Digital Methods in Practice

The Epistemological Implications of Applying Text Re-Use Analysis to the Bloody Accounts of the Conquest of Jerusalem (1099)

by Torsten Hiltmann, Jan Keupp, Melanie Althage and Philipp Schneider*

While the question concerning the possible epistemological impact and implications of digital methods on the production of historical knowledge has been raised time and again, detailed studies in this regard have been scarce. Taking the analysis of the use of the Bible in the accounts on the Conquest of Jerusalem in 1099 as an example, this paper examines the differences between the analogue and the digital approach. Drawing on the method of text re-use analysis and the tool "Tracer," it demonstrates the application of digital methods in practice and shows the consequences this has for historical research.

Digital History is often confronted with high expectations.[1] For many representatives of our discipline, it has yet to prove itself. The question arises of what new insights and groundbreaking results have actually been achieved by the digital approach so far that could justify all the effort and investment involved. After all, computer-based methods are not new to historical

* The research data generated in the project are published as open data under http://doi.org/10.5281/zenodo.4719838. We would like to thank Dr. David West for his final review and correction of the manuscript. All remaining errors are our own.

1 Digital History, in this sense, is much like the Digital Humanities, whose problems it shares; see, e. g., Tom Scheinfeldt, Where's the Beef? Does Digital Humanities Have to Answer Questions?, in: Matthew K. Gold (ed.), Debates in the Digital Humanities, Minneapolis, MN 2012, https://dhdebates.gc.cuny.edu/read/untitled-88c11800–9446–469b-a3be-3fdb36bfbd1e/section/3c03ecdb-2dcf-4597–8fc4-e42f8dcc21e1#p1b2; Gold and Lauren F. Klein, Introduction. Digital Humanities: The Expanded Field, in: Gold and Klein (eds.), Debates in the Digital Humanities 2016, Minneapolis, MN 2016, pp. ix – xv; Cameron Blevins, Digital History's Perpetual Future Tense, in: ibid., pp. 308 – 324. Among those debates, an article in the *The Chronicle of Higher Education* in October 2017 was the cause of intense discussion in the Digital Humanities community. Timothy Brennan struck a nerve when he claimed that Digital Humanities had produced little of substance especially in text analysis; see Timothy Brennan, The Digital Humanities Bust, in: The Chronicle of Higher Education, 15. 10. 2017, https://www.chronicle.com/article/The-Digital-Humanities-Bust/241424. For an overview of the debate, see Monika Barget, "The Real Problem with Digital Humanities." Critical Approaches for Further Discussion, in: Revolts as Communication, 15. 3. 2018, https://revolt.hypotheses.org/1848.

Geschichte und Gesellschaft 47. 2021, S. 122 – 156

ISSN (Printausgabe): 0340-613X, ISSN (online): 2196-9000

https://doi.org/10.13109/gege.2021.47.1.122

research; in fact, they have been around since the late 1960s.[2] Still, they have never really established a foothold in mainstream research.

Maybe these expectations that the digital first has to prove its worth through groundbreaking results before it is applied on a larger scale may also result from a certain hesitation towards the digital as such. In contrast to other methodological renewals in humanities research such as the spatial, the iconic or the material turn, which tended to emerge from within the humanities themselves,[3] the origins of those digital approaches apparently come from the outside, bringing a quite different kind of thinking and way of operating into historical research. Since it draws on methods taken from computer sciences, it seems much further away from the established hermeneutical approaches of the humanities. The hurdle to applying it in one's own work seems to be much higher, combined with the need to acquire completely new skills and competencies. However, at the same time, these new digital methods are perceived as being only an alternative way of doing the same things as before and are evaluated in this sense. Only when it is proven that their application can produce significant results may they be considered part of the methodological toolbox that historians consider useful for their craft.[4]

In order to understand digital approaches and their performance better, we first need to understand the process of digitalization and datafication as such. This means understanding what happens to our texts when they are digitized and how exactly these new digital methods affect how we produce new insights based on these digitized texts. While the impact of digitalization and the use of data-driven methods on the epistemology of historical research have been

2 For early contributions, see, among others, Jerome M. Clubb and Howard Allen, Computers and Historical Studies, in: Journal of American History 54. 1967, pp. 599 – 607, https://doi.org/10.2307/2937409; Vern L. Bullough, The Computer and the Historian. Some Tentative Beginnings, in: Computers and the Humanities 1. 1967, pp. 61 – 64, https://doi.org/10.1007/BF00119888; Carl August Lückerath, Prolegomena zur elektronischen Datenverarbeitung im Bereich der Geschichtswissenschaft, in: HZ 1968, H. 207, pp. 265 – 296; Jean Schneider, La machine et l'histoire. De l'emploi des moyens mécaniques et électroniques dans la recherche historique. XIIIe Congrès international des sciences historiques, Moscou, 16 – 23 Août, 1970, Moscow 1970; Robert P. Swierenga, Clio and Computers. A Survey of Computerized Research in History, in: Computers and the Humanities 5. 1970, pp. 1 – 21, https://doi.org/10.1007/BF02404252; Klaus Arnold, Geschichtswissenschaft und elektronische Datenverarbeitung. Methoden, Ergebnisse und Möglichkeiten einer neuen Hilfswissenschaft, in: Theodor Schieder (ed.), Methodenprobleme der Geschichtswissenschaft, München 1974, pp. 98 – 148.

3 For an overview of the various "turns" that have shaped the humanities and social sciences, see Doris Bachmann-Medick, Cultural Turns. New Orientations in the Study of Culture, Berlin 2016.

4 On these claims and kinds of critique, see Scheinfeldt, What's the Beef?; Blevins, Digital History's Perpetual Future Tense; Malte Rehbein, "L'historien de demain sera programmeur ou il ne sera pas." (Digitale) Geschichtswissenschaften heute und morgen, in: Digital Classics Online 4. 2018, no. 1, pp. 23 – 43, esp. p. 33, https://doi.org/10.11588/dco.2017.0.48491.

repeatedly mentioned, asserted or doubted, systematic investigations have so far been scarce.[5] In our opinion, digital methods are not just another "turn," as it is sometimes put.[6] They are not based on a new theory or research paradigm, but, at least in their current form, on a fundamental media change.[7] Thus, data-driven research follows a completely different logic than our established approaches to historical research, thereby altering the framework of historical scholarship as a whole.

How these methods differ from the way we are used to proceeding in historical research and what they actually imply is what we explore in the following case study. We will do so by using the example of a method called text re-use analysis, which helps to identify intertextual connections such as quotations or allusions in different texts.[8] What we present here is a first workshop report that does not intend to solve a historical question or an old debate. Rather, we want to use this case to understand better the application of digital methods, their specific operating conditions, and subsequent epistemological implications. In doing so, we will demonstrate the methods used step by step in order to be able to make the differing fundamentals as well as the differing methodological questions explicit. For this we will draw on the chronicles of the First Crusade (1096 to 1099) and how they describe the conquest of Jