barbara Nanning discussing her artworks and related creation processes.* *- Knowledge acquisition with Barbara Nanning in which a digital form is used to describe her artworks and related creation processes.* *- Requirements for Data Publication (Moreno-Conde et al., 2019).*

This all typically results in a first set of key terms representing the domain. The granularity of terms and concepts forms a basis for the scope of the ontology. This scope may then further be expanded by using a middle-out approach. This approach allows for high accuracy and increases in control over stability and level of details (Uschold and Gruninger, 1996).

Four types of standard forms are filled in with the artist. The first type specifies information on the artworks themselves, including naming, dimensions, hyperlinks to images, and when possible and legally permitted (AVG; law on privacy) current location. The second type specifies activity-specific information on roles and equipment. The third type specifies individuals involved in various creation processes, with their role and organizational status. The fourth type specifies per creation process of an artwork, the chronological and causal execution order of activities, with corresponding time spans, locations, parties involved and materials processed.

Once done, this results in 946 terms and definitions, known as the glossary of terms (Raven, 2020). The 946 terms cover 794 instances, 121 concepts, and 31 relations or attributes.

User stories and glossary of terms together build up the ontology requirement specification document, as shown in Table 1. This document specifies (semi-formally)

domain, purpose, scope (including the granularity of terms and concepts), sources and level of formality of the ontology to be designed.

3.2 Conceptualization

A set of concept classification trees is constructed based on concepts from the ontology requirement specification document. This results in hierarchies (or sub-class relations) in the 121 concepts, involving entities such as Actors, Periods, Places, Time spans and Representations (Gómez-Pérez et al., 1996; Raven, 2020). An example of a concept classification tree is shown in Figure 3.

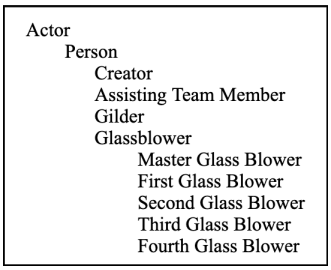


Figure 3: Part of a concept classification tree of the Actor class (indents indicate mutually-disjoint-subclass-of relations)

Based on concept classification trees and glossary of terms a set of data dictionaries, tables of instance attributes, and tables of instances are created (Fernández-López et al., 1997).

Data dictionaries define concepts with related attributes, instances, and relations (Raven, 2020). When constructing data dictionaries, it is important to keep competency questions and requirements in mind, as decisions made compromise information Cultural Heritage experts can query for.

An example of such a decision is not to include concepts related to the fifth requirement defined by Dijkshoorn et al. The support for multiple views may unnecessarily increase complexity when adding concepts like Proxy, View, and Aggregation[13]. Another example is including Creation Events and concepts related to activities part of Creation Events. These concepts allow for structuring information at the level of individual activities, such as materials processed, equipment used, and people involved as shown in Table 2.

Table 2: An example of a data dictionary for the concept Activity with the related relations

Concepts	Instances	Class Attributes	Instance Attributes	Relations
Activity	-	-	-	- Involves - Processes - Requires-Use-Of

Twelve tables are created defining attributes of instances themselves (Raven, 2020). These attributes are all numerical or text-based values (e.g., Has-Name or Has-Depth). An example of such a table is shown in Table 3.

[13] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf

Table 3: An example of an instance attribute table for the attribute Has-Name

Instance Attribute Name	*Has-Name*
Description	*Has-Name is a property which is connected to the name of an entity.*
Value Type	*String*
Unit of Measure	—
Precision	—
Range of Values	—
Default Value	—
Cardinality	*1*
Inferred from Instance Attribute	*Unknown*
Inferred from Class Attribute	*Unknown*
Formula	*Unknown*
To infer	*Unknown*
References	*Unstructured interviews with the artist discussing her artworks and related creation processes.*

A table is defined for each of the 794 instances. These tables define the individual instances by their name, definition, and all relations and attributes with their corresponding values (Raven, 2020). An example of an instance related to New York is shown in Table 4:

Table 4: An example of an instance table for the instance Artwork_17_33

Instance Name	**Description**	**Attributes & Relations**	**Values**
New York	New York is a city in the United States of America	- Location-Of - Located-In - Has-Name - Is-Described-By	- J. Lohmann Gallery, Artwork_44, Collection J. Lohmann Gallery - United States of America - New York - New York is a city in the United States of America.

Fernández et al. also recommend including tables of rules, tables of formulas, tables of class attributes, attribute classification trees, and verb diagrams, on which verb dictionaries and table of conditions are based. However terms, concepts, attributes, and relations provide no suitable basis for constructing these representations. Class attributes (as defined in (Raven, 2020)) are attributes describing the class itself rather than its instances. We identify a limited number of class attributes, including disjointness relations between classes (i.e. City and Country). The majority of class attributes however are not class specific but are generic attributes such as class names, class descriptions, that can later be mapped to standard RDF(S)/OWL properties. They are omitted from the table. Other than these, no additional class attributes were identified due to the high level of variance between instances of classes hindering the possibility of including class-wide attributes. Attribute classification trees and verb diagrams provide overviews of composition relations between attributes and between verbs. These representations are unnecessary as they are not present in the current structure. The domain in question also lacks standard formulas and rules which is why no corresponding representations are defined either.

3.3 Integration

Methontology supports reusing existing ontologies and vocabularies. The integration phase delivers an integration document stating reused terms from other ontologies or vocabularies and corresponding terms conceptualized in earlier stages.

Main sources reused are terminology sources accessible through Termennetwerk, allowing users to easily connect to sources. This is especially useful during ontology construction.

Analysis of the widely-accepted various domain vocabularies and ontologies yields reusable relations, concepts and terms, resulting in links of 124 reused URIs spread over 7 different terminology sources. Specifically, 5 classes were identified as equivalent to CIDOC-CRM classes (Actor, Period, Timespan, Place and Collection). For these, owl:equivalentClass triples are asserted. Likewise 25 classes indicating more specific artistic processes and roles ("Grinding", "Glassblowing", etc.) were mapped to Getty's AAT. Further mappings were made to Wikidata and the Dutch heritage thesaurus indexed by Termennetwerk.

Concepts that remain unmapped include classes that are even more specific. Some of these are more likely to be reused in other cases (e.g. "Assisting_Team_Member", "Cutting_Sheets", "Master_Glass_Blower"), where others are likely very specific to the work and process of Nanning (e.g. "Producing_Coloured_Canes", "AJETO_Glass").

A further 14 triples linking artworks to landing pages and images are added, to allow for further exploration and visualization of query results.

Table 5 shows a part of the integration document (Raven, 2020).

Table 5: Part of the integration document

Ontology related to the works of Barbara Nanning		
Term in Glossary of Terms	**Source to be Reused**	**URI of the term in the Source**
Amsterdam	Geonames	http://sws.geonames.org/2759794
Artwork	Art & Architecture Thesaurus	http://vocab.getty.edu/aat/300133025
Ice Pick	Art & Architecture Thesaurus	http://vocab.getty.edu/aat/300209549
Ice Pick	Wikidata	http://www.wikidata.org/entity/Q586319
Blowing Pipe	Art & Architecture Thesaurus	http://vocab.getty.edu/aat/300022850

3.4 Implementation

The ontology development environment used during the case study is the Protégé ontology editor[14] provided by Stanford University School of Medicine. Protégé meets the requirements of the Methontology methodology and is therefore also used in several ontology engineering studies presenting Methontology based ontologies (Hennig et al., 2009; Moreno-Conde et al., 2019). Representations constructed in the specification, conceptualization, and integration stages are loaded into Protégé as classes, instances, and properties represented in OWL[15].

3.5 Results

Applying Methontology results in an OWL formalization of the ontology, based on constructs generated using Protégé. Additionally, the ontology requirements specification document, the glossary of terms, the set concept classification trees, the data

[14] https://protege.stanford.edu/
[15] https://www.w3.org/OWL/

dictionaries, the table of instance attributes, the table of instances, and the integration document are produced as intermediary results.

The OWL ontology consists of 794 instances, 121 classes, 19 object properties, and 12 data properties. 124 alignments have been made between these entities and URIs from reused sources. The instances populating the ontology specify information describing the artworks of the artist and the related information describing locations, (roles of) individuals, equipment, glass studios, organizations, materials, collections, time spans, representations, creation processes, and the activities part of these creation processes. Figure 4 shows a part of the ontology as visualised using the WebVOWL tool.

A namespace is defined for both the ontology and instances:

<http://semanticweb.cs.vu.nl/test/barbarananning/>.

The resulting IRIs are resolvable: dereferencing these in a browser or application returns the relevant data about the resource identified by that IRI. This makes the knowledge graph a "5-star" Linked Data set. An example of such a resolvable IRI is

<http://semanticweb.cs.vu.nl/test/barbarananning/ARTWORK_17_33_Grinding>

describing the process of grinding for a specific artwork. The complete knowledge graph (ontology and instances) can be accessed through a github repository[16] or browsed via an interactive visualization[17].

Documentation is available in the form of an ontology requirements specification document, the glossary of terms, the set concept classification trees, the data dictionaries, the table of instance attributes, the table of instances and the integration document (Raven, 2020).

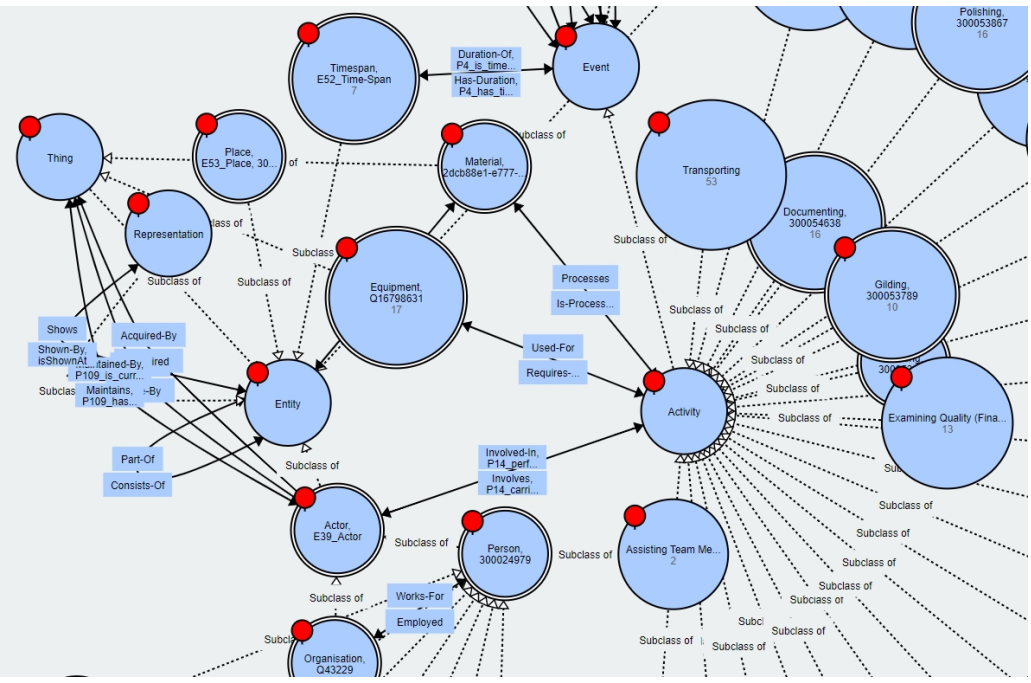


Figure 4: Detail of the WebVOWL visualization of the ontology.

16 https://github.com/biktorrr/barbarananning

17 https://service.tib.eu/webvowl/#iri=https://raw.githubusercontent.com/DRavenNISV/Modeling-Ontologies-for-Individual-Artists/master/BarbaraNanning.owl

4 Validation of the Ontology

The ontology is validated as part of the case study. This is done by appraising how well the ontology requirement specification document adjoins with the ontology implemented. Requirements are based on user stories. Analyzing whether the ontology is able to provide answers to corresponding competency questions forms the basis for validation (Baker and Cheung, 2007). Answers hence corroborate completeness of the ontology development effort.

4.1 User Stories and SPARQL Queries

Giving answers to competency questions requires loading the ontology into a triple store with a SPARQL query environment as a knowledge graph[18]. SPARQL is a semantic query language used to construct queries to RDF-based databases[19].

Examining concepts, instances and properties validates the set of requirements for data publication in the Cultural Heritage domain defined by Dijkshoorn et al.

Interoperability is assured by including an extended top-level concept hierarchy and linking concepts to hierarchically structured vocabularies like the Arts and Architecture Thesaurus, wherever possible[20]. A property hierarchy is not included in the ontology due to a lack of sub-property relations (such as Has-Current-Keeper to Has-Former-Or-Current-Keeper[21]) between individual attributes or relations.

The ontology supports both object and event-based approaches for describing (the creation of) artworks. The Creation Event concept can be used to specify producing artworks and related materials, people and time spans. Object properties like Has-Depth or Made-By can also be used to describe artworks. The Creation Event concept allows for the specification of changes made to artworks over time. Meta-data and representations of artworks are separated by including EDM property Shown-By, connecting (description of) artworks to representations. Context is added to artworks by linking the artwork describing terms to other vocabularies found in Termennetwerk.

As stated in Section 3.2, to include, in this specific case, multiple views of the same artefact from different sources would unnecessarily increase the complexity of the ontology. Therefore, concepts pertinent to the fifth requirement defined by Dijkshoorn et al. are omitted.

4.2 Visualisation of queries

To further illustrate the use of ontologies and knowledge graphs, two examples of SPARQL queries and their visualization. Each is presented in the yasgui.org online SPARQL query builder5 (Rietveld and Hoekstra, 2017), using the aforementioned SPARQL endpoint (yasgui not only allows for querying arbitrary endpoints but also to present results in novel ways).

18 https://semanticweb.cs.vu.nl/test/user/query
19 https://www.w3.org/TR/rdf-sparql-query/
20 http://website.aat-ned.nl/home
21 http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf

Table 6: Overview of competency questions and corresponding queries and prefixes providing complete and reasonable answers

Competency Question	SPARQL Query
CQ-VE-1: What are the top two artworks in the Verre Églomisé collection of Nanning sorted by height?	SELECT DISTINCT ?Artwork ?ArtworkLabel ?Height WHERE { ?Artwork rdf:type bn:Glass_Artwork. ?Artwork rdfs:label ?ArtworkLabel. ?Artwork bn:Has-Height ?Height } ORDER BY DESC(?Height) LIMIT 2
CQ-VE-2: What equipment is used for sandblasting of the artworks within the Verre Églomisé collection of Nanning and how can this equipment be described?	SELECT DISTINCT ?EquipmentLabel ?EquipmentDescription WHERE { ?Activity rdf:type bn:Sandblasting. ?Equipment bn:Used-For ?Activity. ?Equipment rdfs:label ?EquipmentLabel. ?Equipment rdfs:comment ?EquipmentDescription. }
CQ-VE-3: Which artworks in the Verre Églomisé collection of Nanning are currently stored in her private collection?	SELECT DISTINCT ?Artwork ?ArtworkLabel WHERE { ?Artwork rdf:type bn:Glass_Artwork. ?Artwork bn:Part-Of bn:CollectionOfBarbaraNanning. ?Artwork rdfs:label ?ArtworkLabel. }
CQ-VE-4: What is the description of the technique that was last applied in the creation process of the 'Go with the Flow' artwork from the Verre Églomisé collection of Nanning and which parties were involved in this technique?	SELECT ?ActivityType ?DescriptionTechnique ?ActorName WHERE { ?CreationEvent bn:Created bn:Artwork_77. ?Activity bn:Finishes ?CreationEvent. ?Activity rdf:type ?ActivityType. ?ActivityType rdfs:comment ?DescriptionTechnique. ?Activity bn:Involves ?Actor. ?Actor rdfs:label ?ActorName }
CQ-VE-5: What are the links to the digital representations of the artworks in the Verre Églomisé collection?	SELECT ?LabelArtwork ?LabelLinkDigitalImage WHERE { ?Artwork rdf:type bn:Glass_Artwork. ?Artwork rdfs:label ?LabelArtwork. ?Artwork bn:Shown-By ?LinkDigitalImage. ?LinkDigitalImage rdfs:label ?LabelLinkDigitalImage. }
CQ-CS-1: What are the top three artworks in the Coloured Shadows collection of Nanning sorted by descending year of completion?	SELECT DISTINCT ?Artwork ?ArtworkLabel ?LabelYearOfCompletion WHERE { ?Artwork rdf:type bn:Glass_Artwork. ?Artwork rdfs:label ?ArtworkLabel. FILTER(strStarts(?ArtworkLabel, "Coloured Shadows")). ?Artwork bn:Created-By ?CreationEvent. ?CreationEvent bn:Finishes-With ?FinalActivity. ?FinalActivity bn:Has-Duration ?YearOfCompletion. ?YearOfCompletion rdfs:label ?LabelYearOfCompletion. } ORDER BY DESC(?YearOfCompletion) LIMIT 3
CQ-CS-2: What activities are part of the creation process of the 'Beneath the Water' artwork in the Coloured Shadows and where have these activities taken place?	SELECT DISTINCT ?NameActivity ?NameLocation WHERE { ?Artwork rdf:type bn:Glass_Artwork. ?Artwork rdfs:label ?ArtworkLabel. FILTER(strStarts(?ArtworkLabel, "Coloured Shadows Beneath the Water")). ?Artwork bn:Created-By ?CreationEvent. ?CreationEvent bn:Consists-Of ?Activity. ?Activity bn:Located-In ?Location. ?Activity rdfs:label ?NameActivity. ?Location rdfs:label ?NameLocation. }
CQ-CS-3: What are the materials processed in the creation process of the 'Coral Reef' artwork from the Coloured Shadows collection of Nanning and during which activities were these materials processed?	SELECT DISTINCT ?NameActivity ?NameMaterial WHERE { ?Artwork rdf:type BarbaraNanning:Glass_Artwork. ?Artwork rdfs:label ?ArtworkLabel. FILTER(strStarts(?ArtworkLabel, "Coloured Shadows Coral Reef")). ?Artwork BarbaraNanning:Created-By ?CreationEvent. ?CreationEvent BarbaraNanning:Consists-Of ?Activity. ?Activity BarbaraNanning:Processes ?Material. ?Activity rdfs:label ?NameActivity. ?Material rdfs:label ?NameMaterial. }

Company based gallery visualization. Figure 5 shows a gallery view of the objects made at glass house Jiří Pačinek. The query uses links from Wikidata (where this company is defined), through the knowledge graph, using links to the external images hosted on the Web. The results are displayed using yasgui's built-in gallery display.

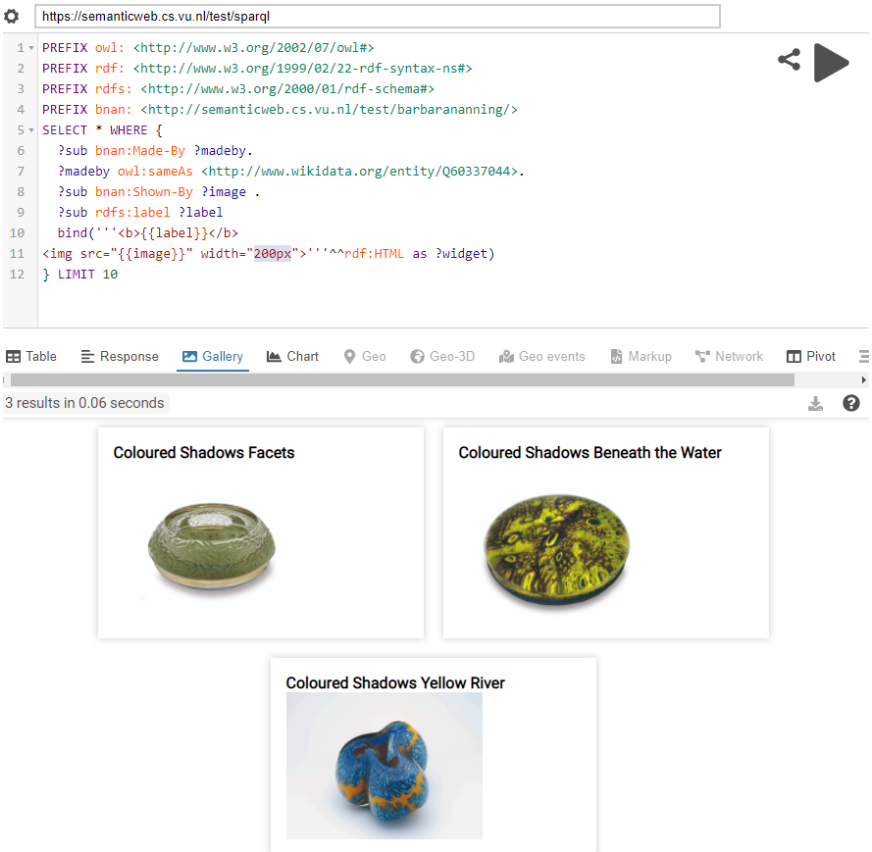


Figure 5: Result for the query to display objects involving a specific person

Geographic visualization. Figure 6 shows the query to retrieve all geographic locations associated with a specific artwork. Through links with the external vocabulary GeoNames, geographic coordinates can be retrieved. This allows using the built-in Geo visualization of yasgui to present the locations on a map.

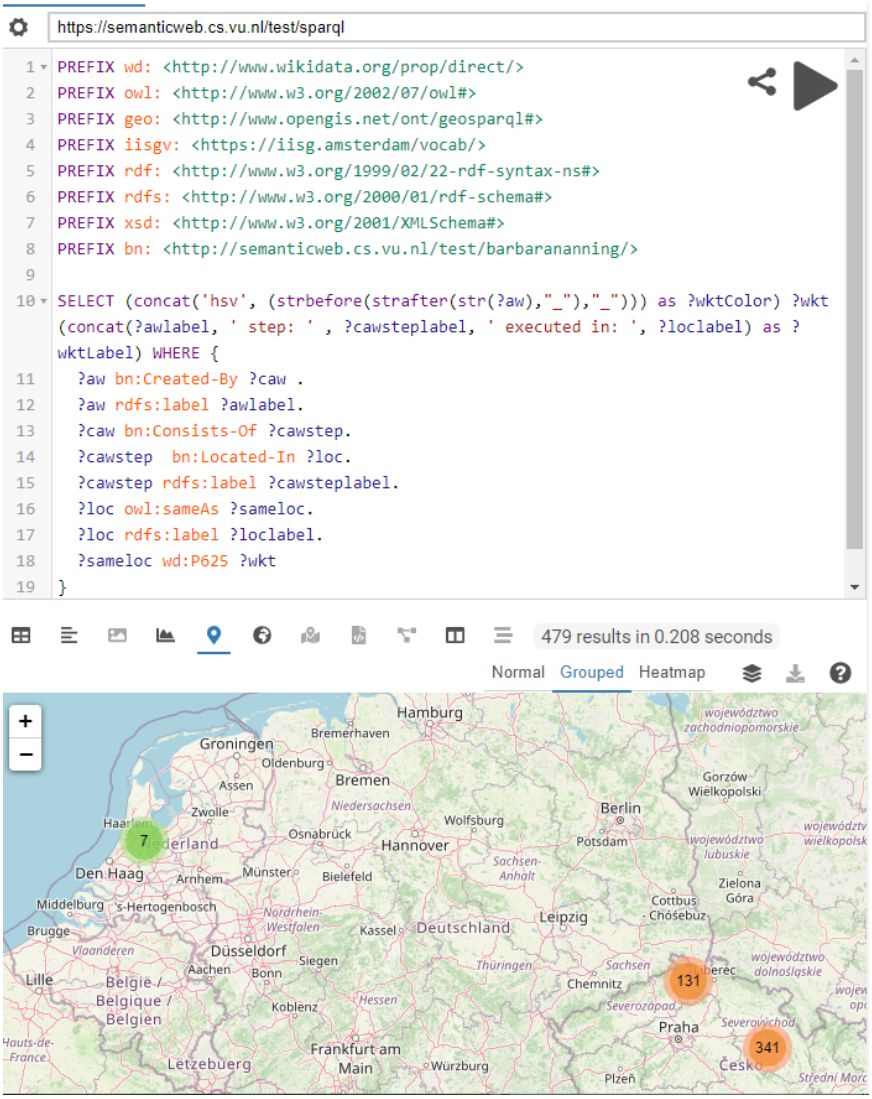


Figure 6: Result for the query to display different locations of a specific artwork

5 Discussion

This paper proposes a re-usable method for developing an individual artist ontology in a specific case.

An adapted version of Methontology is applied to creating an ontology specifying and providing access to enriched information related to the works of the Dutch ceramic-glass artist Barbara Nanning, covering locations, (roles of) individuals, equipment, organizations, materials, collections, time spans, and creation processes.

In contrast to CIDOC-CRM and EDM, the ontology is able to specify information related to activities part of the creation process of an artwork at a high level of detail by creating instances of types Creation Event and Activity.

The method presupposes access to the artist(s) or to detailed resources describing creative and productive processes. Concepts defined at top-level (Entity, Thing, Collection, Artwork, Equipment, Material, Period, Event, Creation Event, Actor, Person, Organization, Place, Country, City, Time span, Representation, and Digital Image) and relations (Processes, Involved-In, Created or Has-Duration) are artist indepen-

dent and can further be extended by including artist-specific concepts similar to the concept classification trees defined in this study. More specific concepts related to glass-processes can be reused by other artists working in the same medium, while concepts that are idiosyncratic to the work and process of Nanning are less likely to be re-used. For new cases, we can therefore expect that the method will result mostly in concepts that are specific to that artist, but that these can be mapped to the already existing more generic ones presented here. Here we do expect a trade-off between the specificity of the ontology for the new artist and the level of re-use of previously defined concepts.

Future work includes research into the application of this and other ontologies in information extraction tasks for different end-users, as well as investigating whether more general design patterns can be derived from the ontology that can be applied to other domains, in both historical and present-day art, craft and design contexts.

Acknowledgments

This research was supported through the Netherlands Institute for Sound and Vision intern program, backing of the Dutch Creative Industries Fund and the Dutch Culture Fund as well as through the CLARIAH-PLUS programme (NWO Grant 184.034.023). Special thanks to Barbara Nanning for participating in the knowledge acquisition activities of the case study.

References

Alia Amin, Jacco Van Ossenbruggen, Lynda Hardman, and Annelies van Nispen. Understanding cultural heritage experts' information seeking needs. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 39–47, 2008.

Christopher JO Baker and Kei-Hoi Cheung. *Semantic web: revolutionizing knowledge discovery in the life sciences*. Springer, 2007.

Chris Dijkshoorn, Lora Aroyo, Jacco Van Ossenbruggen, and Guus Schreiber. Modeling cultural heritage data for online publication. *Applied Ontology*, 13(4):255–271, 2018.

Martin Doerr. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75, 2003.

Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, Carlo Meghini, and Herbert Van de Sompel. The europeana data model (edm). In *World Library and Information Congress: 76th IFLA general conference and assembly*, volume 10, page 15, 2010.

T. Eliëns. *Barbara Nanning*. Waanders uitgeverij, Zwolle, 2019.

Mariano Fernández-López, Asunción Gómez-Pérez, and Natalia Juristo. Methontology: from ontological art towards ontological engineering. *AAAI Technical Report SS-97-06. Retrieved from http://oa.upm.es/5484/*, 1997.

Asunción Gómez-Pérez, Mariano Fernández, and A de Vicente. Towards a method to conceptualize domain ontologies. *European Coordinating Committee for Artificial Intelligence (ECCAI)*, 1996.

Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

Christian Hennig, Nico Adams, and Gesine Hansen. A versatile platform for comprehensive chip-based explorative cytometry. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry*, 75(4):362–370, 2009.

Eero Hyvönen. Publishing and using cultural heritage linked data on the semantic web. *Synthesis lectures on the semantic web: theory and technology*, 2(1):1–159, 2012.

Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, Nurfadhlina Mohd Sharef, et al. An analysis of ontology engineering methodologies: A literature review. *Research journal of applied sciences, engineering and technology*, 6(16):2993–3000, 2013.

Reza Kalbasi, Krzysztof Janowicz, Femke Reitsma, Luc Boerboom, and Ali Alesheikh. Collaborative ontology development for the geosciences. *Transactions in GIS*, 18(6): 834–851, 2014.

Mariano Fernández López, Asunción Gómez-Pérez, Juan Pazos Sierra, and Alejandro Pazos Sierra. Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems and their applications*, 14(1):37–46, 1999.

A Moreno-Conde, Carlos Luis Parra-Calderón, S Sánchez-Seda, GA Escobar-Rodríguez, Marta López-Otero, Lorena Cusso, R del Cerro-García, M Segura-Sánchez, L Herrero-Urigüen, N Martí-Ras, et al. Itemas ontology for healthcare technology innovation. *Health research policy and systems*, 17(1):1–10, 2019.

Kamran Munir and M Sheraz Anjum. The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics*, 14(2):116–126, 2018.

M Cristina Pattuelli. Modeling a domain ontology for cultural heritage resources: A user-centered approach. *Journal of the American Society for Information Science and Technology*, 62(2):314–342, 2011.

Daan Raven. D. raven supplementary material - modeling ontologies for individual artists a case study of a dutch ceramic-glass sculptor, May 2020. URL https://figshare.com/collections/D_Raven_Supplementary_Material_-_Modeling_Ontologies_for_Individual_Artists_A_Case_Study_of_a_Dutch_Ceramic-Glass_Sculptor/4990448/2.

Laurens Rietveld and Rinke Hoekstra. The yasgui family of sparql clients 1. *Semantic Web*, 8(3):373–383, 2017.

C Syamili and RV Rekha. Developing an ontology for greek mythology. *The Electronic Library*, 2018.

T. te Duits. *Barbara Nanning, Evolution*. ISBN:9789090173696, Nanning, 2003.

Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(2):93–136, 1996.

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

Judging a Book by its Criticism: A Digital Analysis of the Professional and Community Driven Literary Criticism of the Ingeborg-Bachmann-Preis

Lore De Greve and Gunther Martens

Ghent University

You should not judge a book by its cover, but by its content. However, in reality many books are judged based on other criteria even before being read, when potential readers rely on book reviews or ratings to decide whether or not to read a specific book. How a book has been judged before may influence how it is judged by others. This literary criticism may originate from either professional or layperson critics. According to Beth Driscoll, two of the increasingly influential phenomena in modern publishing are "literary prizes and social media, both of which draw together participants from multiple areas of literary culture" (Driscoll, 2013, 103). In this article, we shall therefore examine the Ingeborg-Bachmann-Preis, a literary prize which receives a lot of attention in both the academic field and on social media. This prize is awarded during the Tage der deutschsprachigen Literatur (TDDL), an annual and multi-day literary festival. We manually annotated a corpus of TDDL-related Tweets, Instagram posts and Goodreads reviews, as well as the descriptions of the official jury discussions. We then performed a fine-grained aspect-based sentiment analysis (ABSA), which allows us to gain deeper insight in the content of the literary discourse surrounding the Bachmann-Preis and in the evaluative criteria used by both professional (jury discussions) and layperson critics (social media contributions). We argue that there are noticeable differences between the Twitter, Instagram, Goodreads and jury discourse surrounding the Ingeborg-Bachmann-Preis, as well as between the evaluative criteria used by professional and layperson critics. Our analysis will also demonstrate that, contrary to the common prejudice, layperson literary criticism is not necessarily less critical than professional criticism and does in fact concern itself with aesthetic principles, such as form, style etc., instead of solely assessing a literary text based on the judgement of the professional jury.

Keywords: digital humanities; Aspect-Based Sentiment Analysis (ABSA); sentiment mining; Ingeborg-Bachmann-Preis; Tage der deutschsprachigen Literatur (TDDL); literary prize; social media; Twitter, Instagram; Goodreads; axiology; layperson criticism; literary criticism; literature

1 Introduction

Don't judge a book by its cover. Instead, according to the implications of this popular adage, it would be better to judge it by its content. However, these are not the only ways in which books are being judged. Many books are being judged based on other criteria even before being read, when potential readers rely on book reviews or ratings to decide whether or not to read a specific book. How a book has been judged already may influence how it is judged by others; literary criticism is layered and interwoven. This literary criticism may originate from different sources, from either professional or layperson critics. Although Pierre Bourdieu (1993) has argued that the consecration by authorised gatekeepers is decisive for the symbolic capital of a literary text , such as literary prizes, layperson critics act as new literary gatekeepers and cultural transmitters regarding the evaluative talk of literature. As such they rely on the proliferation of social media and peer-to-peer-recommendation platforms responsible for the digitisation of the public sphere to take part in the literary criticism (De Greve and Martens, 2021a,b, in press, 2022; Kostial, 2021). Beth Driscoll calls literary prizes one of two increasingly influential phenomena "[a]mong the rapid changes that characterise publishing in the twenty-first century" (Driscoll, 2013, 103). The other one is social media and she argues that "both [...] draw together participants from multiple areas of literary culture" (Driscoll, 2013, 103). Indeed, in the past decades, the academic interest in particular and in literary prizes in general as consecrators of literature has increased (Auguscik, 2017; Braun, 2014; Chenaux and Beck, 2015; Childress et al., 2017; Dekker and de Jong, 2018; Ducas, 2013; Emmerich, 2012; English, 2009; Heymans, 2001; Irsigler and Lembke, 2014; Kennedy-Karpat and Sandberg, 2017; Meyers, 2007; Sapiro, 2016; Todd, 1996; Ulmer, 2006) and so has the interest – both positive and negative – in layperson literary criticism and social media or social platforms (Allington, 2016; Álvarez-López et al., 2018; Driscoll, 2013, 2014; Jaakkola, 2019; Kellermann and Mehling, 2017; Kellermann et al., 2016; Kousha and Thelwall, 2016; Kousha et al., 2017; Pressman, 2020; Schneider, 2018; Steiner, 2008; Thelwall and Kavyan, 2017; Thomalla, 2018; Walsh and Antoniak, 2021; Wang et al., 2019; Weber and Driscoll, 2019), e.g. book blogs, Goodreads, Twitter, Instagram and Amazon.

One such literary prize which receives a considerable amount of academic interest (Bogaert, 2017; Leinen, 2010; Moser, 2004; Rahmann, 2017; Rebien, 2012; Röhricht, 2016) and social media attention is the German Ingeborg-Bachmann-Preis. This prominent prize is awarded annually during the Tage der deutschsprachigen Literatur or TDDL (translation: "Days of German-language Literature"), a multi-day literary festival and competition that takes place in Klagenfurt, Austria. The professional jury nominates 14 authors to write a short narrative text for the competition. During the event all nominated contenders read their unpublished text in front of a live audience and the jury. Afterwards, the text is discussed and criticised by the professional jury in the presence of the author and the live audience, but increasingly so by an online audience as well, under the #tddl-hashtag. The participation of the audience is stimulated by the organisers, on the one hand because the high monetary and symbolic value of the audience award "raises its prestige and the stakes, as the audience's participation and decision carry more weight" (De Greve and Martens, 2021a, 99) and, on the other hand "indem sie die Verwendung des offiziellen Hashtags, #tddl, bei der Diskussion über den Preis in sozialen Medien zunehmend fördern" (De Greve and Martens, in press, 2022, translation: "By increasingly promoting the use of the official hashtag, #tddl, when discussing the award on social media"). In 2021, they even expanded

the online participation and made "eine Auswahl an Postings [...] Teil der Sendung" (translation: "a selection of posts [...] part of the programme") and included "eine 'Frage des Tages' [...], die über Social-Media-Kanäle öffentlich debattiert werden kann und die in der Mittagspause, moderiert von Cecile Schortmann aufgegriffen und besprochen wird" (translation: "a 'question of the day' [...] which can be publicly debated via social media channels and which is taken up and discussed during the lunch break, moderated by Cecile Schortmann").[1]

In "#Bookstagram and Beyond" (De Greve and Martens, 2021b), we discussed the position and distinctive features of the Ingeborg-Bachmann-Preis within the field of literary prizes, as well as their influence on the online presence of the prize. Our corpus for this article consisted of all Tweets and Instagram posts from 2007 to 2017 that contained a TDDL-related query or hashtag and the Dutch, German and English reviews of all winning texts (as well as resulting novels) on the peer-to-peer recommendation platform Goodreads from this time period. We also argued that these three social media platforms, due to their distinct limitations and expectations concerning the length, type and subject which shape the content of the contributions, have a distinct way of communicating. We therefore examined the corpora by means of a digital corpus analysis and an examination of word frequencies using both Voyant Tools and AntConc.[2] In this article, however, we will employ a different method in order to gain deeper insight in the content of the literary discourse surrounding the TDDL by performing a fine-grained aspect-based sentiment analysis (ABSA) on a smaller manually annotated corpus consisting of TDDL-related Tweets, Instagram posts and Goodreads reviews from 2019, along with descriptions of the jury discussions from the same year. In future steps in our research, the annotated corpora presented here will be used as training data to set up a semi-supervised learning system that will be used to perform an automatic aspect-based sentiment analysis on the corpora containing all data from 2007-2017.[3] We previously employed this annotation method in "Wertung von Literatur 2.0" (De Greve and Martens, in press, 2022), where we exclusively analysed the content of the 2019 TDDL-Twitter-discourse. The results of this analysis will here be supplemented with the previously mentioned three additional corpora, which will enable us to compare the evaluative literary criteria of both professional and layperson critics as well as the differences within layperson criticism across social media platforms. Consequently, we will detect which sentiment is expressed about a certain "aspect" or topic (e.g. nominated author, book, jury, audience etc.) and by whom. In this article, we will thus expand the scope of our research by performing a fine-grained aspect-based sentiment analysis in order to compare the previously annotated 2019 Twitter corpus with three additional newly annotated corpora, consisting of Instagram posts, Goodreads reviews and the jury discussion. We argue that there are noticeable differences between the Twitter, Instagram, Goodreads and jury discourse surrounding the Ingeborg-Bachmann-Preis, as well as regarding the evaluative criteria of professional and layperson critics' literary criticism.

---

[1] "45. Bachmannpreis wieder digital". *Bachmannpreis*, 20 May 2021, `https://bachmannpreis.orf.at/stories/3104649/`, last accessed Oct. 2021.

[2] Open-source digital environments and tools for web-based text reading and analysis as well as corpus analysis.

[3] When training a system for machine learning, the manually annotated training data should be separated from the corpus on which the system is to be run. Instead of selecting and annotating part of our 2007-2017 corpus, it was therefore preferable to annotate a different, but similar, corpus – hence the 2019 data – and use this as training data, so that we can avoid having to exclude any part of the 2007-2007 from the automatised ABSA.

2 Composition of Corpora

Because of our aim to study the evaluative literary criteria used by both professional and layperson critics and to engage with the differences in evaluation practices across platforms and media in the context of the Ingeborg-Bachmann-Preis, these corpora consist, on the one hand, of the official description of the jury discussions, and, on the other hand, of the lay discourse surrounding the prize on Twitter, Instagram and Goodreads.[4] We thus examine four distinct corpora hailing from four different sources and/or platforms. For this case study, in order to narrow down the corpora, we decided to focus on the data and posts from 2019. In the future, this will be expanded to data from 2007 to 2017 as well. As mentioned previously, the jury discussions on the nominated texts are broadcast live. However, an official description and summary of each jury discussion per nominated text is also published on the Prize's website. As only textual data is annotated for our research project, these descriptions serve as a direct representation of the jury discussions. The Twitter corpus consists of all Tweets created during the TDDL (26th-30th June) in 2019 that contained either the query "tddl" or the official #tddl-hashtag, resulting in a total of 4352 Tweets.[5] Similar criteria were used to select the Instagram posts, namely all posts created during the literary festival in 2019 that include the #tddl-hashtag, comprising 191 posts.[6] The reason for only taking those Tweets or posts into account that were posted in this period is twofold.[7] Firstly, in both cases, the majority of Tweets and posts is written during the TDDL and these therefore contain the so-called "Sofortkritik/-kommentierung" or "Stehgreifkritik" (translation: "immediate criticism" or "criticism on the spot") that resembles the set-up of the prize itself, in which the jury immediately and (relatively) spontaneously discuss and criticise the texts that have just been read.[8] Secondly, posts and Tweets posted before or after the event tend to contain more irrelevant information or focus less on the TDDL themselves. Lastly, we extracted all German Goodreads reviews of the novels based on the texts that were nominated for the Bachmann-Preis that year.[9] This constitutes to the reviews for Leander Fischer's *Die Forelle* (Text: "Nymphenverzeichnis Muster Nummer eins Goldkopf"), Tom Kummer's *Von schlechten Eltern*, Lukas Meschik's *Vaterbuch* (Text: "Mein Vater ist ein Baum") and Martin Beyer's *Und ich war da*. It is necessary to bear in mind that because the competing texts are (at the moment of the TDDL) unpublished, short texts that are not always transformed

---

[4] For more information on how we collected data, please read our article "#Bookstagram and Beyond: The Presence and Depiction of the Bachmann Literary Prize on Social Media (2007-2017)" (De Greve and Martens, 2021b).

[5] Please note that this Twitter corpus, as well as its annotation system and the results of this annotation are the same as in our article on the "Wertung von Literatur 2.0: Eine digitale und literatursoziologische Analyse der Online-Twitter-Diskussion zu den Tagen der deutschsprachigen Literatur #tddl" (De Greve and Martens, in press, 2022). They are presented here in comparison to the other three corpora.

[6] Contrary to Twitter, the search function of Instagram only enables the search for hashtags or profiles, not queries.

[7] See also: "Wertung von Literatur 2.0: Eine digitale und literatursoziologische Analyse der Online-Twitter-Diskussion zu den Tagen der deutschsprachigen Literatur #tddl" (De Greve and Martens, in press, 2022).

[8] For more information on the role of "Sofortkritik" regarding the Bachmann-Preis and its Twitter discourse, please see: "ICH WÜRDE AM LIEBSTEN MIT DER JURY DISKUTIEREN! #TDDL" (Bogaert, 2017, 42-45).

Since 1996, the Bachmann-Preis jury receives the texts one week in advance of the author readings, although the principle of spontaneity still applies to the actual discussion between the jury members.

[9] The reviews were collected on 19th May, 2021. Please note that reviews may have been edited, removed or added by users since then.

into a published novel, not all of them may have a Goodreads book page, as is the case here (De Greve and Martens, 2021a, 107-108).

3 Annotation of the Corpora and Results

The annotation system was designed specifically to be applicable not only to Tweets about the Bachmann-Preis, but – providing some adjustments – to both other prizes and social media platforms as well (De Greve and Martens, in press, 2022), which is exactly what was executed here by including Instagram posts, Goodreads reviews and the jury discussions. We distinguish eight main aspect categories, namely "Text", "Reference", "Reading", "Onsite Audience", "Meta", 'Jury", "Irrelevant", "Contender", and 40 subcategories (see Figure 1). The subcategories are relatively self-explanatory but will be briefly touched upon. The "Text"-category refers to the nominated texts and has ten subcategories that encompass different elements of the texts, such as its title, quotes, the point of view or narration, motifs or themes, the language use or style, the general content or plot, the text in general, the form or structure, the flow, rhythm and punctuation and its characters. The "Reference"-category encompasses the references or comparisons to other authors or literary works, musicians or music, film or television etc. and is limited to two subcategories, one for the comparisons made by the layperson critics themselves (which implies an evaluation of the text) and one for those of the professional jury. When the canon or references or the jury are mentioned, this goes hand in hand with an evaluation of the jury discussion and valuation. The third category concerns the author readings and can be divided into mentions of the pronunciation, intonation and understandability, of the reading in general and of the flow, rhythm and punctuation of the reading. "Onsite Audience" represents the live audience that is present at the studio during the event. This category has three subcategories: the audience's behaviour (e.g. coughing, taking pictures, applauding...), the audience in general and its age, appearance and clothing. Then there is the "Meta"-category as well, which refers to any kind of reference to the circumstances of the event or prize itself, for example the environment, namely the weather and location, the video portraits of the competing authors that are shown before the author readings, technology and social media (e.g. the livestream, website troubles, ...), ritualised side events like the TDDL swimming competition in Klagenfurt's topical *Wörthersee*,[10] the opening speech, music played during the event, the montage of the broadcast and livestream, the event or prize itself, literature and literary prizes in general and all aspects related to the competition, such as discussions about the long- and shortlist, the voting, the winner and the award ceremony. It is important to keep in mind that when the competition (the "Meta – Competition"-subcategory) is evaluated, this is at the same time an indirect value judgement regarding the professional jury's own valuation. If the social media user expresses their happiness that a certain competitor won the Bachmann-Preis, they implicitly convey their agreement with the jury's decision. The Tweet "#tddl das freut mich sehr! Bachmannpreis für #birgitbirnbacher" (translation: "#tddl I am very pleased! Bachmann-Preis for #birgitbirnbacher") expresses a positive sentiment about two explicit aspects, namely (the result of) the competition and the contender.[11] The Twitter-user is happy that Birgit Birnbacher was voted the winner

[10] "Das Wettschwimmen", which also qualifies as a pun on the fact that some writers object to the competitive nature of the "live event", the "Wettlesen" (competitive reading).

[11] @Marina_artblue. "#tddl das freut mich sehr! Bachmannpreis für #birgitbirnbacher". *Twitter*, 30 Jun. 2019, https://twitter.com/Marina_Buettner/status/1145260384510763008, last accessed 4

Fig. 1: Aspect Categories	
Main Categories	**Subcategories**
Text	Title Quote Point of View & Narration Motifs & Themes Language & Style General Content & Plot General Form Flow, Rhythm & Punctuation Characters
Reference	Jury Audience
Reading	Pronunciation, Intonation & Understandability General Flow, Rhythm & Punctuation
Onsite Audience	General Behaviour Age, Appearance & Clothing
Meta	Weather & Location Videoportrait Technology & Social Media Side Event Opening Speech Music Montage Main Event Literature & Literary Prizes Competition
Jury	Voice & Language Use Quote General Discussion & Valuation Behaviour Age, Appearance & Clothing
Irrelevant	
Contender	Voice & Language Use Quote General Gender Age, Appearance & Clothing

Figure 1: A table containing all main aspect categories and subcategories.

of the competition. Additionally, this Tweet also implicitly expresses their approval of the jury's value judgement in pronouncing Birnbacher as the winner. Besides this, there is a main category regarding the jury with six subcategories for their voice or language use (the latter often concerns the use of dialect), quotes, the jury in general, the jury discussions and their evaluation of the texts, their behaviour as well as their age, appearance and clothing. The "Contender"-category is used for mentions of the nominated authors and has five subcategories, namely their voice or language use, primarily mentioned with regards to their author readings and video portraits, quotes,[12] the author in general, their gender as well as their age, appearance and clothing. And lastly, there is a category for irrelevant Tweets as well, specifically Tweets that either clearly did not discuss the TDDL or in which it was not immediately evident how they could be connected to the event.

Results from other annotation projects have often shown that annotating fine-grained aspect categories is complex and that the accuracy of the automatic category prediction tends to decrease as complexity increases. However, this does not mean a fine-grained annotation of aspect categories is out of reach. In their article "A Proposal for Book Oriented ABSA: Comparison over Domains", Álvarez-López et al. (2018) have illustrated that ABSA, using multiple subtasks such aspect extraction, category detection, and sentiment analysis, can be used to analyse a data set consisting of Amazon book reviews and to identify multiple aspect categories related to the book and its content, such as "general", "author", "title", "audience", "quality", "structure", "length", "characters", "plot", "genre" etc., similar categories as the ones we employ within the "Text"-category. A lot depends on the consistency of the annotation, especially for categories that are less distinct or whose target words are more varied, such as the "Text – General Content & Plot" or the "Jury – Discussion & Valuation" subcategories.[13] The exact description of and the vocabulary used to describe the content of a text or the jury discussion often consist of a longer span containing more and diverse vocabulary. For many other categories the same target words are repeatedly used to refer to the aspect, such as the variants of "Text" or "Buch" ("book") for the "Text – General"-category, mentions of "Thema(tisch)" ("theme/thematically"), "Motif" ("motif") etc. regarding the "Text – Themes & Motifs"-subcategory, references to contender or jury names, etc. In the case of more "vague" or varied target words, we therefore decided to annotate the parts of these longer spans that are more likely to reoccur and be automatically detected. In this sentence from an Instagram post, for example: "Es geht um den Tod, Vergehen, schlechtes Gewissen, weil man sich vom Vater abwendete" (translation: "It is about death, transgression, bad conscience, because they distanced themselves from the father"),[14] we would solely label "Es geht um" as "Text – General Content & Plot", which is a more frequently recurring phrase. Regarding the jury discussion, if it pertains to a description of the discussion instead of a direct reference of a word like "Diskussion", we can annotate the mention of the name of a jury member in combination with a verb that communicates the expression of an opinion. In "Winkels meint, der Text bringe Mut auf. #tddl" (translation: "Winkels maintains that

Apr. 2022.

[12] In this case we do not include quotes from their nominated text, but only quotes from, for example, interviews etc.

[13] For more information on the exact annotation method, please see our article "Aspect-Based Sentiment Analysis for German: Analyzing 'Talk of Literature' Surrounding Literary Prizes on Social Media" (De Greve et al., 2021).

[14] @literaturwelten_com. '#lukasmeschik präsentiert eine Hommage auf seinen Vater'. *Instagram*, 29 Jun. 2019, `https://www.instagram.com/p/BzSpfsTjAsb/`, last accessed 5 Apr. 2022.

the text inspires courage. #tddl"),[15] we would only label "Winkels meint" as "Jury – Discussion & Valuation". We have already tested the performance of automatic aspect term category prediction using the manually annotated 2019 TDDL-Twitter corpus, achieving a rather high accuracy of 83% for the prediction of the main aspect categories and 73% for the prediction of the fine-grained subcategories (De Greve et al., 2021), illustrating that the application of the categories is indeed fairly reliable and effective.

We annotated the Instagram posts in the exact same manner as the Tweets, as both are generally shorter social media contributions with a length limitation, and tagged the aspects and sentiment on a Tweet or post level and included implied aspects as well (De Greve and Martens, in press, 2022). Some adjustments were made, however, regarding the corpora of Goodreads reviews and the description of the jury discussions. Firstly, both tend to be longer than the Tweets or posts, especially so in the case of the discussion descriptions, and tagging on a review or article level would therefore exclude more information. Consequently, for these two corpora, each aspect was tagged, regardless of any repeated mentions. Secondly, in the description of the jury discussion, the jury members are naturally mentioned frequently. However, as we employ the description as a stand-in for the actual jury discussion, mentions of other jury members in this corpus were only tagged and annotated if they were either mentioned in a quote or if a jury member commented on, agreed or disagreed with the other members of the jury. This means that in the sentence "Klaus Kastberger vermisste eine österreichische Note des Texts" (translation: "Klaus Kastberger missed an Austrian touch in the text") the word group "Klaus Kastberger vermisste" did not receive a "Jury – Discussion/Valuation"-label, as it is only a pointer attributing the utterance to a jury member.[16] Contrary to this, the word groups such as "Hildegard Keller verstand den Einwand" (translation: "Hildegard Keller understood the objection")[17] and "Dem widersprach Wilke" (translation: "Wilke contradicted this")[18] do describe jury members (dis)agreeing with each other and thus evaluating each other's arguments in the discussion. As a consequence, these would be tagged as "Jury – Discussion/Valuation", with a positive and a negative sentiment expression respectively.

To facilitate the comparison of the differently sized corpora, the results will be shown as percentages, as opposed to absolute numbers, and the irrelevant Tweets are excluded from the graph (De Greve and Martens, in press, 2022).[19] The first four graphs (Figures 2, 3, 4 and 5) visualise how often a main aspect category is mentioned in each of the corpora and which percentage of the mentions is either positive, neutral or negative. When comparing the results, it becomes clear that the corpora have different focal points. In the Twitter discourse, attention is divided over several main categories, mainly the "Meta"-, "Text"-, "Jury"- and "Contender"-categories. In her

[15] @literaturcafe. 'Winkels meint, der Text bringe Mut auf. #tddl.' *Twitter*, 28 jun. 2019, https://twitter.com/literaturcafe/status/1144542048311218178, last accessed 5 Apr. 2022.

[16] "Jurydiskussion Andrea Gerster". *Bachmannpreis.orf.at*, 27 Jun. 2019, https://bachmannpreis.orf.at/v3/stories/2987586/, last accessed 8 Oct. 2021.

[17] ibid.

[18] ibid.

[19] Underneath each graph we have included a table displaying these percentages. How often a certain aspect is mentioned in total (as %) can be seen in the graph and was calculated by adding the percentage (not rounded) of positive, neutral and negative mentions of this aspect. The total percentage of positive, neutral and negative mentions can be calculated by adding all percentages of a specific sentiment across the aspect main (Figures 2, 3, 4 and 5) or subcategories (see Figures 6, 7, 8 and 9; Figures 10, 11, 12 and 13; Figures 14, 15, 16 and 17; as well as 18, 19, 20 and 21) that are displayed in a specific graph.

research on the Bachmann-Preis, Xiana Bogaert compared the thematic tendencies of the jury discussions between 2010 and 2014 with hand-selected Tweets (Bogaert, 2017, 7) and came to the conclusion that Twitter-users "hauptsächlich die Jurydiskussionen des Bachmannpreises zum Gegenstand ihrer Kritik herananziehen" (Bogaert, 2017, 54, translation: "mainly draw on the jury discussions of the Bachmann-Preis as the object of their critique") in order to comment and criticise the texts indirectly. Consequently, their evaluative process and criteria are influenced by those of the jury (Bogaert, 2017, 56).[20] Although they indeed discuss the jury most frequently (28,63%), the large percentage of "Text"-mentions (25,72%) indicates that the Twitter-users also discuss and criticise the texts themselves (De Greve and Martens, in press, 2022). This discourse also contains comparatively more negative sentiments than the others: 41,83% of the mentions are negative in the Twitter discourse, in comparison to 28,51% (more than 10% less) in the jury discussions, and only 8,89% on Instagram and 12,03% on Goodreads. The Instagram corpus, on the other hand, appears to focus most on the event itself – the "Meta"-category (54,13%)- and to a smaller degree on the text (16,21%), jury members (10,71%), nominated authors (9,48%) and author readings (7,35%). Besides this, the mentions on Instagram are primarily neutral (59,94%).[21] The Goodreads reviews and jury discussions, however, are more positive (55,69% and 46,62%) and both focus predominantly on the participating texts (87,98% and 79,49%). As a result, the Goodreads reviews and the valuation of the professional jury appear to be more similar in their shared fixation on the texts themselves. In itself, this aligns with the expectations regarding the task of the jury as well as the focus of a book recommendation platform: it is the jury's prerogative and responsibility to discuss the text – and, to a lesser extent, the author's reading. Similarly, on a peer-to-peer book recommendation platform, users are expected to write book reviews, focusing on the literary work. Twitter- and Instagram-users, on the other hand, do not have this restriction and they are therefore free to comment on more aspects of the literary competition. For this research, we will now expound on the four most dominant main aspect categories, namely the text, which will grant insight into the various literary criteria at work in the different corpora, as well as the "Meta"-category, the jury and the contenders.

[20] See also: "Wertung von Literatur 2.0: Eine digitale und literatursoziologische Analyse der Online-Twitter-Diskussion zu den Tagen der deutschsprachigen Literatur #tddl" (De Greve and Martens, in press, 2022).

[21] Neutral mentions can be interpreted as informative statements (cf. Bogaert, 2017, 59-60).

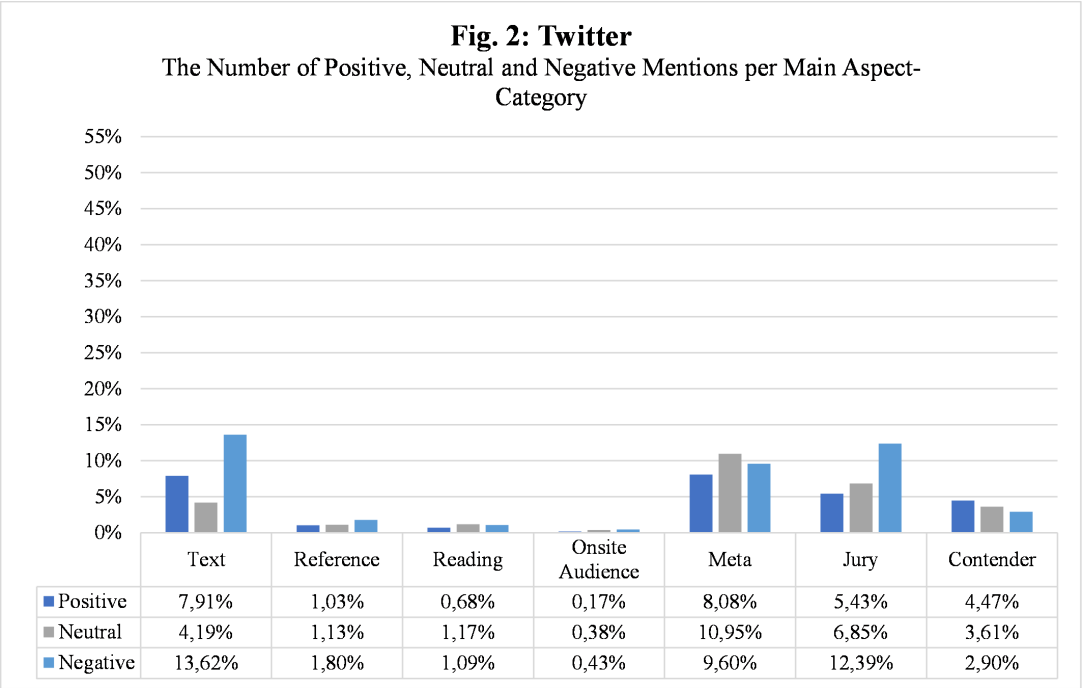


Figure 2: A graph containing the percentage of positive, neutral and negative mentions of each main aspect category in the Twitter discourse.

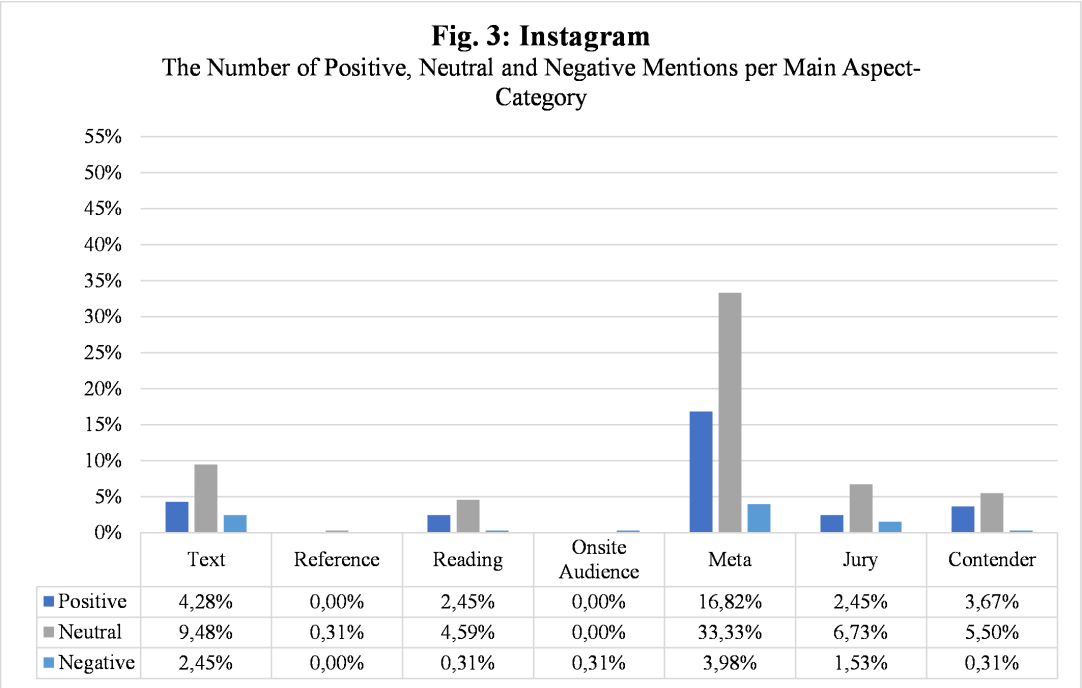


Figure 3: A graph containing the percentage of positive, neutral and negative mentions of each main aspect category in the Instagram discourse.

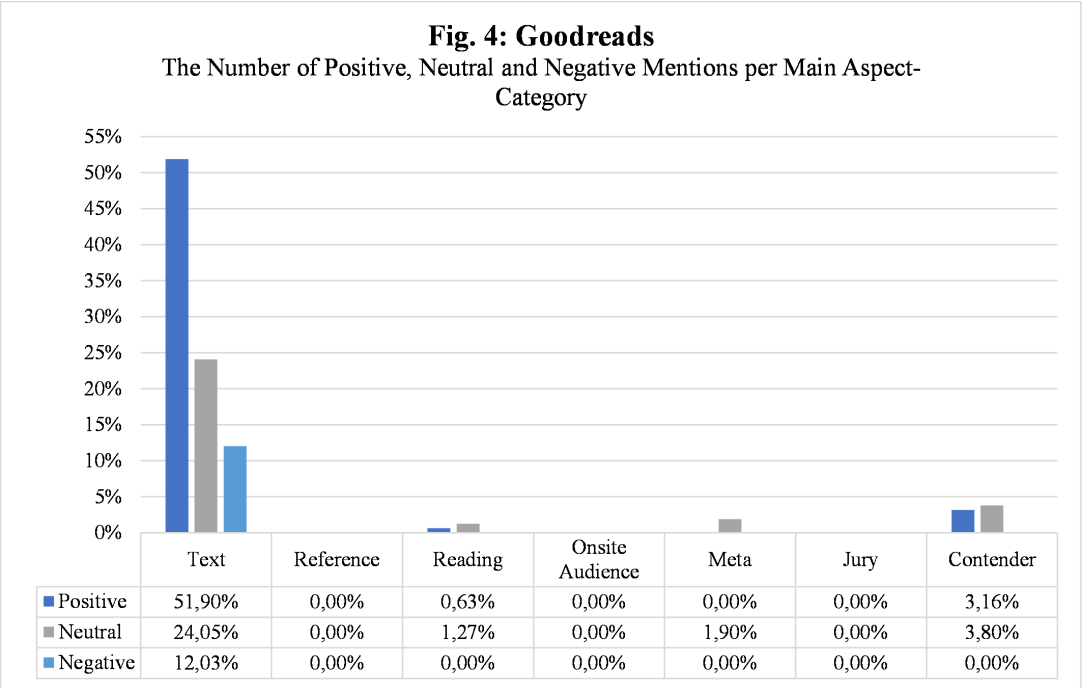


Figure 4: A graph containing the percentage of positive, neutral and negative mentions of each main aspect category in the Goodreads reviews.

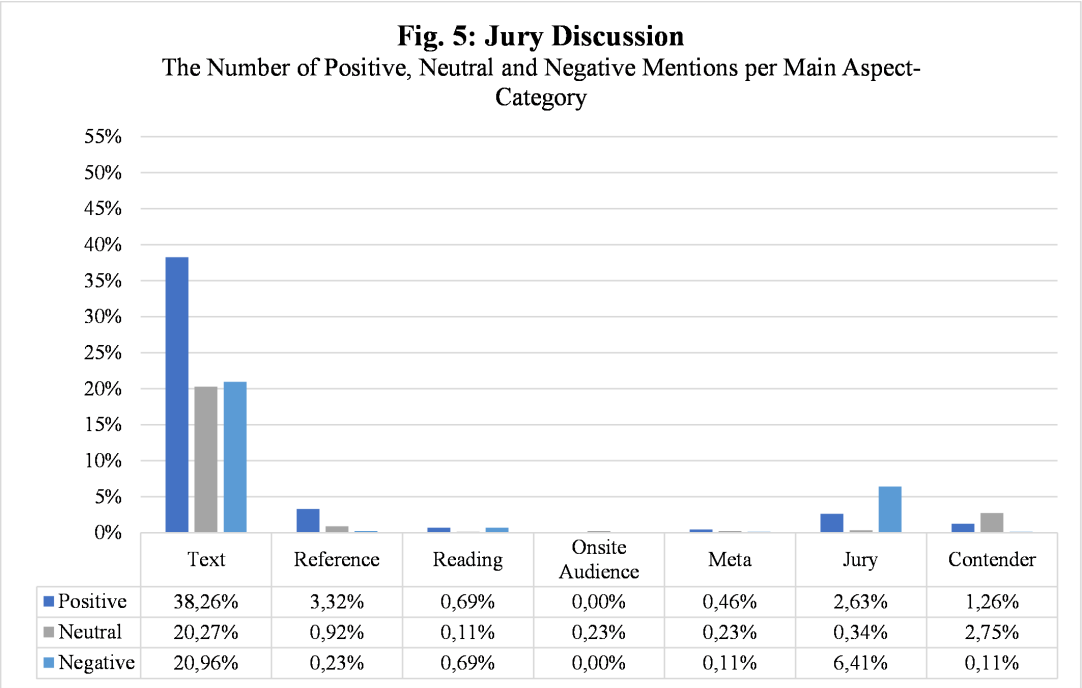
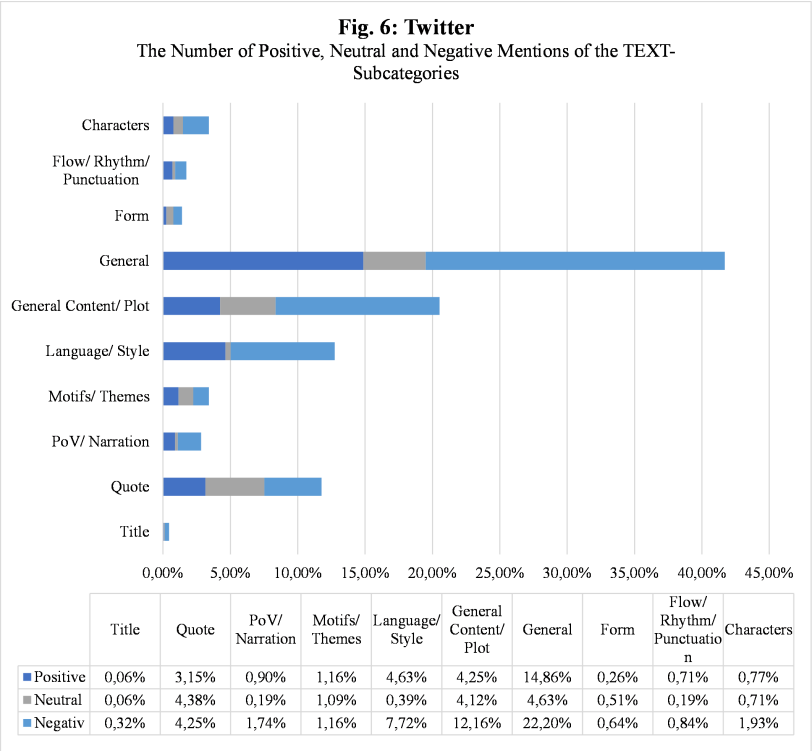


Figure 5: A graph containing the percentage of positive, neutral and negative mentions of each main aspect category in the jury discussions.

By addressing and examining the "Text"-subcategories, it is possible to examine which aspects of a literary text are being mentioned most often, which implies how

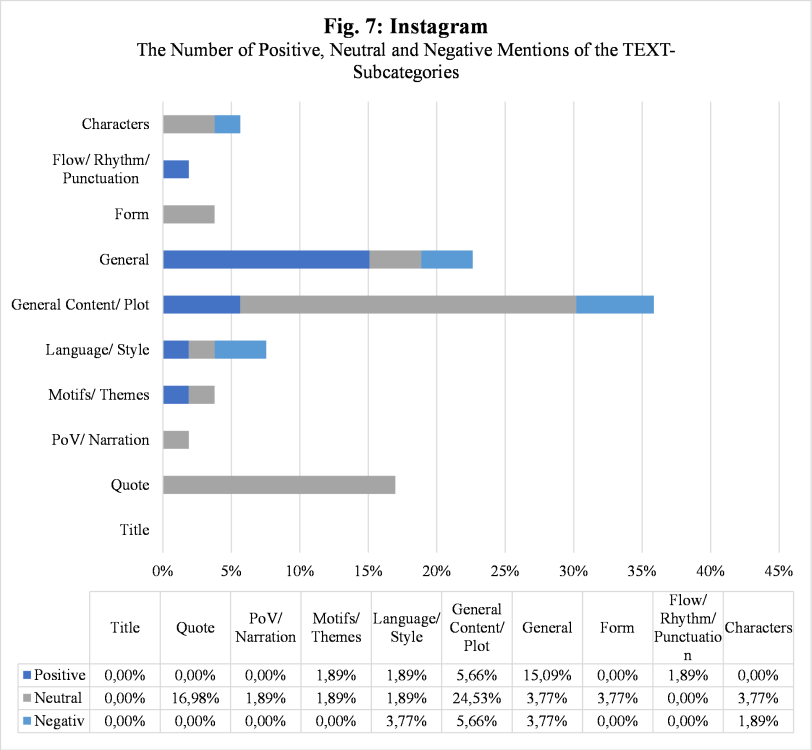
relevant this aspect is to a certain group, as well as whether it is mainly criticised or praised. The diagrams (Figures 6, 7, 8 and 9) illustrate how often a certain "Text"-subcategory is brought up in the Tweets, posts, reviews and jury discussions. In contrast to the previous graphs, in the context of the "Text"-category, the similarity between the Twitter discourse and the jury discussion is slightly more pronounced when compared to the other corpora. And in fact, contrary to common prejudice against fan communities and layperson criticism, the Twitter discourse is generally more negative (52,96%) than the actual jury discussion (26,37%), containing twice as many negative mentions in terms of percentage. Both of these discourses are comparatively more negative than the others (Instagram: 15,09%; Goodreads: 13,67%). Furthermore, the aspect distributions in the Tweets and jury discussion correspond more closely to one another. In both cases, they primarily discuss the text in general (Twitter: 41,70%; jury: 37,03%), secondly the content or plot (Twitter: 20,53%; jury: 24,78%) and thirdly the language use and style (Twitter: 12,74%; jury: 12,39%). The other aspect subcategories are not mentioned as often. Nevertheless, despite the similarities, there are some differences as well. The jury pays somewhat more attention to the characters (Twitter: 3,41%; jury: 5,76%), the motives and themes (Twitter: 3,41%; jury: 5,76%), the form of the text (Twitter: 1,42%; jury: 4,90%) as well as the narration (Twitter: 2,83%; jury: 6,77%), whilst the Twitter-users focus more on the quotations (Twitter: 11,78%; jury: 1,59%). The literary criteria expressed in the Tweets do not appear to be more superficial than those of the professional jury.

We can conclude that the literary evaluation criteria of the Twitter-users and the professional jury are in fact relatively similar, apart from some minor differences, and that these lay critics are even more critical than the professional jury, thus contradicting Bogaert's assumption that "das Potenzial eines Textes [...] nicht beim Bewerten in Betracht gezogen" is (Bogaert, 2017, 48, translation: "the potential of a text (...) not taken into account when evaluating it"). The subcategory aspects in Instagram posts are generally more often mentioned in a neutral context (58,49%) and are frequently informative statements. Only the text in general (15,09% positive out of 22,64%, equalling 66,65% of this total percentage) and the text's flow and rhythm are (predominantly) being praised, and the style and language (3,77% negative out of 7,55%, equalling 49,93% of this total percentage) are criticised more often. In comparison to the other corpora, they mention the plot and content of the text more frequently (35,85%). When examining the Goodreads reviews, the percentage of positive aspect mentions stands out: this is the most positive corpus regarding the "Text"-category (58,99%). In this corpus the text in general (26,62%) and its content (24,46%) are discussed almost equally often. Striking, however, is the fact that the characters are mentioned so frequently (14,39%), as they receive far less attention in the other three corpora (Twitter: 3,41%; Instagram: 5,66%; jury: 5,76%). The diagrams of the Instagram posts and Goodreads reviews illustrate a somewhat different hierarchy of aspect importance than illustrated by those of the Tweets and jury discussions: the text in general loses significance in comparison to its content. From these data can be concluded that many aspects of the competing texts are being discussed in all of the corpora, though each corpus somewhat has its own focus, once again (De Greve and Martens, in press, 2022) disproving Wegmann's thesis that "Auseinandersetzungen mit ästhetischen Formprinzipien, mit der Poetik von literarischen Texten, ihrer Stilistik, ihren rhetorischen Mitteln" in the Web 2.0 "[t]endenziell eher unterrepräsentiert sind" (Wegmann, 2012, 287, translation: "Discussions of aesthetic principles of form, of the poetics of literary texts, their stylistics, their rhetorical devices (...) (t)end to be underrepresented").



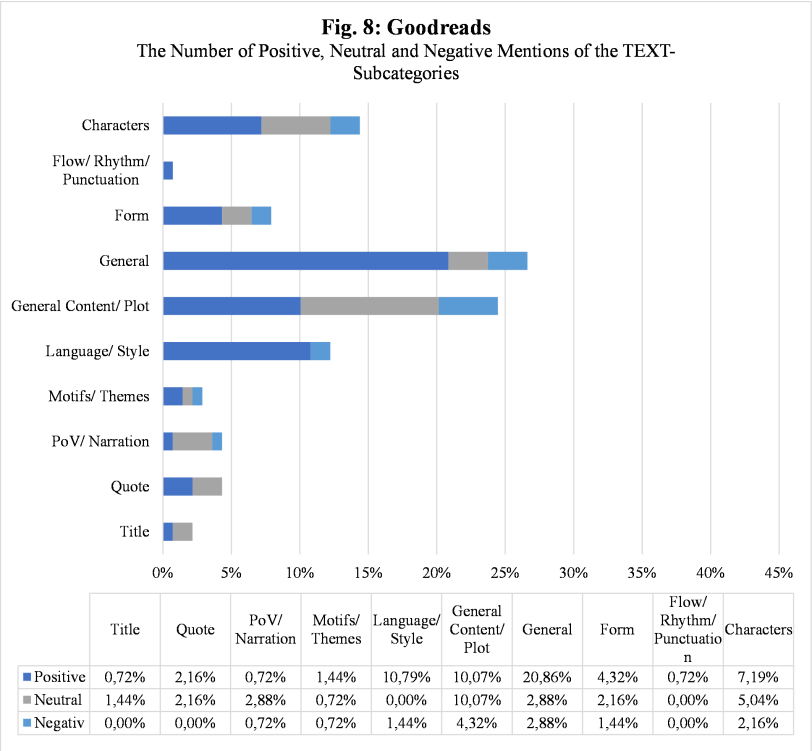
	Title	Quote	PoV/ Narration	Motifs/ Themes	Language/ Style	General Content/ Plot	General	Form	Flow/ Rhythm/ Punctuation	Characters
Positive	0,06%	3,15%	0,90%	1,16%	4,63%	4,25%	14,86%	0,26%	0,71%	0,77%
Neutral	0,06%	4,38%	0,19%	1,09%	0,39%	4,12%	4,63%	0,51%	0,19%	0,71%
Negativ	0,32%	4,25%	1,74%	1,16%	7,72%	12,16%	22,20%	0,64%	0,84%	1,93%

Figure 6: A graph containing the percentage of positive, neutral and negative mentions of the "Text"-subcategories in the Twitter discourse.



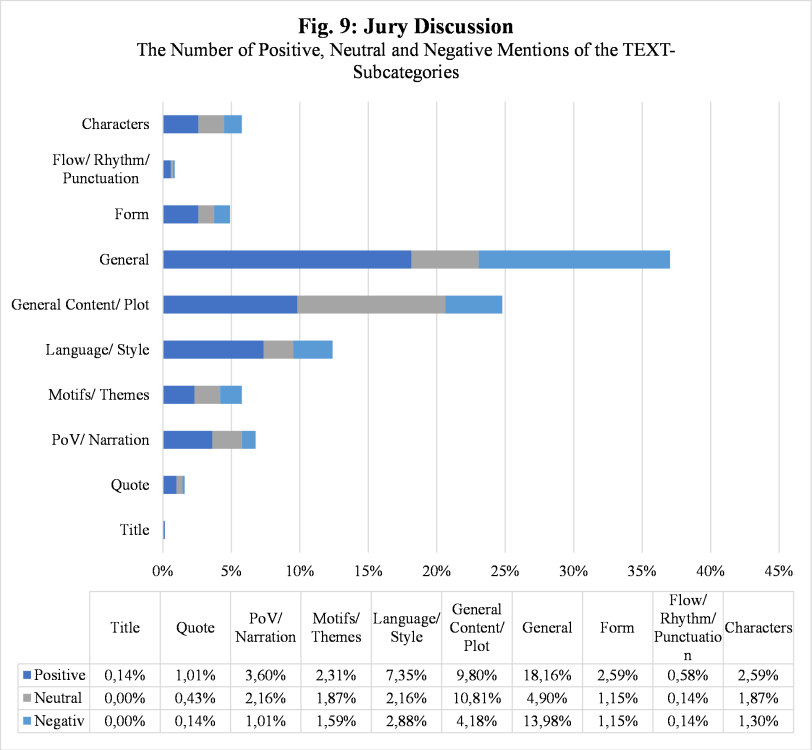
	Title	Quote	PoV/ Narration	Motifs/ Themes	Language/ Style	General Content/ Plot	General	Form	Flow/ Rhythm/ Punctuation	Characters
Positive	0,00%	0,00%	0,00%	1,89%	1,89%	5,66%	15,09%	0,00%	1,89%	0,00%
Neutral	0,00%	16,98%	1,89%	1,89%	1,89%	24,53%	3,77%	3,77%	0,00%	3,77%
Negativ	0,00%	0,00%	0,00%	0,00%	3,77%	5,66%	3,77%	0,00%	0,00%	1,89%

Figure 7: A graph containing the percentage of positive, neutral and negative mentions of the "Text"-subcategories in the Instagram discourse.



	Title	Quote	PoV/ Narration	Motifs/ Themes	Language/ Style	General Content/ Plot	General	Form	Flow/ Rhythm/ Punctuation	Characters
Positive	0,72%	2,16%	0,72%	1,44%	10,79%	10,07%	20,86%	4,32%	0,72%	7,19%
Neutral	1,44%	2,16%	2,88%	0,72%	0,00%	10,07%	2,88%	2,16%	0,00%	5,04%
Negativ	0,00%	0,00%	0,72%	0,72%	1,44%	4,32%	2,88%	1,44%	0,00%	2,16%

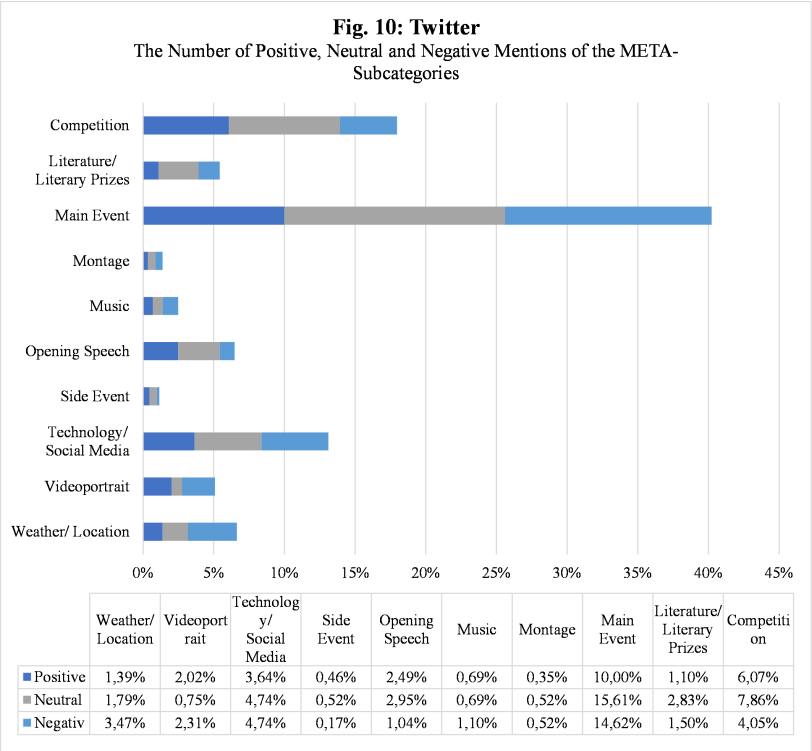
Figure 8: A graph containing the percentage of positive, neutral and negative mentions of the "Text"-subcategories in the Goodreads reviews.



	Title	Quote	PoV/ Narration	Motifs/ Themes	Language/ Style	General Content/ Plot	General	Form	Flow/ Rhythm/ Punctuation	Characters
Positive	0,14%	1,01%	3,60%	2,31%	7,35%	9,80%	18,16%	2,59%	0,58%	2,59%
Neutral	0,00%	0,43%	2,16%	1,87%	2,16%	10,81%	4,90%	1,15%	0,14%	1,87%
Negativ	0,00%	0,14%	1,01%	1,59%	2,88%	4,18%	13,98%	1,15%	0,14%	1,30%

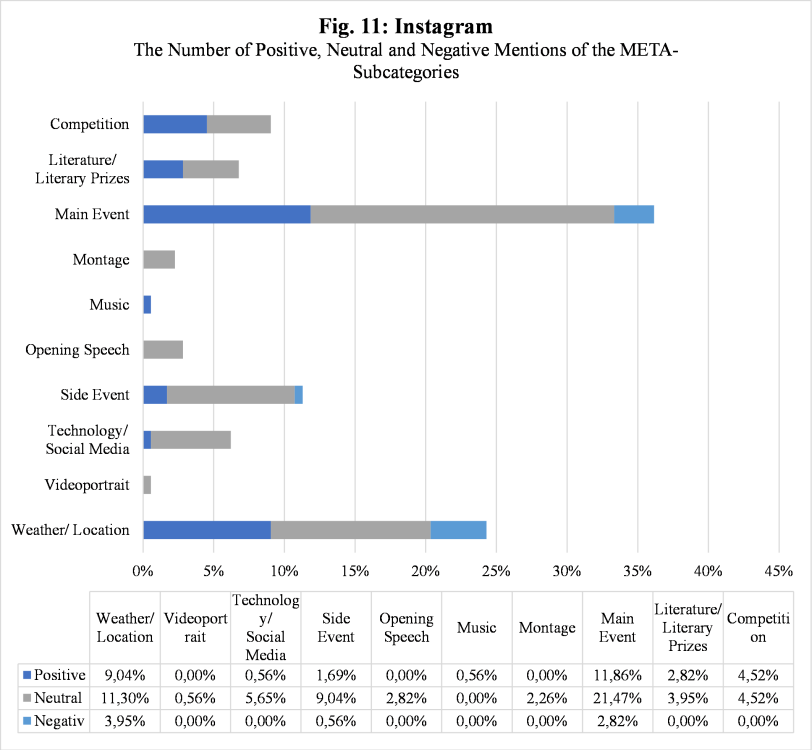
Figure 9: A graph containing the percentage of positive, neutral and negative mentions of the "Text"-subcategories in the jury discussions.

Moving onwards to the "Meta"-category (Fiure. 10, 11, 12 and 13), the differences between the corpora are more pronounced. Contrary to the previous results, in this context the resemblance in focus is greater between the Twitter and Instagram corpus and between the Goodreads reviews and the jury discussions, respectively. The former mainly focus on the TDDL and the Ingeborg-Bachmann-Preis itself (Twitter: 40,23%; Instagram: 36,16%), the "main event", and discuss every other subcategory as well, though in varying degrees – and that is where their differences lie. Once again, the Twitter discourse contains the largest percentage of negative mentions (33,53%), although the majority are neutral (38,27%), and the Instagram posts are mostly neutral (61,58%). The competition (Twitter: 17,98%; Instagram: 9,04%), concerning the voting, long- and shortlist as well as the award ceremony -mostly in a neutral (Twitter: 7,86%; Instagram: 4,52%) or positive (Twitter: 6,07%; Instagram: 4,52%) context-, the opening speech (Twitter: 6,47%; Instagram: 2,82%) as well as the technology and social media (Twitter: 13,12%; Instagram: 6,21%) are mentioned almost twice as frequently in the Tweets, whereas the Instagram posts more often address the side events (Twitter: 1,16%; Instagram: 11,30%) and, especially so, the weather and location (Twitter: 6,65%; Instagram: 24,29%). The main focus on the event itself for the Twitter and Instagram corpora in addition to the greater focus on location on Instagram than on Twitter confirm the preliminary results from our preceding digital corpus analysis, as "[f]or such a location-oriented visual social media platform [=Instagram], the frequent occurrence of place names is to be expected" (De Greve and Martens, 2021b, 15). The Goodreads reviews and jury discussions on the other hand appear to concentrate mostly on the event itself (Goodreads: 33,33%; jury: 28,57%), similar to the Twitter and Instagram corpus, the competition (Goodreads: 33,33%; jury: 28,57%), the technology and social media (on Goodreads, 33,33%) and on the video portraits and the weather and location (jury discussion, two times 14,29%). However, contrary to the Tweets and Instagram posts, they do not address any of the other subcategories. The Goodreads corpus solely contains neutral mentions, whereas the sentiment varies depending on the subcategory in the jury discussions, with a total of 57,14% positive, 28,57% neutral and 14,29% negative mentions. However, looking at the percentages presented here, it is important to keep in mind that the total percentage of mentions of the "Meta"-category is in fact very low for both of these corpora (cf. Figures 4 and 5) and consequently are relatively negligible, in accordance with the conclusion that there were "no explicit references to the Ingeborg-Bachmann-Preis or the TDDL" in the Goodreads reviews (De Greve and Martens, 2021b, 19).



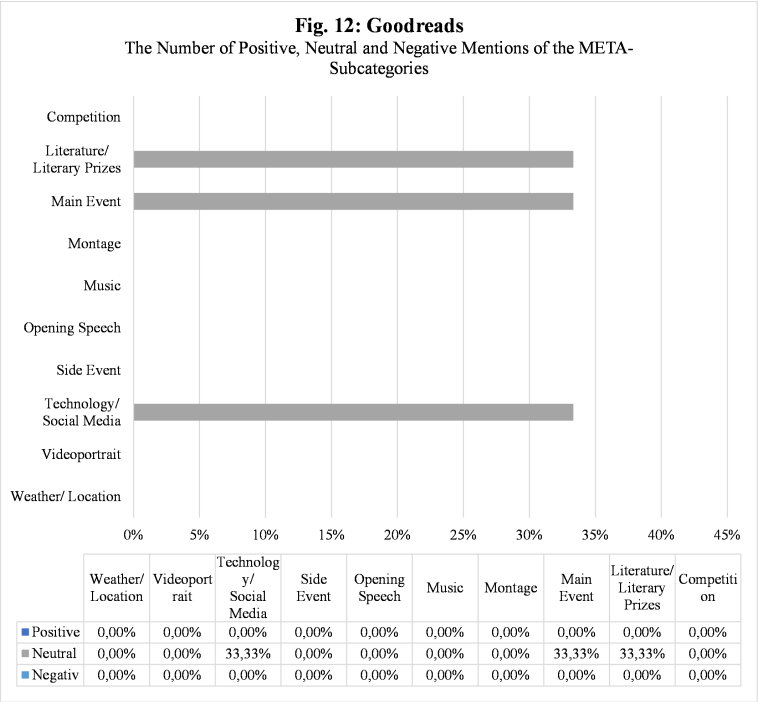
	Weather/ Location	Videoportrait	Technology/ Social Media	Side Event	Opening Speech	Music	Montage	Main Event	Literature/ Literary Prizes	Competition
Positive	1,39%	2,02%	3,64%	0,46%	2,49%	0,69%	0,35%	10,00%	1,10%	6,07%
Neutral	1,79%	0,75%	4,74%	0,52%	2,95%	0,69%	0,52%	15,61%	2,83%	7,86%
Negativ	3,47%	2,31%	4,74%	0,17%	1,04%	1,10%	0,52%	14,62%	1,50%	4,05%

Figure 10: A graph containing the percentage of positive, neutral and negative mentions of the "Meta"-subcategories in the Twitter discourse.



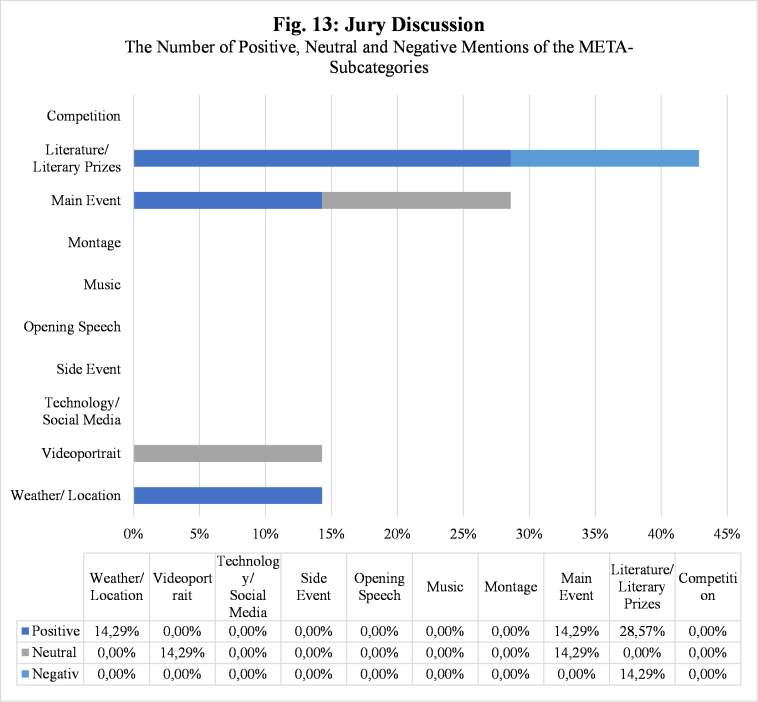
	Weather/ Location	Videoportrait	Technology/ Social Media	Side Event	Opening Speech	Music	Montage	Main Event	Literature/ Literary Prizes	Competition
Positive	9,04%	0,00%	0,56%	1,69%	0,00%	0,56%	0,00%	11,86%	2,82%	4,52%
Neutral	11,30%	0,56%	5,65%	9,04%	2,82%	0,00%	2,26%	21,47%	3,95%	4,52%
Negativ	3,95%	0,00%	0,00%	0,56%	0,00%	0,00%	0,00%	2,82%	0,00%	0,00%

Figure 11: A graph containing the percentage of positive, neutral and negative mentions of the "Meta"-subcategories in the Instagram discourse.



	Weather/ Location	Videoportrait	Technology/ Social Media	Side Event	Opening Speech	Music	Montage	Main Event	Literature/ Literary Prizes	Competition
Positive	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
Neutral	0,00%	0,00%	33,33%	0,00%	0,00%	0,00%	0,00%	33,33%	33,33%	0,00%
Negativ	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

Figure 12: A graph containing the percentage of positive, neutral and negative mentions of the "Meta"-subcategories in the Goodreads reviews.



	Weather/ Location	Videoportrait	Technology/ Social Media	Side Event	Opening Speech	Music	Montage	Main Event	Literature/ Literary Prizes	Competition
Positive	14,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	14,29%	28,57%	0,00%
Neutral	0,00%	14,29%	0,00%	0,00%	0,00%	0,00%	0,00%	14,29%	0,00%	0,00%
Negativ	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	14,29%	0,00%

Figure 13: A graph containing the percentage of positive, neutral and negative mentions of the "Meta"-subcategories in the jury discussions.

The next set of graphs (Figures 14, 15, 16 and 17) shows the percentage of positive, neutral and negative mentions of the jury in the four corpora. Both the Instagram posts

and the jury discussions exclusively discuss the jury discussion and valuation (Instagram: 71,43%; jury: 96,34%) and the jury in general (Instagram: 28,57%; jury: 3,66%). Although the discussion of the professional Bachmann-Preis jury focuses mainly on the nominated texts, they occasionally also discuss topics related to themselves. This usually consists of criticising the arguments or opinions of the other jury members or agreeing with them (total of 68,29% negative and 28,05% positive mentions). In comparison, the percentage of neutral mentions (62,86%) is much higher in the Instagram corpus and they address the jury in general more frequently: 28,57% compared to 3,66%. A greater variety of jury-related aspects is being discussed in the Tweets, although their main focus is on the jury discussion and valuation (71,83%) as well. This subcategory is most mentioned in a negative context (37,56% negative out of 71,83%, equalling 52,29% of this total percentage), which indicates that the Twitter-users generally disagree with the jury's evaluation, even though they were not as condemning of the jury's decision regarding the competition (Figure 10). In fact, most of the mentions for the "Jury"-category are negative (50,23%). It must be taken into consideration, however, that the Twitter-users use the jury discussions as a stepping stone to take part in the online discussion of the competing texts, by evaluating them indirectly (Bogaert, 2017, 54-56). Besides this, the Tweets also simply mention (10,19%) and quote (10,33%) the jury members. In addition, the Twitter discourse is also interested in some non-text evaluation-related aspects of the professional jury, such as their voice or language use (2,35%),[22] their behaviour (2,88%) and their age, appearance and clothing (2,41%). This focus on secondary aspects, such as age, appearance, attire and gender also extends to the discussion surrounding the contenders (cf. Figure 18).[23] The reason for this may be twofold: on the one hand, the online audience watching the TDDL-livestream is able to see or is confronted with these secondary aspects and may simply comment on them, just like they comment on the montage of the livestream or the interior decor of the studio. On the other hand, however, the online community also uses Twitter and Instagram (e.g. regarding the gender of the authors) for social activism by means of striking symbols. Comparable in this context is, for example, the role of Hanna Engelmeier's T-shirt at the Deutscher Buchpreis.[24] This attention to detail might strike one as superfluous or as an aberration. But the dress-code of the jurors signals their habitus, just as the major critics of past eras like Reich-Ranicki and Fritz J. Raddatz did with their more formal attire. While everything can be said to be political, the more outspoken nature of politically engaged discourse also coincides

---

22 In this specific Twitter corpus this mainly concerns the dialect of the jury members; e.g. "Sobald ich Ankowitsch höre, habe ich imaginäre Gummibänder im Mund und kann österreichisch. #tddl" (translation: "As soon as I hear Ankowitsch [=TDDL moderator], I have imaginary rubber bands in my mouth and I speak Austrian. #tddl").

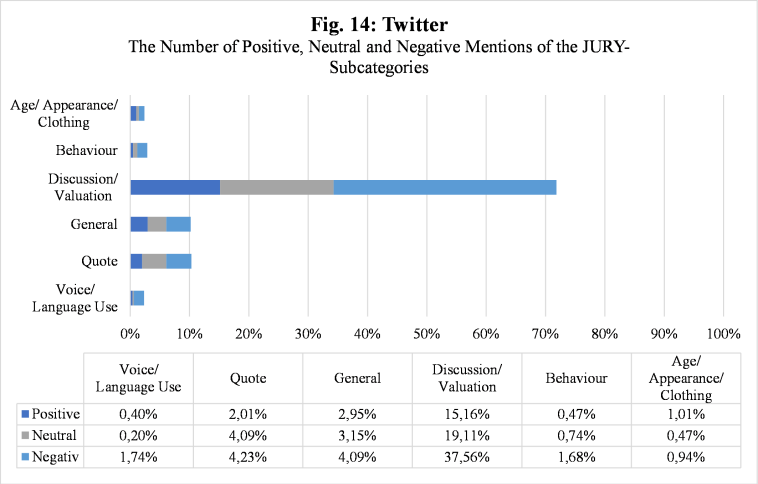
@slowtiger. 'Sobald ich Ankowitsch höre, habe ich imaginäre Gummibänder im Mund und kann österreichisch. #tddl'. *Twitter*, 27 Jun. 2019, `https://twitter.com/slowtiger/status/1144154475931799553`, last accessed 5 Apr. 2022.

23 E.g. following a Tweet in which a Twitter-user comments on the fact that the German-language media called the first TDDL-day a "women's day" due to the fact that only female authors read that day: "Habe ich in den letzten zwanzig Jahren das Wort 'Männertag' gelesen? Ich glaube, nein. Wie wär's mal mit Nachdenken. [...] @ORF @3sat @tddlit . #tddl pic.twitter.com/j6utPeSRck" (translation: "Have I ever read the word 'Men's Day' in the last twenty years? I think not. So how about some reflection. [...] @ORF @3sat @tddlit . #tddl pic.twitter.com/j6utPeSRck").

@Dschungoerl. 'Habe ich in den letzten zwanzig Jahren das Wort „Männertag" gelesen? Ich glaube, nein. Wie wär's mal mit Nachdenken. [...] @ORF @3sat @tddlit. #tddl'. *Twitter*, 29 Jun. 2019, `https://twitter.com/Dschungoerl/status/1145002713152937984`, last accessed 5 Apr. 2022.

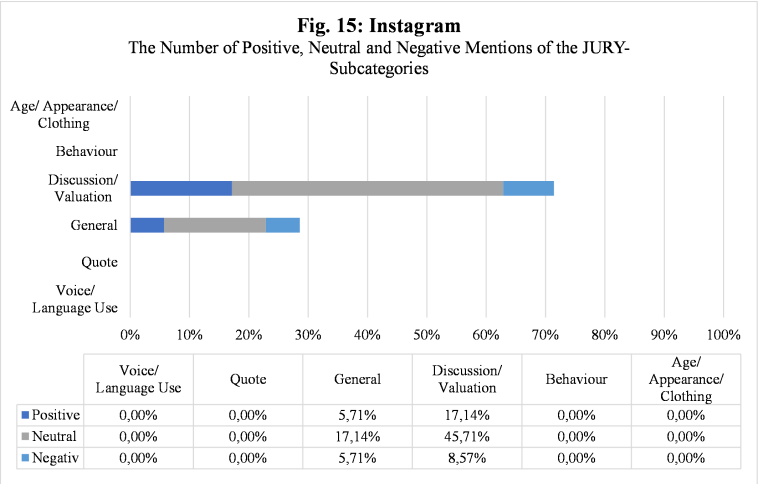
24 Cf. "Ein T-Shirt sorgt für eine Debatte: Plappern mit Jürgen Habermas" (Knipphals, 2020) and "Deutscher Buchpreis: Lesen Sie zuerst das T-Shirt!" (Küveler, 2020)

with online movements such as #frauenzählen (translation: "#countingwomen").[25] Critics on social media are more preoccupied with the visibility of women and gender equality in the literary field, e.g. "Federer und Heitzler auf der Shortlist und Birkhan fehlt? Unverständlich. Das muss ein Fall von Quotenmännern sein. #tddl" (translation: "Federer and Heitzler on the shortlist and Birkhan missing? Incomprehensible. This must be a case of token men. #tddl").[26] In comparison, the jury is not mentioned at all in the corpus of Goodreads reviews (De Greve and Martens, 2021b), which focus mainly on the text itself.



	Voice/ Language Use	Quote	General	Discussion/ Valuation	Behaviour	Age/ Appearance/ Clothing
Positive	0,40%	2,01%	2,95%	15,16%	0,47%	1,01%
Neutral	0,20%	4,09%	3,15%	19,11%	0,74%	0,47%
Negativ	1,74%	4,23%	4,09%	37,56%	1,68%	0,94%

Figure 14: A graph containing the percentage of positive, neutral and negative mentions of the "Jury"-subcategories in the Twitter discourse.



	Voice/ Language Use	Quote	General	Discussion/ Valuation	Behaviour	Age/ Appearance/ Clothing
Positive	0,00%	0,00%	5,71%	17,14%	0,00%	0,00%
Neutral	0,00%	0,00%	17,14%	45,71%	0,00%	0,00%
Negativ	0,00%	0,00%	5,71%	8,57%	0,00%	0,00%

Figure 15: A graph containing the percentage of positive, neutral and negative mentions of the "Jury"-subcategories in the Instagram discourse.

25 http://www.frauenzählen.de/, last accessed 4 Apr. 2022.

26 @gedankentraeger. "Federer und Heitzler auf der Shortlist und Birkhan fehlt? Unverständlich. Das muss ein Fall von Quotenmännern sein. #tddl". *Twitter*, 30 Jun. 2019, See:https://twitter.com/gedankentraeger/status/1145258994136690688, last accessed 4 Apr. 2022.

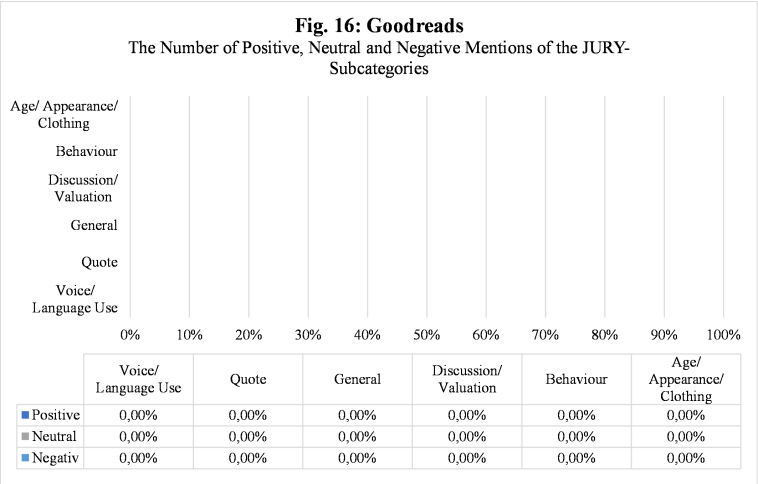


Figure 16: A graph containing the percentage of positive, neutral and negative mentions of the "Jury"-subcategories in the Goodreads reviews.

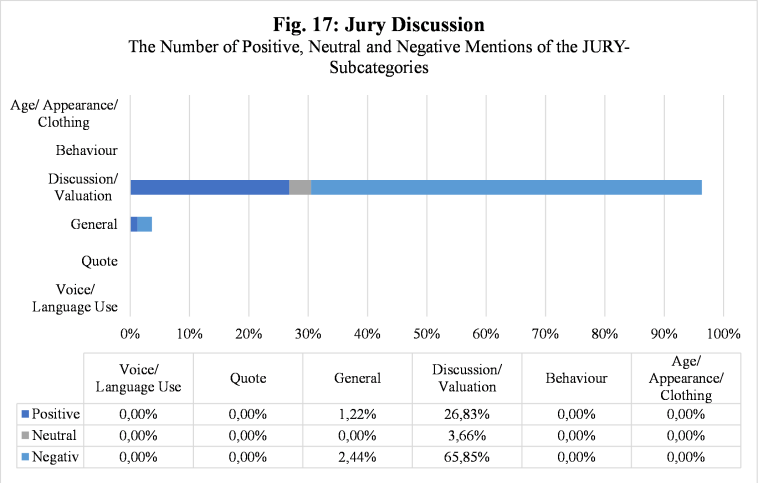


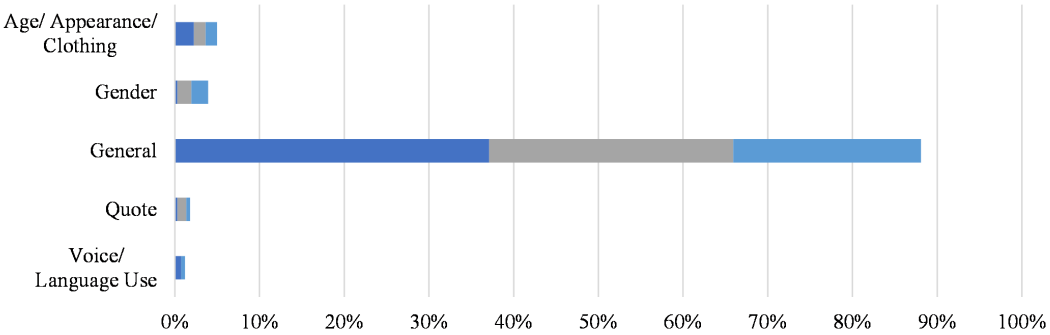
Figure 17: A graph containing the percentage of positive, neutral and negative mentions of the "Jury"-subcategories in the jury discussions.

The final main aspect category we discuss in this article concerns the nominated authors (Figures 18, 19, 20 and 21). Once more, the Twitter discourse remains the corpus with the greatest percentage of negative mentions by far (Twitter: 26,40%; versus Instagram: 3,23%; Goodreads: 0%; and jury: 2,78%), even though the percentage of positive (40,72%) and neutral (32,88%) mentions of this main aspect category prevail. The Twitter corpus is comparatively more varied and discusses a wider range of topics, namely the competitors' voice and language use (1,21%), quotes (1,81%), and – as addressed in the previous paragraph – their gender (3,92%), as well as their age, appearance and clothes (4,98%). Nevertheless, the main focus for all four corpora are the contenders in general (Twitter: 88,08%; Instagram: 93,55%; Goodreads: 100%; jury: 88,89%). In the corpora of Instagram posts, Goodreads reviews and jury discussions, the mentions are mostly neutral, informative statements (Instagram: 58,06%; Goodreads: 54,55%; jury: 66,67%) or positive references (Instagram: 38,71%; Goodreads: 45,45%; jury: 30,56%), with little to no negative sentiment. In the Goodreads reviews, no other aspects related

to the nominated authors are mentioned, but the Instagram posts periodically discuss the gender (cf. previous paragraph) as well (6,45%), and the jury sometimes quotes the contender (2,78%) or addresses his or her voice and language use (8,33%), which can be connected to the author readings.

Fig. 18: Twitter

The Number of Positive, Neutral and Negative Mentions of the CONTENDER-Subcategories

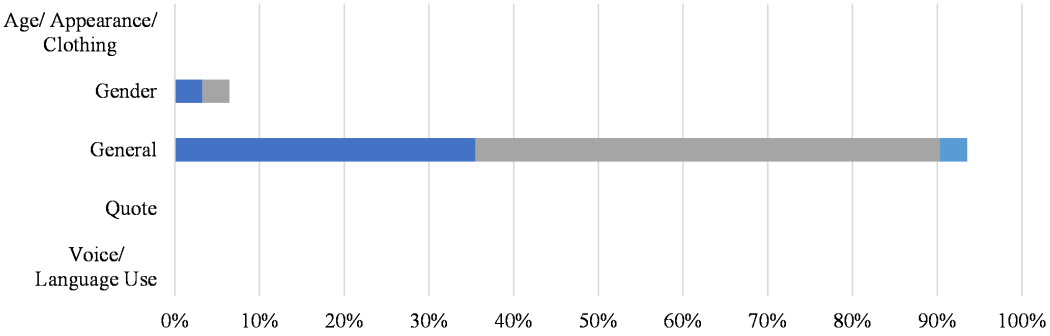


	Voice/ Language Use	Quote	General	Gender	Age/ Appearance/ Clothing
Positive	0,75%	0,30%	37,10%	0,30%	2,26%
Neutral	0,00%	1,06%	28,81%	1,66%	1,36%
Negativ	0,45%	0,45%	22,17%	1,96%	1,36%

Figure 18: A graph containing the percentage of positive, neutral and negative mentions of the "Contender"-subcategories in the Twitter discourse.

Fig. 19: Instagram

The Number of Positive, Neutral and Negative Mentions of the CONTENDER-Subcategories



	Voice/ Language Use	Quote	General	Gender	Age/ Appearance/ Clothing
Positive	0,00%	0,00%	35,48%	3,23%	0,00%
Neutral	0,00%	0,00%	54,84%	3,23%	0,00%
Negativ	0,00%	0,00%	3,23%	0,00%	0,00%

Figure 19: A graph containing the percentage of positive, neutral and negative mentions of the "Contender"-subcategories in the Instagram discourse.

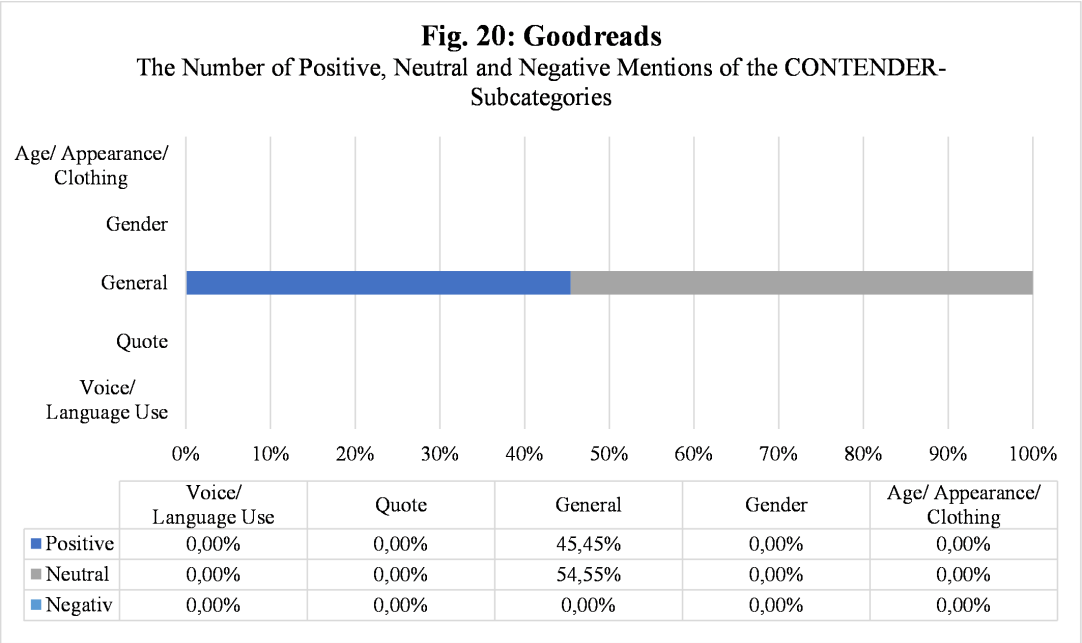


Figure 20: A graph containing the percentage of positive, neutral and negative mentions of the "Contender"-subcategories in the Goodreads reviews.

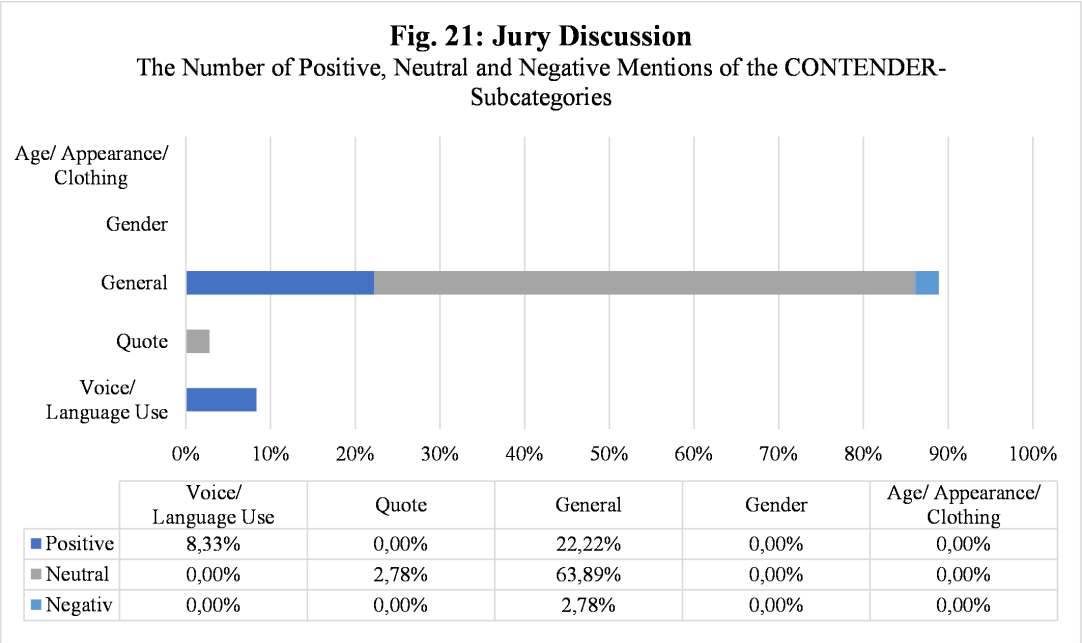


Figure 21: A graph containing the percentage of positive, neutral and negative mentions of the "Contender"-subcategories in the jury discussions.

4 Conclusion

This article has addressed the literary criticism surrounding the German Ingeborg-Bachmann-Preis in order to gain deeper insight into the content of the evaluative talk

about literature by professional and layperson critics 'in real life' as well as on various social media platforms by performing a fine-grained aspect-based sentiment analysis on a manually annotated corpus consisting of Tweets, Instagram posts, Goodreads reviews and descriptions of the jury discussions. The goal of this article was to identify which topics regarding the TDDL (e.g. nominated texts, jury, authors, the prize itself etc.) were being discussed in the different corpora and which sentiment was expressed about them. We wanted to analyse the possible differences between the Twitter, Instagram, Goodreads and jury corpora and between the evaluative literary criteria of professional and layperson critics, as well as to expand upon and compare the new findings to the results presented in our previous articles on the Bachmann-Preis (De Greve and Martens, 2021b, in press, 2022). In addition to this, we also briefly discussed the corpus composition and annotation method.

First, we compared the four corpora based on seven main aspect categories before zooming in on the subcategories of the four most frequently discussed main categories, namely the "Text"-, "Meta"-, "Jury"- and "Contender"-categories. The visualisations focusing on the main aspect categories illustrated the main overarching differences and similarities regarding discussed topics and sentiment between the discourses. We discovered that, on an overarching level, there is a greater similarity between the Twitter and Instagram discourses on the one hand and between the Goodreads reviews and the jury discussions on the other hand. Due to the lack of restrictions and expectations regarding the main focus of the discourse, the Tweets and Instagram posts are free to focus on more diverse topics, whereas both the Goodreads reviews and jury discussions focused predominantly on the nominated texts. The latter were, percentage-wise, comparatively more positive than the others. Overall, the Instagram corpus contained more neutral mentions and the Twitter-users were the most critical. The examination of the "Text"-subcategories enabled us to gain a deeper insight into the evaluative literary criteria at work in each discourse. We discovered that these were relatively similar, although the corpora each had their own focal points to which they paid more attention than the others. The differences were more substantial, however, for the other three subcategories, regarding the event itself, the jury and the competing authors, where the Tweets tended to discuss more diverse topics than the other corpora. The analysis of the manual annotation of the four corpora consequently enabled us to confirm that there are multiple distinctions between the literary discourse and the respective evaluative literary criteria on Twitter, Instagram, Goodreads and of the professional jury discourse, as well as to disprove the prejudice and hypothesis that layperson literary criticism does not concern itself with aesthetic principles, such as form, style etc. (Wegmann, 2012), or only makes an assessment based on the judgement of the professional jury and is generally less critical. This study has delved into the content of the literary discourse surrounding the Ingeborg-Bachmann-Preis and has provided an in-depth qualitative and quantitative analysis of the selected and manually annotated corpora. A larger quantitative analysis will be needed to investigate whether these findings also apply to the discourse of other editions of the TDDL. However, the data presented here can be used to train a semi-supervised learning system to automatically examine and explore the TDDL-discourse over a larger time period.

References

Daniel Allington. "'Power to the Reader' or 'Degradation of Literary Taste'? Professional Critics and Amazon Customers as Reviewers of 'The Inheritance of Loss"'. *Language and Literature*, 25(3):254–278, 2016.

Tamara Álvarez-López, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, and Patrice Bellot. A Proposal for Book Oriented Aspect Based Sentiment Analysis: Comparison over Domains. In *Natural Language Processing and Information Systems. NLDB 2018*, pages 3–14, Paris, France, 2018. URL https://hal.archives-ouvertes.fr/hal-01958697.

Anna Auguscik. *Prizing Debate: The Fourth Decade of the Booker Prize and the Contemporary Novel in the UK*. Transcript, 2017.

Xiana Bogaert. 'ICH WÜRDE AM LIEBSTEN MIT DER JURY DISKUTIEREN! #TDDL' - Der Ingeborg-Bachmann-Preis: Ein Vergleich zwischen der Jury- und Laienkritik auf Twitter. Master's thesis, Ghent University, 2017.

Pierre Bourdieu. *The Field of Cultural Production. Essays on Art and Literature*. Polity Press Blackwell Publishers, 1993.

Rebecca Braun. PRIZE GERMANS? CHANGING NOTIONS OF GERMANNESS AND THE ROLE OF THE AWARD-WINNING AUTHOR INTO THE TWENTY-FIRST CENTURY. *Oxford German Studies*, 43(1):37–54, 2014.

Christine Chenaux and Annina Beck. *Schweizer Literaturpreise: Preisträgerinnen und Preisträger 2015*. Bundesamt für Kultur, 2015.

Clayton Childress, Craig M. Rawlings, and Brian Moeran. Publishers, Authors, and Texts: The Process of Cultural Consecration in Prize Evaluation. *Poetics*, 60:48–61, 2017.

Lore De Greve and Gunther Martens. The audience (dis)agrees : studying the impact of award-winning books on lay literary value judgements using social media data. In Pieternella Broomans, Mathijs Sanders, and Jeanette den Toonder, editors, *Literary Prizes and Cultural Transfer*, volume 9 of *Studies on Cultural Transfer and Transmission*, pages 85–130. Barkhuis, 2021a. ISBN 9789493194380.

Lore De Greve and Gunther Martens. #Bookstagram and Beyond: The Presence and Depiction of the Bachmann Literary Prize on Social Media (2007-2017). *DIGITAL HUMANITIES BENELUX JOURNAL*, 3:81–102, 2021b. ISSN 2666-6952.

Lore De Greve and Gunther Martens. Wertung von Literatur 2.0 : eine digitale und literatursoziologische Analyse der Online-Twitter-Diskussion zu den Tagen der deutschsprachigen Literatur #tddl. In Oliver Ruf and Christoph H. Winter, editors, *Small critics: Transmediale Konzepte feuilletonistischer Schreibweisen der Gegenwart*, volume 3 of *Mikrographien / Mikrokosmen. Kultur-, literatur- und medienwissenschaftliche Studien; Band 3*. Königshausen & Neumann, in press, 2022.

Lore De Greve, Pranaydeep Singh, Cynthia Van Hee, Els Lefever, and Gunther Martens. Aspect-based sentiment analysis for German : analyzing 'talk of literature' surrounding literary prizes on social media. *COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS JOURNAL*, 11:85–104, 2021. ISSN 2211-4009.

Erwin Dekker and Marielle de Jong. What Do Book Awards Signal? An Analysis of Book Awards in Three Countries. *Empirical Studies of the Art*, 36(1):3–21, 2018.

Beth Driscoll. Twitter, literary prizes and the circulation of capital. In Emmett Stinson, editor, *By the book? Contemporary publishing in Australia*, pages 103–119. Monash University Press, 2013.

Beth Driscoll. *The New Literary Middlebrow: Tastemakers and Reading in the Twenty-first Century*. Palgrave Macmillan, 2014.

Sylvie Ducas. *La littérature à quel(s) prix? Histoire des prix littéraires*. La Découverte, 2013.

Wolfgang Emmerich. Konsekration Und Kanonisierung. Tendenzen Der Vergabe von Literaturpreisen in Der Bundesrepublik 1950 Bis 1989. *Amsterdamer Beiträge Zur Neueren Germanistik*, 82(1):113–132, 2012.

James F. English. *The Economy of Prestige: Prizes, Awards, and the Circulation of Cultural Value*. Harvard University Press, 2009.

Frans Heymans. *Het Goud van de Vlaamse Letteren: 170 Jaar Prijzen voor de Nederlandse Literatuur in België (1830-2000)*. Snoeck-Ducaju, 2001.

Ingo Irsigler and Gerrit Lembke. *Spiel, Satz und Sieg: 10 Jahre Deutscher Buchpreis*. Berlin university press, 2014.

Maarit Jaakkola. From Re-Viewers to Me-Viewers: The #Bookstagram Review Sphere on Instagram and the Uses of the Perceived Platform and Genre Affordances. *Interactions: Studies in Communication & Culture*, 10(1/2):91–110, 2019.

Holger Kellermann and Gabriele Mehling. Laienrezensionen auf amazon.de im Spannungsfeld zwischen Alltagskommunikation und professioneller Literaturkritik. In Andrea Bartl and Markus Behmer, editors, *Die Rezension: Aktuelle Tendenzen der Literaturkritik*, pages 173–202. Würzburg: Königshausen und Neumann, 2017.

Holger Kellermann, Gabriele Mehling, and Martin Rehfeldt. Wie bewerten Laienrezensenten? Ausgewählte Ergebnisse einer inhaltsanalytischen Studie. In Stefan Neuhaus and Uta Schaffers, editors, *Was wir lesen sollen: Kanon und literarische Wertung am Beginn des 21. Jahrhunderts*, pages 229–238. Würzburg: Königshausen und Neumann, 2016.

Colleen Kennedy-Karpat and Eric Sandberg. *Adaptation, Awards Culture, and the Value of Prestige*. Palgrave Macmillan, 2017.

Dirk Knipphals. Ein T-Shirt sorgt für eine Debatte: Plappern mit Jürgen Habermas. *Die Tageszeitung*, 2020. URL https://taz.de/!5717509/.

Vera Kostial. Der Deutsche Buchpreis auf der Social Reading-Plattform LovelyBooks: Wahrnehmung, Diskussion und Wertung. In Christoph Jürgensen and Antonius Weixler, editors, *Literaturpreise: Geschichte und Kontexte*, pages 183–197. J.B. Metzler, 2021.

Kayvan Kousha and Mike Thelwall. Can Amazon.com Reviews Help to Assess the Wider Impacts of Books? *Journal of the Association for Information Science and Technology*, 67(3):566–581, 2016.

Kayvan Kousha, Mike Thelwall, and Mahshid Abdoli. Goodreads Reviews to Assess the Wider Impacts of Books. *Journal of the Association for Information Science and Technology*, 68(8):2004–2016, 2017.

Jan Küveler. Deutscher Buchpreis: Lesen Sie zuerst das T-Shirt! *Die Welt*, 2020. URL https://www.welt.de/kultur/literarischewelt/plus217744960/Deutscher-Buchpreis-Lesen-Sie-zuerst-das-T-Shirt.html.

Angela Leinen. *Wie man den Bachmannpreis gewinnt. Gebrauchsanweisung zum Lesen und Schreiben*. Heyne, 2010.

Jeffrey Meyers. The Literary Politics of the Nobel Prize. *The Antioch Review*, 65(2): 2014–223, 2007.

Doris Moser. *Der Ingeborg-Bachmann-Preis. Börse, Show, Event*. Böhlau Verlag, 2004.

Jessica Pressman. *Bookishness. Loving Books in a Digital Age*. Columbia University Press, 2020.

Kathrin Rahmann. *Von der Wirkung zur Wertung. Formal-ästhetische Werte in den Diskussionen des Ingeborg-Bachmann-Wettbewerbs 1999-2009*. PhD thesis, Georg-August-Universität Göttingen, 2017.

Kristin Rebien. Literary Awards and the Practice of Aesthetic Judgment: Lessons of the Ingeborg Bachmann Preis. *Journal of Austrian Studies*, 45(3/4):113–134, 2012.

Karin Röhricht. *Wettlesen um den Ingeborg-Bachmann-Preis. Korpusanalyse der Anthologie Klagenfurter Texte (1977-2011)*. Studienverlag, 2016.

Gisèle Sapiro. The Metamorphosis of Modes of Consecration in the Literary Field: Academies, Literary Prizes, Festivals. *Poetics*, 59:5–19, 2016.

Ute Schneider. Bücher zeigen und Leseatmospären inszenieren. Vom Habitus enthusiastischer Leserinnen und Leser. In Steffen Martus and Carlos Spoerhase, editors, *Gelesene Literatur: Populäre Lektüre im Zeichen des Medienwandels*, pages 113–123. München: Edition Text+Kritik, 2018.

Ann Steiner. Private Criticism in the Public Sphere: Personal Writing on Literature in Readers' Reviews on Amazon'. *Participations*, 5(2), 2008.

Mike Thelwall and Kousha Kavyan. Goodreads: A Social Network Site for Book Readers. *Journal of the Association for Information Science and Technology*, 68(4):972–983, 2017.

Erika Thomalla. Bücheremphase: Populäre Literaturkritik und Social Reading im Netz. In Steffen Martus and Carlos Spoerhase, editors, *Gelesene Literatur: Populäre Lektüre im Zeichen des Medienwandels*, pages 124–136. München: Edition Text+Kritik, 2018.

Richard Todd. *Consuming Fictions: The Booker Prize and Fiction In Britain Today*. Bloomsbury, 1996.

Judith S. Ulmer. *Geschichte des Georg-Büchner-Preises: Soziologie eines Rituals*. De Gruyter, 2006.

Melanie Walsh and Maria Antoniak. The Goodreads "Classics": A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism. *Journal of Cultural Analytics*, 4:243–287, 2021.

Kai Wang, Xiaojuan Liu, and Yutong Han. Exploring Goodreads reviews for book impact assessment. *Journal of Informetrics*, 13(3):874–886, 2019.

Millicent Weber and Beth Driscoll. Playful Twitter Accounts and the Socialisation of Literary Institutions. *First Monday*, 24(3):1–23, 2019.

Thomas Wegmann. Warentest und Selbstmanagement. Literaturkritik im Web 2.0 als Teil nachbürgerlicher Wissens- und Beurteilungskulturen. In Matthias Beilein, Claudia Stockinger, and Simone Winko, editors, *Kanon, Wertung und Vermittlung: Literatur in der Wissensgesellschaft*, pages 279–292. De Gruyter, 2012.

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

When No News is Bad News. News-Based Change Detection during COVID-19

Kristoffer L. Nielbo[1,2], Frida Hæstrup[1], Kenneth C. Enevoldsen[1], Peter B. Vahlstrup[1], Rebekah B. Baglini[2], and Andreas Roepstorff[2]

[1]Center for Humanities Computing Aarhus, Aarhus University, Denmark
[2]Interacting Minds Centre, Aarhus University, Denmark

Abstract

During the first wave of COVID-19, a peculiar phenomenon could be observed in the flow of news media content both within and between media platforms: as the pandemic spread, news media transformed into *Corona-news*. The information corollary of the content alignment, was that the novelty of news content went down as media focused monotonically on the pandemic event, which resulted in the, from a news media's perspective, paradoxical situation where the same news was repeated over and over. This information phenomenon, where a decrease in novelty persists, has previously been used to track change in media, but in this paper we use a Bayesian approach to test the claim that a persistent change in novelty can be used to detect change in news media content originating in negative events.

Keywords: newspapers; pandemic response; Bayesian change detection; information theory

1 Introduction

A peculiar behavior could be observed in news media when the first wave of COVID-19 virus spread across the world. In response to the pandemic event, the ordinary rate of change in news content was disrupted because nearly every story became associated with COVID-19. On the one hand, content novelty went down, because nearly every story became more similar to previous stories (i.e., news suddenly became 'Corona news'), but on the other hand, the COVID-19 association became more prevalent, resulting in, at least initially, an increase in content persistence. A recent study (Nielbo et al., 2021b) argues that this behavior is an example of the *news information decoupling* (NID) principle, according to which information dynamics of news media are (initially) decoupled by temporally extended catastrophes such that the content novelty

decreases as media focus monotonically on the catastrophic event, but the resonant property of said content increases as its continued relevance propagate throughout the news information system. The same study further indicated that NID can be used to detect significant change in news media that originate in catastrophic events. From the perspective of cultural dynamics, the COVID-19 pandemic provides a natural experiment that allows us to study the effect of a global catastrophe on the the dynamics of news media's information. While news media are neither unbiased nor infallible as sources of events, they do reflect preferences, values, and desires of a wide socio-cultural and political user spectrum. As such, news media coverage of COVID-19 functions as a proxy for how cultural information systems respond to unexpected and dangerous events.

Several studies have shown that the associative structure of news media is sensitive to socio-cultural dynamics, for instance the rise and intricacies of modernity as reflected in historical newspapers (Guldi, 2019; van Eijnatten and Ros, 2019), and value-based differential response to negative events such as instability and war (Daems et al., 2019). In a similar vein it has been shown that the ordered one-dimensional representation of the word co-occurrence structure quite accurately captures historically relevant trends in newspapers (Newman and Block, 2006). By embedding the co-occurrence structure in a low-dimensional space, it has been shown that newspaper content reflect fundamental cultural movements and understandings from the 19th century onward (van Eijnatten and Ros, 2019), and that these context-dependent representations are sensitive to cultural bias as reflected in newspapers (Wevers, 2019). In continuation of the 'Culturomics' trend that used Google Books to show how lexical variation is sensitive to events (Michel et al., 2011), a wide range of studies has demonstrated that simple word and concept frequencies are sufficient for robust offline detection of major historical events (Kestemont et al., 2014) and can be used to model the evolution of complicated cultural processes such as the historical interdependencies between media and politics (Bos et al., 2016). Fluctuations of time dependent word frequencies have been shown to discriminate between classes of events that have class-specific fractal signatures, where the social-cultural class display non-stationary and on-off intermittent behavior (Gao et al., 2012). Even within the social-cultural class, different types of events (or stories about events) seem to show fine-grained differences in their degree of self-affinity in newspapers (Wevers et al., 2020).

In line with recent developments in information theory, studies have used information theoretic measures to track the states and dynamics of socio-cultural systems as reflected in lexical data (Barron et al., 2018; Guldi, 2019; Murdock et al., 2015; Nguyen et al., 2020; Nielbo et al., 2019b). One paradigmatic study used relative entropy to study the development of Darwin's thinking in relation to his cultural context (Murdock et al., 2015). Both Shannon entropy and relative entropy have similarly been used in other studies to detect changes in prevalent mental states due to the socio-cultural context (e.g., state censorship, degree of recognition, religious observation) (Nielbo et al., 2019a,b). One specific information theoretic approach applies windowed relative entropy to dense low-dimensional text representations in order to generate signals that capture information *novelty*, $\mathcal{N}$, as a documents reliable content difference from the past; *transience*, $\mathcal{T}$, the documents content difference from future documents; and *resonance*, $\mathcal{R}$, the difference between *novelty* and *transience*, or conceptually, the degree to which future information conforms to a documents *novelty* (Barron et al., 2018; Nguyen et al., 2020). Taking a more dynamic perspective on this approach, one study has shown that discussion boards on social media where the novelty signal displays

both short-range correlations only and a particularly strong association with resonance are more likely to contain trending content (Nielbo et al., 2021a). Using the same approach combined with event detection, has also been shown to reliably predict major change points in historical data (Vrangbæk and Nielbo, 2021).

On the specific intersection between news media, COVID-19 and uncertainty, a group of economists has developed an index for economic policy uncertainty based on dense probabilistic representations of newspaper articles (Bess et al., 2020). The index correlates with existing market indices (e.g., VIX and BBO) and can accurately identify phase one of COVID-19 as well as other events associated with increased economic uncertainty. By using similar newspaper representations of front pages during COVID-19 and applying the above mentioned information theoretical approach, one study has argued that the information dynamics of news during COVID-19 were reflective of societal and value-based responses to the pandemic, and further argued that news media's response to the pandemic is a decoupling of the news content's novelty and resonance (i.e., news information decoupling) and, furthermore, that the decoupling may reflect political alignment with the current government (Nielbo et al., 2021b).

This study specifically tests the claim of Nielbo et al. (2021b), that NID-like behavior can provide input for change point detection algorithms. Specifically, we propose to test the claim that two change points are observable in news media during the first phase of COVID-19, *Lockdown* and *Opening* respectively, using a Bayesian approach to change point detection.

2 Methods

2.1 Data and Normalization

The data set consists of the linguistic content (title and body text) from front pages[1] of six Danish newspapers (Berlingske, BT, Ekstrabladet, Jyllands-Posten, Kristeligt Dagblad, and Politiken), see Table 1. All newspapers are national and published daily, with the exception of Kristeligt Dagblad that is only published six times per week from Monday to Saturday. Kristeligt Dagblad is kept in the sample because it is a national newspaper with a substantial circulation and represents a specific Danish reader segment. All newspapers were sampled from December 1, 2019 to June 30 2020 resulting in a corpus of $1,271,004$ tokens.

In order to normalize linguistic content, numerals and highly frequent function words were removed, and the remaining data were casefolded and lemmatized using a language-specific neural model trained on the Danish Dependency Treebank (Qi et al., 2020). Subsequently, the data were represented as a bag-of-words (BoW) model using latent Dirichlet allocation in order to generate a dense low-rank representation for each front page. Parameter sweep was used for hyperparameter optimization and leave-p-out cross-validation was used for testing generalization. Note that with appropriate modifications to the divergence measures, see equations (4) and (5) in the following section, the approach to change detection in text presented in this paper is applicable to any probabilistic or geometric vector-representation of documents. Novelty and resonance were estimated for windows of seven days, $w = 7$, representing the weekly news cycle.

[1] Front pages are used because they condense the most important news content and are, in comparison to full newspapers, more similar across conditions. Qualitatively similar results can be obtained from full newspapers although the results are subject to considerable more noise.

For validation purpose we used the official dates for the first lockdown March 13 and opening April 15, 2020. While it is uncontroversial to assume that the lockdown date was point-like, the opening was actually more gradual with several restrictions remaining and local lockdowns. The change detection model proposed is however sensitive to this gradual change in the interval of τ (see the next sections).

Source	Type	Political alignment
Berlingske	compact	center-right
BT	tabloid	center-right
Ekstrabladet	tabloid	independent
Jyllands-Posten	compact	center-right
Kristeligt Dagblad	broadsheet	independent evangelical
Politiken	broadsheet	center-left

Table 1: Danish newspaper data set. Column one contains the name of the newspaper, column two the type of newspaper (Broadsheet, Compact, or Tabloid), and three approximate political alignment of the newspaper. It is important to note that the newspapers do not have direct affiliation with political parties today and that the alignment reflect their own classification. In some cases, this may not represent the perception of the readers.

2.2 Novelty and Resonance

Two related information signals were extracted from the temporally sorted BoW model: *Novelty* ($\mathcal{N}$) as an article $s^{(j)}$'s reliable difference from past articles $s^{(j-1)}, s^{(j-2)}, \ldots, s^{(j-w)}$ in window w:

$$\mathcal{N}_w(j) = \frac{1}{w} \sum_{d=1}^{w} JSD(s^{(j)} \mid s^{(j-d)}) \tag{1}$$

and *resonance* ($\mathcal{R}$) as the degree to which future articles $s^{(j+1)}, s^{(j+2)}, \ldots, s^{(j+w)}$ conform to article $s^{(j)}$'s novelty:

$$\mathcal{R}_w(j) = \mathcal{N}_w(j) - \mathcal{T}_w(j) \tag{2}$$

where $\mathcal{T}$ is the *transience* of $s^{(j)}$:

$$\mathcal{T}_w(j) = \frac{1}{w} \sum_{d=1}^{w} JSD(s^{(j)} \mid s^{(j+d)}) \tag{3}$$

The novelty-resonance model was originally proposed in Barron et al. (2018), but here we apply a symmetrized and smooth version by using the Jensen–Shannon divergence (JSD):

$$JSD(s^{(j)} \mid s^{(k)}) = \frac{1}{2} D(s^{(j)} \mid M) + \frac{1}{2} D(s^{(k)} \mid M) \tag{4}$$

with $M = \frac{1}{2}(s^{(j)} + s^{(k)})$ and D is the Kullback-Leibler divergence:

$$D(s^{(j)} \mid s^{(k)}) = \sum_{i=1}^{K} s_i^{(j)} \times \log_2 \frac{s_i^{(j)}}{s_i^{(k)}} \tag{5}$$

2.3 Nonlinear Adaptive Filtering

Nonlinear adaptive filtering is applied to the information signals because of the their inherent noisiness (Gray, 2007). First, the signal is partitioned into segments (or windows) of length $w = 2n + 1$ points, where neighboring segments overlap by $n + 1$. The time scale is $n + 1$ points, which ensures symmetry. Then, for each segment, a polynomial of order D is fitted. Note that $D = 0$ means a piece-wise constant, and $D = 1$ a linear fit. The fitted polynomial for ith and $(i + 1)th$ is denoted as $y^{(i)}(l_1), y^{(i+1)}(l_2)$, where $l_1, l_2 = 1, 2, ..., 2n + 1$. Note that the length of the last segment may be shorter than w. We use the following weights for the overlap of two segments.

$$y^{(c)}(l_1) = w_1 y^{(i)}(l + n) + w_2 y^{(i)}(l), l = 1, 2, \ldots, n + 1 \tag{6}$$

where $w_1 = (1 - \frac{l-1}{n}), w_2 = 1 - w_1$ can be written as $(1 - \frac{d_j}{n}), j = 1, 2$, where d_j denotes the distance between the point of overlapping segments and the center of $y^{(i)}, y^{(i+1)}$. The weights decrease linearly with the distance between point and center of the segment. This ensures that the filter is continuous everywhere, which ensures that non-boundary points are smooth.

2.4 Change Point Detection

In order to model media response to COVID-19, we follow Nielbo et al. (2021a) and state that it is only when a reliable difference in $\mathcal{N}$ can be observed that we inspect the relationship between $\mathcal{N}$ and $\mathcal{R}$. We propose a simple Bayesian approach to model changes in the mean of $\mathcal{N}$, a Bayesian mean-shift model. The simplest case of news information decoupling is a temporary state change corresponding to Denmark open $\rightarrow$ lockdown $\rightarrow$ open. This model implies three means, two of which are approximately identical, and two shifts, decoupling start and decoupling end, in $\mathcal{N}$, corresponding the beginning and the end of the lockdown respectively. Or more formally, we assume that the time series $\mathcal{N}$ contains two change points in time, τ_1 and τ_2, that can be located anywhere in the signal. While the last part of the the assumption clearly disregards information about the occurrence of the lockdown, it simplifies the model and is sufficient to detect one reliable decoupling, decoupling start and decoupling end, in $\mathcal{N}$. Aside from the change points, we assume that the series is stable and follow a normal distribution with varied mean, μ_i, and singular variance, σ. Notice that these assumptions follow from a mean-shift model. This gives us the following model for the observed $\mathcal{N}_i$:

$$\mathcal{N}_t = \begin{cases} \text{Normal}(\mu_1, \sigma) \text{ for } t < \tau_1 \\ \text{Normal}(\mu_2, \sigma) \text{ for } \tau_1 \leq t < \tau_2 \\ \text{Normal}(\mu_3, \sigma) \text{ for } t \geq \tau_2 \end{cases} \tag{7}$$

for which we wish to estimate the location of the change points τ_i, and the value of the means μ_i and variance σ, resulting in the following posterior:

$$P(\mu_i, \sigma, \tau_i | \mathcal{N}_t) = P(\mu_1, \mu_2, \mu_3, \sigma, \tau_1, \tau_2 | \mathcal{N}_t) \tag{8}$$

For estimation of the posterior, we used NUTS sampling with 4000 samples (Salvatier et al., 2015). Again, the estimation was done using slightly conservative priors assuming that the change points, τ_i, can be anywhere in the sequence (with $\tau_2 > \tau_1$) and that the variance, σ, is stable across change points. Note that the half Cauchy prior

distribution has several beneficial properties most importantly fat tail which allows for extreme values (Gelman et al., 2013; Polson and Scott, 2012). These assumptions were modelled using the following priors:

$$\begin{aligned} \mu_i &\sim \text{Normal}(0, 0.5) \\ \sigma &\sim \text{Half Cauchy}(0.5) \\ \tau_1 &\sim \text{Uniform}(0, \max(\mathcal{N}_t)) \\ \tau_2 &\sim \text{Uniform}(\tau_1, \max(\mathcal{N}_t)) \end{aligned} \tag{9}$$

In order for the model to be valid, that is detect the change points in $\mathcal{N}$, the following is required for : 1) the relevant dates (lockdown and opening) should be contained in the estimate intervals (i.e., the actually events should detected); and 2) the intervals of the decoupling start and decoupling end should be non-overlapping (i.e., there model should detect a reliable state change).

2.5 News Information Decoupling

Finally, in order to describe the information states before and after an event and confirm if a change point reflects a decoupling (i.e., that novelty decreases while resonance increases), we fit resonance on novelty to estimate the $\mathcal{NR}$ slope β_1 before and during the event in question (e.g., the Danish lockdown):

$$\mathcal{R}_i = \beta_0 + \beta_1 \mathcal{N}_i + \epsilon_i, \;\; i = 1, \ldots, n. \tag{10}$$

where β_0 is the intercept and ϵ is a random variable representing the errors of the fit.

3 Results

To establish a baseline for novelty and resonance, we computed the per newspaper linear slope for resonance on novelty ($\mathcal{NR}$) from December 01, 2019 to February 26, 2020 (the first incidence of COVID-19 in Denmark was registered on February 27, 2020). As can be observed from Figure 1, the slopes are remarkably similar, indicating a medium to strong association between novelty and resonance ($M = 0.56$, $SD = 0.06$) before the national outbreak of COVID-19. In the normal state of affairs, novelty and resonance therefore seems to be coupled such that novel news items resonate more than overused and repetitive items and vice versa. This general news dynamic confirms the intuition that news media, all things being equal, maintain their relevance by propagating news.

Figure 2 displays a prototypical example of NID during the first phase of COVID-19 (Nielbo et al., 2021b). Although COVID-19 news items date back to December 2019, '*Wuhan*', newspaper content is not impacted until the period after the first national outbreak (in this case in Denmark). '*Virus*'. From the phase 1 lockdown '*Lockdown*' to the opening, '*Opening*', the newspaper shows a valley in novelty and, initially, a peak in resonance until both processes approximately return to normal after the opening. Figure 3 shows the same trend, but with a noticeable difference in the time it takes *Berlingske* to return to a normal state of affairs and (i.e., '*Opening*' then is less pronounced).

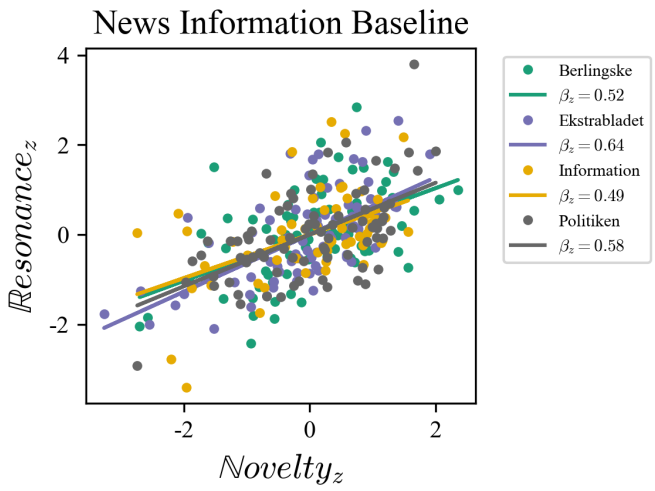


Figure 1: $\mathcal{NR}$ slope baseline for four national newspapers that represents the left-right political spectrum. Data are sampled before COVID-19 phase 1 in Denmark initiated. Information has been included in the graph, which is a left-wing broadsheet newspaper.

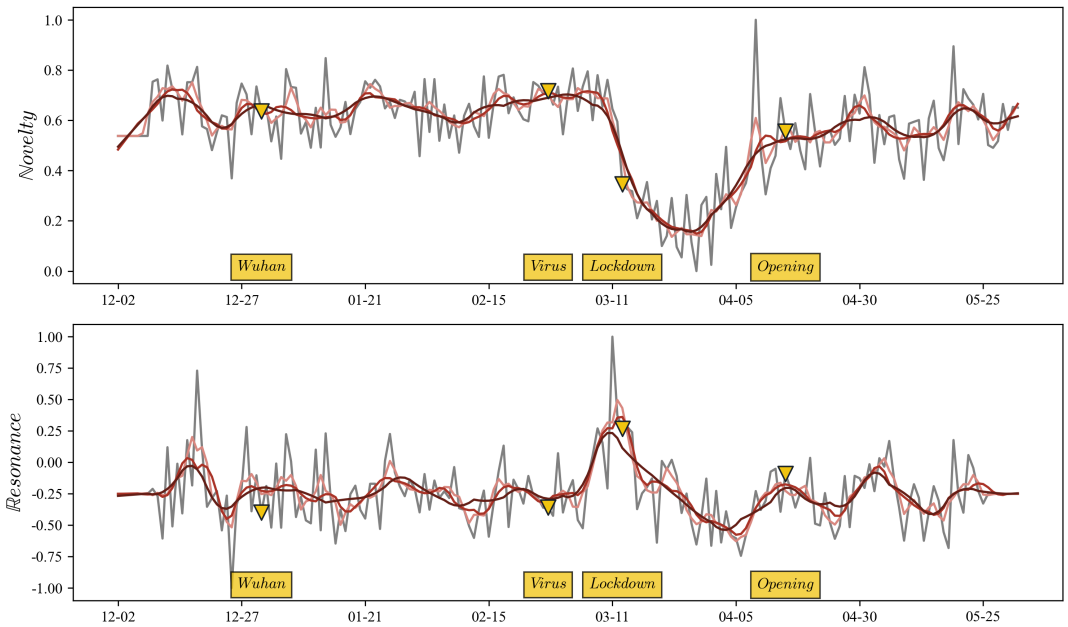


Figure 2: Novelty (upper panel) and resonance (lower panel) for the center-left newspaper *Politiken* before and during COVID-19 phase 1. Trend lines in the upper and lower panel are estimated using a nonlinear adaptive filter for three different levels of smoothing to emphasize the robustness of the trend, see Equation 6.

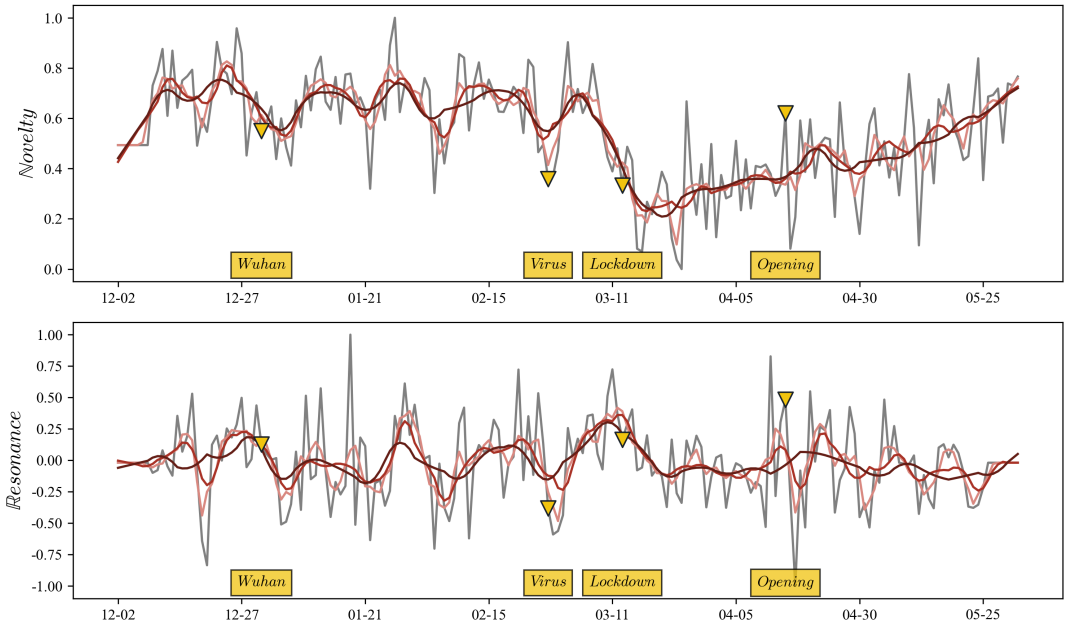


Figure 3: Novelty (upper panel) and resonance (lower panel) for the center-right newspaper *Berlingske* before and during COVID-19 phase 1. Trend lines in the upper and middle panel are estimated using a nonlinear adaptive filter, see Equation 6.

To validate the observed behavior, we tested for two change points in novelty using a Bayesian model. The first change point, '*NID* Start' should separate pre-lockdown from lockdown centered on week 11 (March 9-15), and the second lockdown, '*NID* End' from post opening (centered on week 16, April 13-19). Table 2 shows the estimated change points for six national newspapers, two of which are *T*abloid newspapers (Class) and the remainder *B*roadsheet/*C*ompact. From the model, it can be observed that all broadsheet/compact newspapers seem to support the NID principle in novelty. The first change point is placed in weeks 10-11, the second, however, is more a matter of contention. The opening change point lies within April and displays a month's delayed response. Finally, it can be observed that tabloid press shows no indication of NID behavior. Table 3 shows the posterior distributions for novelty, their means and highest density intervals, for broadsheet/compact newspapers, which clearly indicates that they do conform to NID.

Source	Type	*NID* Start	*NID* End	*NID*
Berlingske	*C*	03.07 [03.03, 03.09]	04.28 [04.09, 05.08]	*True*
BT	*T*	04.10 [12.30, 09.01]	07.25 [04.22, 09.03]	*False*
Ekstrabladet	*T*	01.28 [01.02, 03.17]	05.08 [01.16, 07.22]	*False*
Jyllands-Posten	*C*	03.10 [03.08, 03.14]	05.25 [05.21, 06.06]	*True*
Kristligt Dagblad	*B*	03.07 [03.05, 03.12]	04.15 [04.11, 04.17]	*True*
Politiken	*B*	03.13 [03.12, 03.13]	04.08 [04.05, 04.08]	*True*

Table 2: Estimated temporal change points at 94% high density intervals for novelty. Column one contains the name of the newspaper, columns two its type (*B*roadsheet, *C*ompact *T*abloid), NID Start and End is the beginning and end of the lockdown as represented in the newspaper, and the final column indicated if the specific source supported the NID principle.

Source	$\mathcal{N}_{pre}$	$\mathcal{N}_{NID}$	$\mathcal{N}_{post}$
Berlingske	0.36 [0.35, 0.37]	0.29 [0.27, 0.31]	0.34 [0.34, 0.35]
Jyllands-Posten	0.29 [0.28, 0.30]	0.23 [0.22, 0.24]	0.27 [0.26, 0.28]
Kristligt Dagblad	0.27 [0.26, 0.28]	0.19 [0.18, 0.21]	0.26 [0.25, 0.27]
Politiken	0.27 [0.26, 0.28]	0.15 [0.14, 0.17]	0.26 [0.25, 0.26]

Table 3: Estimates of mean $\mathcal{N}$ values at 94% high density intervals before during and after the lockdown for the four broadsheet newspapers that supported the NID principle, see Table 1. All newspapers show a reliable reduction of $\mathcal{N}$ during the lockdown.

That novelty decreases during a catastrophic event is nevertheless only half the story. For NID to be supported by the data, resonance should increase during the lockdown such that the medium to strong association between novelty and resonance is momentarily weakened. Following Nielbo et al. (2021b) we inspected the time-windowed linear fits of resonance on novelty, $\mathcal{NR}$ slopes, in order to confirm this, see Figure 4. All broadsheet newspapers display a slope decrease during the lockdown, thereby conforming to the NID principle 4. Tabloids on the other hand, follow an inverse pattern, such that the $\mathcal{NR}$ slope increases during the lockdown period.

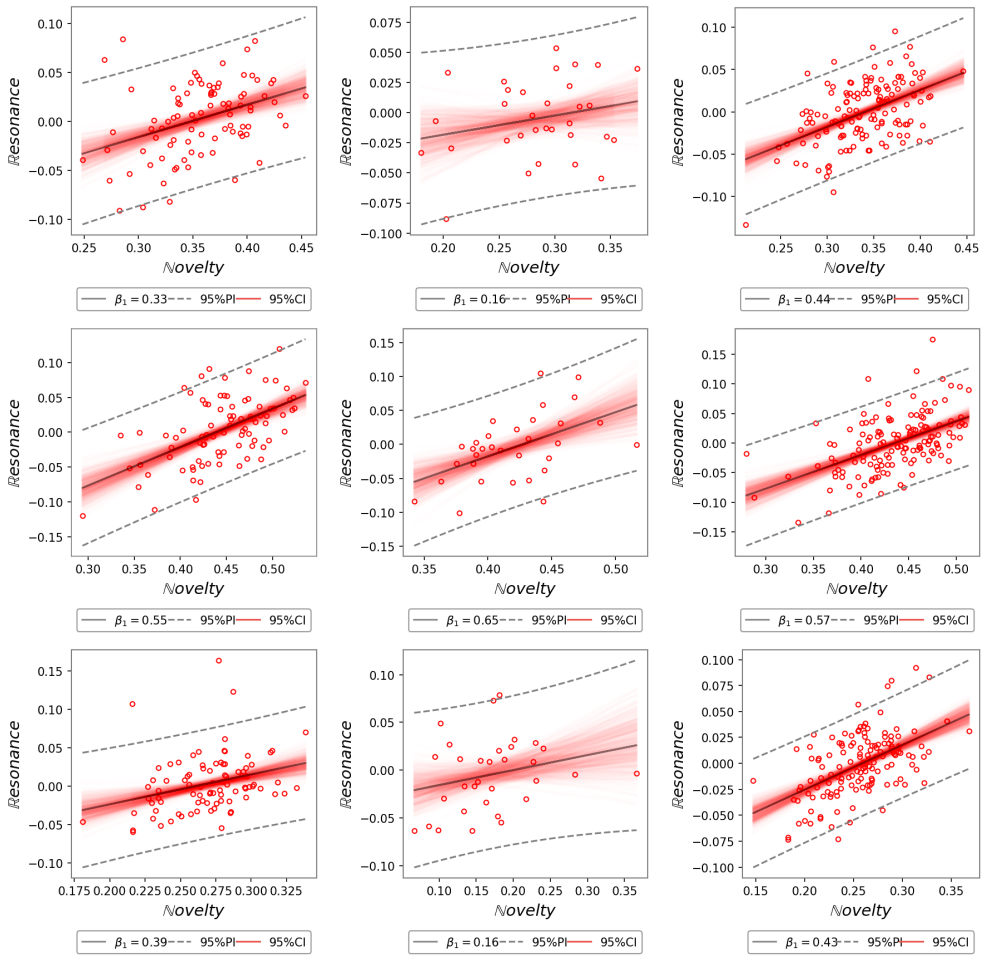


Figure 4: $\mathcal{NR}$ slopes before during and after the lockdown for Berlingske (upper row), Ekstrabladet (middle row), and Politiken (lower row) during COVID-19 phase 1. Punctured lines represents the 95% prediction intervals, and semi-transparent lines the 95% confidence intervals using a bootstrap approach with 500 samples.

Source	Type	$\mathcal{NR}_{pre}$	$\mathcal{NR}_{NID}$	$\mathcal{NR}_{post}$
Berlingske	*C*	0.33 $[0.17, 0.51]$	0.16 $[-0.07, 0.38]$	0.44 $[0.32, 0.58]$
BT	*T*	0.49 $[0.29, 0.66]$	0.55 $[0.28, 0.83]$	0.26 $[0.08, 0.43]$
Ekstrabladet	*T*	0.55 $[0.38, 0.72]$	0.65 $[0.26, 1]$	0.57 $[42, 0.71]$
Jyllands-Posten	*C*	0.42 $[0.24, 0.63]$	0.31 $[0.04, 0.56]$	0.39 $[0.25, 0.51]$
Kristligt Dagblad	*B*	0.57 $[0.34, 0.78]$	0.43 $[0.06, 0.78]$	0.76 $[0.55, 0.95]$
Politiken	*B*	0.39 $[0.14, 0.61]$	0.16 $[-0.05, 0.37]$	0.43 $[0.32, 0.54]$

Table 4: $\mathcal{NR}$ coefficients at 95% confidence intervals before during and after the lockdown for all newspapers in the sample. Column two contains the newspaper type (*B*roadsheet, *C*ompact *T*abloid).

4 Discussion

This study has sought to validate the news information decoupling (NID) principle on a sample of six national newspapers from Denmark during the first phase of COVID-19. Using a Bayesian approach to change point detection, we showed that content novelty in broadsheet newspapers does indeed display statistically reliable points of change coinciding with the COVID-19 lockdown and opening. NID was further corroborated by the $\mathcal{NR}_{pre}$ slopes that indicated a decoupling of resonance from novelty during the lockdown. Several direct observations can be made from the findings. First, the estimated change points for the 'Pre-lockdown $\rightarrow$ Lockdown' are spread over a two week interval, which indicates that a lockdown could be reasonably predicted already from the first COVID-19 incident in Denmark. Second, in a similar vein, the 'Lockdown $\rightarrow$ Opening' change point interval is spread over a full month from April 8 to May 8, which may reflect disagreement about if and when the lockdown ended. The Danish government, during the first phase of COVID-19, was center-left and the model's uncertainty in determining the opening may reflect political alignment of the newspapers (Nielbo et al., 2021b), where center-right newspapers (e.g., Berlingske and Jyllands-Posten) were more sceptical towards the government's implementation of an opening than the center-left (e.g., Politiken). In other words, the center-right may have been more reluctant to acknowledge the opening as a return to normal. Third, tabloid newspapers did not show any indication of a news decoupling. On the contrary, their $\mathcal{NR}_{pre}$ slopes momentarily increased during the lockdown. This increase in slopes does, however, not provide any useful information, because, as shown by our change point detection model, the periodization is not meaningful for the two tabloid newspapers.

As already mentioned, validation of the NID principle is still needed for multilingual data and its value for crisis management should be further tested. For change detection, the scope of the principle needs additional testing; does NID generalize beyond a small set of negative events to, for instance, temporally extended significant events (e.g., moon landing, fall of the Berlin Wall). Finally, several contrast already hinted at need to be tested, e.g., left vs. right-wing newspapers, tabloid vs. broadsheet newspapers, silly season and other seasonal effects, are all interesting venues for media and journalism researchers.

We advise caution in developing rich (domain-specific) interpretations of the model's behavior, because the behavior primarily shows that something is happening (i.e., change detection), but not what is happening (i.e., change characterization) during the lockdown. There are however two possible interpretations of NID that we would like to foreground in this context: an unmediated and a mediated interpretation

of the observed NID behavior. According to the unmediated interpretation, NID is a context-independent response to negative events, where news focus monotonically on the event in question (e.g., COVID-19 lockdown), thereby lowering the the overall uncertainty or unpredictability of the regular news cycle. The unmediated account does not imply that NID is a deterministic response to negative events as such. It is possible that NID only captures a class of events (e.g., temporally extended catastrophic events) similar to how the persistence of information processes can classify event types (Gao et al., 2012; Wevers et al., 2020). Against this account is the observed differences between newspaper types, the broadsheet/compact vs. tabloid contrast. If NID is truly unmediated, we are likely to expect a more uniform response from all newspaper types. According to the alternative account, NID is a mediated response to negative events given certain cultural and societal conditions. Mediating factors could be societal coordination mechanisms reflected in complex social variables such as trust in government, social uncertainty, and economic equality. Denmark belongs to the Nordic group of universal welfare states that is characterized by high levels of trust in the government, and low levels of social uncertainty and inequality. It may be that NID reflects the close coordination between media, population and government under these specific conditions. The differential NID behavior observed in the contrast between political alignment, the centre-left vs centre-right contrast, may well support this interpretation. That centre-right newspapers were more reluctant than centre-left newspapers to acknowledge the return to a normal state of affairs, may very well indicate a negotiation of trust in government. To properly validate and decide between these interpretations, we need to develop additional models for relevant contrasts (e.g., news data from other Nordic countries).

From a more theoretical perspective, we would like to propose an interpretation of the combination of information-theoretical and change detection models as providing a potential state variable or indicator of societal uncertainty. To properly understand the behavior of complex socio-cultural systems in response to a catastrophe like COVID-19, we have to continuously monitor the states of said system. Ideally, we want to monitor all intrinsic variables related to uncertainty, and potentially trust and inequality given our above mentioned un-/mediated interpretations. This strategy however is not feasible and instead we suggest to rely on the fundamental embedding theorem of chaos theory, which states that the detailed dynamics of a system that has an underlying attractor can be readily studied by reconstructing a suitable phase space of a scalar time series recorded from the system (Packard et al., 1980; Sauer et al., 1991; Takens, 1981). Chaos theory offers an elaborate scheme for generating aperiodic, highly irregular data from a deterministic system that can be characterized by only very few state variables instead of a random system with infinite numbers of degrees of freedom (Gao and Xu, 2021). While the evolution of a complex social system may not be modeled as a dynamical system with a single attractor, we can assume that the dynamics of a large-scale social system can be approximated by switching between a large number of attractors, some of which may be simple, such as fixed points that may be associated with the dynamics of cultural information, while others may be complicated, including chaotic attractors (Gao et al., 2007; Ott, 2002). In order to understand societal uncertainty in face of the pandemic, we need to find an adequate continuous variable related to cultural information that is shared by members of the society during the event in question. We propose that a model of uncertainty in news media could very well provide such an adequate variable. On this account, the approach to news information dynamics presented here provides a valuable tool for

media-based indices that can supplement existing economic and policy-based indices of uncertainty (Bess et al., 2020).

5 Online Resources

All data are proprietary and have been collected through Infomedia's API: https://infomedia.dk/. For inquiries regarding models and derived data, please contact chcaa@cas.au.dk. The source code for methods is available on Github: https://bit.ly/3beahFd. More details on NID detection can be found at NeiC's NDHL website: https://bit.ly/3bfeW9C.

6 Acknowledgments

This research was supported the "HOPE - How Democracies Cope with COVID-19"-project funded by The Carlsberg Foundation with grant CF20-0044, NeiC's Nordic Digital Humanities Laboratory project, and DeiC Type-1 HPC with project DeiC-AU1-L-000001. The authors would like to thank Berlingske Media, JP/Politkens Hus, and Kristeligt Dagblad for providing access to proprietary data.

References

Alexander T. J. Barron, Jenny Huang, Rebecca L. Spang, and Simon DeDeo. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612, 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1717729115.

Mikkel Bess, Erik Grenestam, Alessandro Tang-Andersen Martinello, and Jesper Pedersen. Uncertainty and the real economy: Evidence from denmark. *Working Paper - Danmarks Nationalbank*, 165, 2020.

Patrick Bos, Huub Wijfjes, Maaike Piscaer, and Gerrit Voerman. Quantifying "pillarization": Extracting political history from large databases of digitized media collections. *Proceedings of the 3rd HistoInformatics Workshop*, page 10, 2016.

Joke Daems, Thomas D'haeninck, Simon Hengchen, Tecle Zere, and Christophe Verbruggen. 'workers of the world'? a digital approach to classify the international scope of belgian socialist newspapers, 1885–1940. *Journal of European Periodical Studies*, 4(1):99–114, 2019. ISSN 2506-6587. doi: 10.21825/jeps.v4i1.10187.

Jianbo Gao and Bo Xu. Complex Systems, Emergence, and Multiscale Analysis: A Tutorial and Brief Survey. *Applied Sciences*, 11(12):5736, June 2021. ISSN 2076-3417. doi: 10.3390/app11125736.

Jianbo Gao, Yinhe Cao, Wen-wen Tung, and Jing Hu. *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond*. Wiley-Interscience, 1 edition edition, 2007. ISBN 978-0-471-65470-4.

Jianbo Gao, Jing Hu, Xiang Mao, and Matjaž Perc. Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. *Journal of The Royal Society Interface*, 9(73):1956–1964, 2012.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, 3 edition edition, 2013. ISBN 978-1-4398-4095-5.

Katharine Lynn Gray. Comparison of Trend Detection Methods. *Graduate Student Theses, Dissertations, & Professional Papers*, 228:98, 2007.

Jo Guldi. The measures of modernity: The new quantitative metrics of historical change over time and their critical interpretation. *International Journal for History, Culture and Modernity*, 7(1):899–939, 2019. ISSN 2666-6529, 2213-0624. doi: 10.18352/hcm.589.

Mike Kestemont, Folgert Karsdorp, and Marten Düring. Mining the twentieth century's history from the time magazine corpus. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (*LaTeCH*), pages 62–70. Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-0609.

J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1199644.

Jaimie Murdock, Colin Allen, and Simon DeDeo. Exploration and exploitation of victorian science in darwin's reading notebooks. *arXiv:1509.07175*, 2015.

David J. Newman and Sharon Block. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, 57(6):753–767, 2006. ISSN 15322882, 15322890. doi: 10.1002/asi.20342.

Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, 3:62, 2020. ISSN 2624-8212. doi: 10.3389/frai.2020.00062.

K.L. Nielbo, P.B. Vahlstrup, and A. Bechmann. Trend reservoir detection: Minimal persistence and resonant behavior of trends in social media. *Proceedings of Computational Humanities Research*, 1, 2021a.

Kristoffer L Nielbo, Katrine F Baunvig, Bin Liu, and Jianbo Gao. A curious case of entropic decay: Persistent complexity in textual cultural heritage. *Digital Scholarship in the Humanities*, 34(3), 2019a. ISSN 2055-7671. doi: 10.1093/llc/fqy054.

Kristoffer L. Nielbo, M. L. Perner, C. P. Larsen, Jonas Nielsen, and D. Laursen. Automated compositional change detection in saxo grammaticus' gesta danorum. In *DHN*, pages 320–332, 2019b.

Kristoffer L. Nielbo, Rebekah B. Baglini, Peter B. Vahlstrup, Kenneth C. Enevoldsen, Anja Bechmann, and Andreas Roepstorff. News information decoupling: An information signature of catastrophes in legacy news media. *arXiv:2101.02956* [*cs*], 2021b.

Edward Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 2 edition, 2002. ISBN 978-0-521-01084-9. doi: 10.1017/CBO9780511803260.

N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Lett.*, 45(9):712–716, 1980. doi: 10.1103/PhysRevLett.45.712. Publisher: American Physical Society.

Nicholas G. Polson and James G. Scott. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 2012. ISSN 1936-0975. doi: 10.1214/12-BA730.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv:2003.07082* [*cs*], 2020.

John Salvatier, Thomas Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using PyMC. *arXiv:1507.08050* [*stat*], 2015.

Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. *Journal of Statistical Physics*, 65(3):579–616, 1991. ISSN 1572-9613. doi: 10.1007/BF01053745.

Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381. Springer Berlin Heidelberg, 1981. ISBN 978-3-540-38945-3.

Joris van Eijnatten and Ruben Ros. The eurocentric fallacy. a digital-historical approach to the concepts of 'modernity', 'civilization' and 'europe' (1840–1990). *International Journal for History, Culture and Modernity*, 7(1):686–736, 2019. ISSN 2666-6529, 2213-0624. doi: 10.18352/hcm.580.

E.E.H Vrangbæk and K.L. Nielbo. Composition and change in de civitate dei: A case study of computationally assisted methods. *Studia Patristica*, 2021.

Melvin Wevers. Using word embeddings to examine gender bias in dutch newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-4712.

Melvin Wevers, Jianbo Gao, and Kirstoffer L. Nielbo. Tracking the consumption junction: Temporal dependencies between articles and advertisements in dutch newspapers. *Digital Humanities Quarterly*, 014(2), 2020.

2022. This work is licensed under a Creative Commons "Attribution 4.0 International" license.

Combining Tools with Linked Data: a Social History Example

Ivo Zandhuis[1,2]

[1]Ivo Zandhuis Research & Consultancy
[2]International Institute of Social History

This papers presents a workflow on how to link, retrieve and create data for historical research in a Linked Data format. Through this workflow, one can create a 'web of data' that contains a representation of all the things that are relevant to a particular history research project. For example, the workflow can accommodate publications and archival sources (linked to online catalogues or retrieved from data set providers and registered in Zotero, Tropy and Recogito) and organizations, persons and songs (created with CoW and burgerLinker). As an example a 19th century labour history project is presented on the development of collective action among print labourers. While describing the workflow, the paper advocates the use of Linked Data and investigates how well tools are equipped to realize interoperable data with this technique, using the three Linked Data principles introduced by Tim Berners-Lee in 2009.

Keywords: Linked Open Data; methodology; reproducibility

1 The Web of Data as our research infrastructure

Collecting, organizing and analysing sources and references without computational tools is unthinkable in current historical scholarship. While finding and combining these sources and references, tools can help scholars to iteratively develop interesting research questions and arguments that support or contradict their hypotheses. Ultimately, these arguments always refer to the sources they are based upon.

Every scholar prefers their own tool set to support this process and some of the tools are unavoidable in a certain domain. Personally, I prefer to use Zotero[1] for storing and organizing references, Tropy[2] to organize photos I take in the reading room, and Recogito[3] to organize the machine readable representation of a text. Additionally, I create and use data on individuals where I wish to keep a close relation with the original source. These micro observations are published online in data sets and the

1 https://zotero.org
2 https://tropy.org
3 https://recogito.pelagios.org

aforementioned combination of tools helps me to relate every digital representation and observation to the right archival material or literature reference.

Preferably, research is thought of as creating data to substantiate an argument, instead of using tools investigate a topic. The use of tools is only a means to develop a network of data that refers to the various building blocks of the research and their provenance. This network of data, or in mathematical terms the 'graph', helps to create a thorough and reproducible historical explanation, independent of the tools we use. Ideally, this data should be processable in an automatic way. To be more precise: it needs to satisfy the FAIR-principles of being Findable, Accessible, Interoperable and Reusable (Wilkinson et. al., 2016). Unfortunately, the tools I use each have their own environment to store and manipulate data. Without devoting special attention to creating *one* network of data, using the various tools would result in various unrelated data sets. That means the data would not comply to the FAIR-principles.

These principles prescribe that the data must be exported and published. By applying a standard for syntactical representation of the data and standardized protocols for publishing the data, most principles are met. Findability, Accessibility and Reusability are improved massively by doing this. Interoperability of data is achieved by adding machine readable semantics and relations: instead of a human readable code book, a data provider should add a definition of the elements in their data according to standardized syntax and protocols as well.

Because of its gradual adoption and semantic potential, Linked Data is the most obvious technique to use for creating the network of FAIR, but especially interoperable, data that we envision. The technique consists of specifications for (1) identifying concepts in the data and (2) relating them to other concepts. These specifications are independent of tools or computer platforms and enable decentralized data definition and storage. For that reason they are especially suitable for creating interoperable data.

Concretely, publishing information resources as Linked Data can be done by transforming the data stored in the used tool to the format as described by the Resource Description Framework (RDF) specification (Cyganiak et al., 2014). To explain RDF to a more broader audience than computing scientists, Tim Berners-Lee introduced three main principles for Linked Data (Berners-Lee, 2009). For a concise introduction into Linked Data I explain them here briefly and I use them to evaluate the implementation of Linked Data in the tools I discus in this paper.

1. *All conceptual things should have a name starting with HTTP.* This means that every entity you want to publish an information resource for, must have a web address, more precisely a so-called Uniform Resource Identifier ('URI'). The entities could, for example, be a book, the member of an association or a location.

2. *Looking up an HTTP name should return useful data about the thing in question in a standard format.* If you call the web address, the web server must serve the data about the referred conceptual thing in RDF format. In the RDF format, the data properties are expressed. These properties are for instance the title of the book, the date of birth of a person or the coordinates on a map of a location.

3. *Anything else that that same thing has a relationship with through its data, should also be given a name beginning with HTTP.* In the data, for instance, the book is related with its author using the URI of that author (e.g. `http://www.wikidata.org/entity/Q80`) instead of the label ("Berners-Lee, T.J."). This means that you can follow the link and use the web address of the author to collect data about

the author. In this way you can, for instance, retrieve their birth date (for Tim Berners-Lee this would be "8 June 1955"), without having to register this yourself. Linking data like this results in a "Web of Data".

In this paper, I explain how I use the tools and export the data in Linked Data format to create one graph containing a representation of all the concepts relevant to my research, like publications, archival sources, organizations and persons. I introduce an example from my 19th century Labour History project on the development of collective action among print labourers. The paper explicitly does not provide a blue print for doing (social) historical research with digital tools. The described tools are my personal (and therefore rather arbitrary) choice. The paper advocates the use of Linked Data and investigates how well these specific tools are equipped to realize interoperable data with this technique as an inspiration for tool developers to learn how this functionality can be improved in their own project. It might help researchers to understand the criteria they need while evaluating the tool set they use.

The scripts I made and the resulting graph are available on github.[4] Remember this is a work in progress, so the repository changes over time. The remainder of this paper is organised as follows. In Section 2, I introduce the research I am currently undertaking and use as an example in the rest of the paper. After that, in Section 3, I investigate the way I am able to link and retrieve Linked Data from cultural heritage collections and data repositories and discuss my findings. Section 4 concentrates on the creation of Linked Data from the data stored in the tools I use. And again I discuss my findings. In Sections 5 and 6, I discuss the use and usefulness of Linked Data in general. Finally, Section 7 sums up the overall conclusions.

2 An example from 19th century labour history

For my project, I'm building a web of data for research into a Dutch phenomenon called 'typografische verenigingen' (typographical associations). During the first half of the 19th century print labourers in The Netherlands organized themselves in these local associations, comparable to the English 'friendly societies' and the French 'sociétés mutuelles' (Linden, 1996). These associations were founded to ensure health benefits and they organized a yearly feast to celebrate their identity as 'children of Laurens Coster', the Dutchman they believed invented printing. They were connected in a nationwide social network and organized the erection of a statue in Coster's honour in 1856 (see Figure 1). Eventually, this led to the establishment of the first national trade union in The Netherlands in 1866 (Giele, 1972). I'm interested in how this phenomenon originated.

Modern information technology helps me to process more details to form a complete picture of the development of collective action among print labourers. I can create lists of people involved and register their attributes, like their age and marital status. If I am able to find their (family-)relations, I can use Social Network Analysis to find patterns in the diffusion of the phenomenon. Collecting the songs the members of the associations sang during their feast (which they obviously printed), helps me study potential shifts in their interests. Minutes of their meetings refer to persons and befriended associations, and provide detailed insights in their relations as well. Besides these quantitative opportunities, references to publications enable us to use historiography about the subject and refer to details in other publications.

[4] https://github.com/ivozandhuis/typografische-verenigingen.

Figure 1: Lithograph showing the unveiling of the statue of Laurens Jansz. Coster in Haarlem, 1856 (Noord-Hollands Archief, https://hdl.handle.net/21.12102/3E03EEFEFB8F11DF9E4D523BC2E286E2)

In this paper, I present as an example, the data about J. H. Regenboog, mentioned frequently in the sources in the 1850s. We learn about his activities and find relevant family relations. This results in a part of the web of data visualized in Figure 2.

3 Linking and retrieving Linked Data

The current 'web of documents' is a base for the creation of the 'web of data' that is anticipated. Important online sources can be organized by providing simple and stable web addresses for every element in the collection of an heritage institution. Some online catalogues and data repositories already help us to link the right source.

3.1 Linking data in online catalogues

Most of the sources I need to study for my project are held by the International Institute of Social History (IISH), the Library of the University of Amsterdam (UB-UvA) and the Noord-Hollands Archief (NHA) in Haarlem. I found a lot of sources with a full-text search on Worldcat[5] for publications and with a search on Archives Portal Europe[6] for archival material.

I register my findings in Zotero. Zotero is a convenient tool for recording resources and for creating footnotes in the historical papers I want to write. Thanks to existing import scripts I was able to easily feed the Worldcat data into Zotero. For archival

5 https://www.worldcat.org/
6 https://archivesportaleurope.net/

materials I needed to insert the data in Zotero by hand. Unfortunately Zotero is unable to handle the hierarchical nature of archival descriptions, so I had to design a work-around for that. I published the overview of the sources on the Zotero website.[7]

For every item I registered I returned to the original online catalogue of the institute holding the material. I checked the data and stored the web address for reference to this original catalogue. Both the IISH and the NHA provide a stable web address according to the Handle principle, while the UB-UvA has chosen to use ARK.[8] That way these institutes made a step towards implementing Berners-Lee's first Linked Data principle. The same book might be held by various institutions or published on Google Books. When this is the case, I store the web addresses of the other manifestations of the same publication in Zotero as an 'attachment' typed 'Link to URI'.

For example, one of the sources that are relevant for my research, is the nationwide yearbook print labourers created in 1856 (Mommaas, 1856). This book contains information about all the associations that existed at that moment. A short history of its origination and development is presented, as well as the founders, board members and members. I registered this yearbook in Zotero and stated that it has a digitized version on Google Books and that the original is held by the UB-UvA. Additionally, Zotero refers to a tabular data-file I created and published on GitHub, listing all members mentioned in the yearbook. One of these members is 'J.H. Regenboog', a board member of the typographical association in The Hague.

The UB-UvA has published information resources from their catalogues in RDF format and the URIs that they have introduced can be used to request the data on the item in the catalogue in RDF format.[9] That way they have fully implemented Berners-Lee's second Linked Data principle. Unfortunately, not all data is converted into RDF and the approach suggests that the RDF-version of the data is static and must be updated periodically instead of converting the data to RDF on-the-fly (Koster, 2021). Therefore we might be dealing with outdated metadata.

The IISH has introduced separate URIs for the RDF-version of the data in their catalogue.[10] This means that the persistent web addresses that were introduced according to the Handles-principle do not resolve into a representation of the data in RDF-format. As a consequence, not all requirements to implement the second Linked Data principle are met.

The NHA has no RDF representation of their data whatsoever and does not meet the requirements of the second principle.

3.2 Retrieving data from online data repositories

I want to include the people that organized themselves in the typographical associations into my social historical research: what was their origin and what were their relations? To study the individual life courses of important participants in the 'typografische verenigingen', I use civil registries held by various archival institutes across The Netherlands. Fortunately, these registries are accessible through online indexes created by mostly volunteers and used by genealogical enthusiasts. On a website called Open Archives, Bob Coret has aggregated all the indexes that are published as online, open data.[11] On this platform Coret provides an Application Programming Interface

7 https://www.zotero.org/groups/2707622/typografische-verenigingen/library
8 For a thorough introduction into persistent web adresses see Koster (2020).
9 https://lod.uba.uva.nl/UB-UVA/Catalogue/
10 https://druid.datalegend.net/IISG/iisg-kg
11 https://www.openarch.nl/

(API) as well, which enables me to obtain the data of a particular registration.[12]

Every certificate that is included in this database has its own unique web address. This web address can be stored for a reference to the certificate on the Open Archives website. Using this web address in a web browser results in a human readable web page, but I can request the data in RDF as well. This means that both the first and second Linked Data principles are met.

Thankfully, the name of my research subject 'Jan Hendrik Regenboog' can also be found in the indexes on openarch.nl. Through his marriage certificate we know his age, profession and the names of his parents, wife, and parents-in-law.[13] I collect all the relevant references to important participants in the 'typografische verenigingen' in the civil registries in a tabular data file and retrieve the accompanying data in RDF. First I obtain a list of all civil registry certificates mentioning a person with one of the main occupations in the typographical domain: 'letterzetter', 'boekbinder', 'boekdrukker' and 'drukker'. Step two is to harvest all data of the certificates by resolving the URI of the certificate.

The names of properties and classes in the data are derived from the A2A standard, originally developed for exchanging personal data in XML. There is no ontology linked to the data to be found, however. This means that only the standardization of syntax and protocol of the data is established. The RDF representation refers to various elements in the data by means of internal links. Links to other sources, like places or archival institutions are not added. Therefore the third principle is not met.

The Linked Data representation of the data set 'History of Work' by the IISH enables us to create an external link with a clever trick.[14] By adding a base URI to the lexical title of an occupation in the source, e.g. by combining the string 'letterzetter' with the prefix 'https://iisg.amsterdam/resource/hsn/occupation/' into the URI `https://iisg.amsterdam/resource/hsn/occupation/letterzetter`, a link is created to the 'History of Work' dataset. Thanks to this link, I relate additional data about an occupation. This contains for instance the social status of the occupation in various standards, translations of the occupational title into other languages and the HISCO grouping (Zijdeman and Lambert, 2010).

3.3 Discussion

In this section I investigated the four data sources that are the most relevant for my research (see Table 1) and concluded that all of them made serious steps towards unique web addresses for reference to a resource. Three out of four provided data per resource in RDF.

Table 1: Overview of the described data sources and their compliance to the Linked Data principles

Data source	Principle 1	Principle 2	Principle 3
Int. Inst. of Social History	yes	yes, but	
Library Univ. of Amsterdam	yes	yes	
Noord-Hollands Archief	yes	no	
Openarch.nl	yes, but	yes	no

12 https://www.openarch.nl/api/docs/
13 https://www.openarch.nl/hga:F6619DCA-BE62-4A37-B09D-960CB8900393
14 https://druid.datalegend.net/HistoryOfWork

There are important issues with regard to the persistency of the web addresses. And these issues result in some extra 'buts' in the overview. Openarch.nl does not pretend to mint persistent web addresses for every resource it publishes. The web addresses are depending on the data provided by the archive holding the original data. If the identification at the archive changes, so does the web address openarch.nl provides.

The IISH provides two different web addresses for a resource: one persistent identifier according to the Handle system and one that enables providing data in RDF-format. Those two web addresses are technically unrelated.

An interesting question here that would move us beyond the scope of this paper, is whether more institutes in the Benelux and beyond use Linked Data to provide metadata about their collections. The Web of Data of heritage collections is growing fast and an overview given here would be outdated the moment it is published. At least national libraries of The Netherlands, Belgium, Great Britain and France have Linked Data facilities or plan to have them in the near future. In both The Netherlands and in Flanders, governments are stimulating the use of Linked Data in the heritage domain. This policy is executed by the Netwerk Digitaal Erfgoed[15] in The Netherlands and meemoo[16] in Flanders.

The observant reader notices I did not include the third Linked Data principle in my analysis. This third principle was only investigated for data I retrieved. Most of the sources I describe here, are only linked to my graph. As a reference management application, Zotero uses its own, locally stored metadata to create a reference in a paper. I chose to use the metadata of the objects conveniently stored in Zotero as the starting point for my graph rather than using the metadata that is provided in RDF via the URI of the object in the online catalogue. The alternative, i.e. using the metadata directly from the online catalogues, would have been a less practical choice. Furthermore, we have to take into account that the various catalogues each provide metadata according to different metadata standards. If I wanted to use that data, I would have needed to convert all the various types of linked data into my own model. A third argument for using the Zotero database as a starting point is that not all cultural heritage institutions — in our case for example the Noord-Hollands Archief — is providing Linked Data.

The curated data in Zotero must be transformed into Linked Data to be part of our web of data. The next section will study how this can be done, along with connecting data from other tools.

4 Creating Linked Data

Besides linking and retrieving the Linked Data available online, I create my own Linked Data for the graph of all relevant assets in my research. Ideally I want to create this data in convenient tools that I planned to use anyway, like the aforementioned Zotero, Tropy and Recogito. To some extent, these tools all provide the possibility to export and/or convert data into RDF. The results of these conversions will ultimately be combined into my graph.

Apart from the metadata about sources, my research needs data about entities like persons (e.g. the individual print labourers), organisations (e.g. 'typografische verenigingen') and songs (sung during their yearly feasts). This data is constructed in tabular form in a spreadsheet application and must be transformed into Linked Data.

15 https://www.netwerkdigitaalerfgoed.nl/
16 https://www.meemoo.be/

Finally, I use an application to derive additional data computationally. This constructed data must be available in the graph as well.

This section studies how easy I can construct the Linked Data and the links between data created in the different applications.

4.1 Zotero

Zotero is a reference management application. This means a scholar can create and manage records to create footnotes or end notes in their publications. Fortunately, you do not need to enter all data into Zotero, but the application has functionality to obtain the data from important catalogue websites, like worldcat.org, with the click of a button. After that you can curate the metadata to your liking. That way I collected data about all relevant sources for my research and I shared the result on the Zotero-website as a 'public library'.[17]

By publishing my collection of references through Zotero, a web address is available for every item. The registration of the yearbook 1856, for instance, can be found using `https://www.zotero.org/groups/2707622/items/P5HRJ669`. In my web of data I use this web address as a basic URI to relate all information I know about this object. It relates to manifestations of this book in the Library of Leiden University, Google Books, the Library of the University of Amsterdam and refers to data available on Worldcat.

I can use a so-called Zotero Translator, built in JavaScript, to export the data from the local Zotero storage. To create Linked Data I took an available Translator and adjusted it to my needs. The script exports the data with properties standardized on Schema.org[18] and uses the Zotero web address as a URI.

The possibility to create your own script to export the data collected in Zotero is very important. It complies with the idea of making the data more important than the tool. The available techniques meet the Linked Data principles halfway.

By creating the web address on the Zotero website for an item, I can use this as a URI for the item I want to register. On the other hand, the stability of the URI remains questionable. That way only some of the necessary aspects of the first Linked Data principle are met.

Concerning the second principle, the web address only refers to a human readable web page and is not able to deliver the underlying data in RDF. To mitigate this problem, I have created my own data set of data with the Translator. I need to publish this separately from the Zotero website.

According to the third principle I must use URIs to refer to other relevant resources. One of these is the holding institute of the item. To link items to their holding institute I use the list of International Standard Identifier for Libraries (ISIL) codes (for Standardization, 2019). The Dutch Royal Library[19] and the Dutch National Archive[20] are responsible for issuing these codes for institutes in The Netherlands. The IISH, for instance, has code 'NL-AsdIISG'. While entering data into Zotero I use the characters 'isil:' in front of an ISIL code, and my script creates a URI for this data point in the RDF export. A similar trick is used to refer to typographical associations. Here I use the characters 'typo:' which refers to a list of typographical associations I created in a spreadsheet.

---

17 `https://www.zotero.org/groups/2707622/typografische-verenigingen/library`

18 `http://schema.org/`

19 `https://www.bibliotheeknetwerk.nl/landelijke-digitale-infrastructuur-ldi/isil-codes`

20 `https://www.nationaalarchief.nl/archiveren/kennisbank/isil-codes`

Zotero helps to create web addresses for items, but the publication of Linked Data, both according to the second and third principles, is not incorporated. Nevertheless, creating Linked Data for my graph is possible thanks to the flexible export possibility and my own built-in conversion of links to other resources.

4.2 Tropy

In Tropy I can organize the pictures I take in the reading rooms of the institutes I visit. To do so, I just create a 'project', and combine pictures into 'items'. Of every item I can add some basic metadata, but the most important thing I register is the Zotero-URI I created. With this URI I can link the data in Tropy to the data in Zotero. One object I digitized is a booklet with songs, written on the occasion of the erection of the Coster statue in Haarlem in 1856 (Regt and Breeman, 1856). Different members from different associations submitted their text on existing melodies. Most prominent songwriter, with six songs, was J.H. Regenboog, boardmember of the association in The Hague.

Tropy has an export function, exporting the data directly into a Linked Data format, more specifically JSON-LD. Unfortunately, the export function does not construct URIs for the Items and therefore the data does not comply to the first Linked Data principle. I have to create an additional Python script to correct this. In the same script I must convert the property with my reference to the Zotero URI into a property expressing the semantics that the photos are a representation of the work described in Zotero.

4.3 Recogito

The third tool I use is Recogito. Recogito helps users to create textual or image sources with markings for entities in the text or on the image. The entities can be typed: you can state that the marking is a person, place or event. Entities for places can be reconciled to a standardized list of geographical names, like geonames.[21] In Recogito, I can add my own standardized list of relevant concepts, like important persons, or the typographical associations.

I created a machine-readable version of the yearbook of 1856 in Recogito and related this representation with the URI created by Zotero. In the text I marked the persons and associations that were mentioned. One of these persons is J.H. Regenboog, again as a board member of the typographical association in The Hague. He was awarded a silver medal in 1851 for his commitment to the association. From Recogito, this type of marking can be exported into a Linked Data file. Moreover, the text is hosted on the Recogito-website and the marking of J.H. Regenboog has a working web address.[22] I can use this as an anchor for linking other information to the mentioning of "J.H. Regenboog" in the yearbook.

Of all tools discussed in this paper so far, Recogito implements the Linked Data principles best. Documents and annotations all have their own URIs, and these URIs result in a presentation on the web and can be linked to other URIs for more information on the subject it links to. Some wishes remain, though. Unfortunately the link to the marking is lost if a new version of the text is uploaded, so I am unable to correct an error in the text without breaking the existing URIs. Besides that, the annotation ontology used in the Linked Data export is not used correctly (there is no such thing as a 'oa:Tag class') or incomplete ('oa:hasBody' is missing if the marking is not reconciled) (Sanderson et al., 2017).

---

21 https://www.geonames.org/

22 https://recogito.pelagios.org/annotation/51e9cac0-a07a-472d-9e9b-eef16f95995f

4.4 LDWizard and CoW

Some of the resources I need to link to in my graph are new and not available online. For example, I need resources for cultural heritage institutes, (for which I used the ISIL-codes, but which are not available as URIs online) and the typographical associations I study.

To create Linked Data for these resources, I have developed a CSV file with columns for each relevant property. For the typographical association that would for instance be the year the association was founded, or its location. To create Linked Data from this CSV file, I use the tool CoW, which stands for 'CSV on the Web'.[23] This tool, created in the CLARIAH program, converts the CSV into Linked Data. So the links I created in Zotero to express that an item is about a certain typographical association, refer to the URIs created in this process.

CoW needs a mapping that secures which Linked Data property must be used for each column. This mapping is stored in a JSON format, of which the creation by hand might be error prone. For that reason I used the LDWizard tool.[24] This tool provides a more user friendly environment to create such a mapping.

With my home brewed URIs, I am able to state that the yearbook of 1856 has information about the typographical association called "Door Eendracht t' Zaam Verbonden" in The Hague. While transforming my CSV into Linked Data I complied to the first Linked Data principle because I added URIs to the things (i.e. typographical associations) I want to describe. I did not create a web server that could resolve the URIs (yet) and do not deliver the data belonging to the things I described through this URI. So Berners-Lee's second principle is not met.

4.5 burgerLinker

Finding all family relations of Jan Hendrik Regenboog, — maybe scattered over the entire country — is very cumbersome. With help of the data retrieved from the openarch-website and a tool, developed in the CLARIAH program, called burgerLinker, I am able to find these family relations.[25]

For this I use the RDF data retrieved from openarch.nl. After harvesting the data I need to convert it into the semantics the burgerLinker tool needs. After that the transformed Linked Data is fed to the burgerLinker-tool, which finds links between mentions of the same person. burgerLinker is able to calculate family relations as well: persons with the same parents are (obviously) siblings.

In the resulting data set with family relations I'll find a brother of J.H. Regenboog, called Christiaan Regenboog. Christiaan was born in The Hague but moved to Amsterdam and apparently took the idea of founding a typographical association and his organizational skills with him. He became the co-founder and secretary of the typographical association 'Voorzorg en Genoegen' in Amsterdam. His name is mentioned in several publications on the early development of the organization of labourers. (Bos, 2001; Giele, 1972) Although in these publications Amsterdam is considered a hotspot in labour organization, the relation with The Hague should not be neglected. And the family relations might be a good source to map relations to other towns as well.

Creating this data set with family relations requires programming skills, firstly to obtain the right data from OpenArch, and then to convert it into the Linked Data

23 https://github.com/CLARIAH/CoW
24 https://ldwizard.netwerkdigitaalerfgoed.nl/
25 https://github.com/CLARIAH/burgerLinker

with the properties prescribed by burgerLinker. There are initiatives that try to create more user-friendly interfaces for this process, like the aforementioned LDWizard. Furthermore, the heritage institutes in The Netherlands have the ambition to provide Linked Data directly, which simplifies and abridges this programming phase. If this development takes off, we need to agree on standardized Linked Data classes and properties, and implement an automatic transformation into the data model in burgerLinker. Otherwise the researcher still needs programming skills to convert the data from the web.

Again there is no web server that could resolve the URIs I created and deliver the data belonging to the persons I described. So I do not comply to the second Linked Data principle. I am able, though, to link mentioned persons to their occupation in the History Of Work data set. Because I use a URI to do this I do comply to Berners-Lee's third Linked Data principle.

4.6 Discussion

Currently, the creation of Linked Data from the tools a researcher uses, depends on the ability to export data in a syntax complying to the RDF standard. The tool must help us create a URI for the resource that is modelled, enable the possibility to retrieve data via this URI, and encode relations by the use of URIs in other data. An overview of these three principles is presented in Table 2 for the tools where I needed to export data from. In Table 3 an overview is presented of the tools I used to created extra Linked Data.

Table 2: Overview of the described tools and the ability to comply to the Linked Data principles while exporting data

Tool	Principle 1	Principle 2	Principle 3
Zotero	yes, but	no	yes, but
Tropy	no	no	yes, but
Recogito	yes, but	no	yes

In the case of Zotero, coding skills are needed to construct RDF. With these skills one is able to construct a URI that results in a human readable web page, but the sustainability of the URI remains questionable. Tropy does not provide a URI at all. Therefore, I needed to write an extra script to create a URI to link to. In Recogito URIs are available and resolvable, but after updating the text in the system, the links are broken.

None of the three tools have the ability to provide data based on the URI. Finally, the links that are available in the data of resource depends heavily on the possibility to include the creation of these links in the scripting that is developed. Tools creating Linked Data from scratch are all able to comply to two out of three principles if the users walk the extra mile of using the tools with these principles in mind.

Table 3: Overview of the described tools and the ability to comply to the Linked Data principles while creating data

Tool	Principle 1	Principle 2	Principle 3
LDWizard and CoW	yes, but	no	yes, but
burgerLinker	yes, but	no	yes, but

5 Using the graph

At some point in time the data and source collection phase of the research project will be finished and a more or less complete graph is available. The moment my graph is complete, I will be able to validate the coherence of the data or select data from it and do an analysis. For that, I can upload the data into a Triple Store, an application specialized in storing and retrieving Linked Data. I could use the online CLARIAH Linked Data environment 'Druid'[26], which is an instantiation of TriplyDB[27] by Triply.[28] I could use a Triple Store on my laptop, like GraphDB,[29] with which I created Figure 2. Both systems include the option of using the query-language SPARQL to select a relevant subset for an analysis (DuCharme, 2013). The same SPARQL queries can be used in various programming environments.

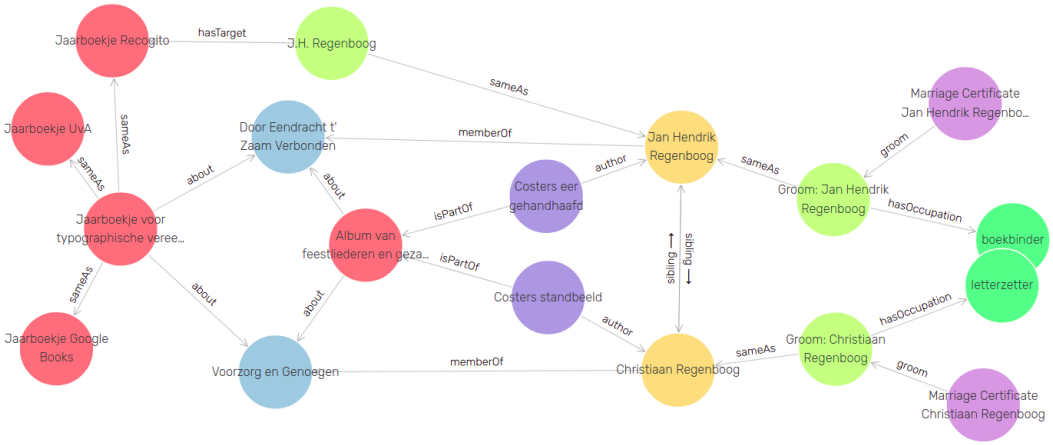


Figure 2: Part of the 'Web of Data' concerning Jan Hendrik and Christiaan Regenboog and their relations to the sources

6 Discussion

This paper demonstrates that creating Linked Data can, with some extra consideration and coding, be achieved with standard research tools. It therefore contributes to the FAIR principles by adding interoperability to the data.

The resulting 'web of data' can be considered a 'knowledge graph' about the subject of the research at hand. That way the knowledge about the subject is encoded

26 https://druid.datalegend.net/
27 https://triplydb.com/
28 https://triply.cc/
29 https://graphdb.ontotext.com/

in a computationally processable way and can be reused and extended by future researchers.

But why is that useful and necessary?

6.1 Reproducibility

If the underlying sources and literature of the arguments that support or contradict the hypotheses of my research are organized in Linked Data, this means that future researchers are able to validate the research. Future researchers can easily trace conclusions back to its original sources and evaluate the quality of the path in the argumentation. It results into reproducible research: scholars who use this research can retrieve references very quickly and check them. In addition, anyone can use the graphy I created (see Figure 2) to check which sources I used to conclude that J.H. Regenboog in The Hague was related to Ch. Regenboog in Amsterdam, important for the development of the phenomenon because it says something about why it was transferred from The Hague to Amsterdam.

6.2 Precision

For me, the main reason for creating this graph concerning the typographical associations is to investigate whether I can store all relations I find into one system, comprising of various applications. This forces me to be extremely precise about the relations between sources and conclusions I draw from them. Being extremely precise benefits the quality of the research.

6.3 Short term usefulness

The short term benefits of using Linked Data has to do with the data analysis that can be applied on a data set. In our example, the use of burgerLinker and the use of the links to HISCO enable me to add extra data to the graph already available.

6.4 Computer Aided Historical Research

A more visionary take on the Linked Data approach is that the future might bring a kind of Computer Aided Historical Research, instead of the current googling with search terms. A system might be able to retrieve more sources relevant to your research, thanks to the semantic relations added to the web of data by your colleagues in the past. We might develop a user interface were scholars are able to 'link up' their own subset of the web of data, and construct and publish new links, that lead to new suggestions in other research projects. If sources can be found and combined more quickly and more precisely thanks to this automation, the historian has more time to draw conclusions. I leave the dreaming about more possibilities in the more distant future to your own imagination.

7 Conclusion

In this paper I investigated the available Linked Data functionality in existing tools. I did this by creating a web of data with these tools.

Every specialized tool is good in its particular function: reference management, photo management, data conversion or source enrichment. For that reason I use this

combination of tools. The data they contain is one big data system, with links between things 'living' in different tools. For that reason I create one graph covering this data system. Creating the combined graph needed thorough thinking and additional coding, because the links between the data exported from the tools can not be easily made into the URIs that encode the link. Future development of this setup should take the decentralised data principle of Linked Data into account and enable fields to be filled with references to URIs. This, to comply to Berners-Lee third Linked Data principle: *anything else that that same thing has a relationship with through its data, should also be given a name beginning with HTTP*.

A big issue is: how do I publish my home brewed URIs as resolvable URIs, according to the second principle: *Looking up an HTTP name should return useful data about the thing in question in a standard format.* These URIs should be persistent and dereferenceable, but at the moment they are not. I need an institution to facilitate the creation of my own dereferenceable URIs quick and easy.

After this experiment, I conclude that big steps have already been taken towards implementing the Linked Data principles. It still needs further focusing on 'data' instead of the tool, however. At this moment some unnecessary programming is needed, so developers should take the Linked Data functions in their tools to the next level. The benefits and future development of Linked Data are important enough to do so.

References

Tim Berners-Lee. The next web, 2009. URL https://www.ted.com/talks/tim_berners_lee_the_next_web.

Dennis Bos. *Waarachtige volksvrienden: de vroege socialistische beweging in Amsterdam, 1848-1894*. Bakker, Amsterdam, 2001. ISBN 978-90-351-2315-1.

Richard Cyganiak, David Wood, and Markus Lanthaler. RDF 1.1 Concepts and Abstract Syntax, February 2014. URL https://www.w3.org/TR/rdf11-concepts/.

Bob DuCharme. *Learning SPARQL: Querying and Updating with SPARQL 1.1*. O'Reilly, 2013. ISBN 978-1-4493-7143-2.

International Organization for Standardization. ISO 15511:2019(en), Information and documentation — International standard identifier for libraries and related organizations (ISIL), 2019. URL https://www.iso.org/obp/ui/#iso:std:iso:15511:ed-4:v2:en.

Jacques Giele. Het ontstaan van de typografen-vakorganisatie in Nederland (1837-1869). *Mededelingenblad NVSG*, 42:2–55, 1972. URL https://hdl.handle.net/10622/FEC3F7D7-F7CB-4DC8-9C80-D8AFA0B7B9DF.

Lukas Koster. Persistent identifiers for heritage objects. *The Code4Lib Journal*, 47, February 2020. ISSN 1940-5758. URL https://journal.code4lib.org/articles/14978.

Lukas Koster. Infrastructure for heritage institutions – Open and Linked Data, June 2021. URL http://purl.org/cpl/3227.

Marcel van der Linden. *Social security mutualism: the comparative history of mutual benefit societies*. Peter Lang, Bern; New York, 1996. ISBN 978-3-906755-85-4.

C. Mommaas, editor. *Jaarboekje voor typographische vereenigingen*. A.W. Sythoff, 1856. URL https://pid.uba.uva.nl/ark:/88238/b19900035721202051 31.

J.K. de Regt and J. Breeman, editors. *Album van feestliederen en gezangen, te zingen door de typographische vereenigingen, die deel zullen nemen aan de onthullingsfeesten op den 16den Julij 1856 te Haarlem*. Brederode, Jacobus Johannes van Haarlem, Haarlem, 1856. URL https://hdl.handle.net/21.12102/5FB118A70935449D99D2EC40A029D267.

Robert Sanderson, Paolo Ciccarese, and Benjamin Young. Web Annotation Vocabulary, February 2017. URL https://www.w3.org/TR/annotation-vocab/.

Mark D. Wilkinson et. al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL https://www.nature.com/articles/sdata201618.

Richard Zijdeman and Paul Lambert. Measuring social structure in the past: A comparison of historical class schemes and occupational stratification scales on Dutch 19th and early 20th century data. *Journal of Belgian History/ Belgisch Tijdschrift voor Nieuwste Geschiedenis/ Revue Belge de Histoire Contemporaine*, 40(1-2):111–141, 2010. ISSN 0035-0869.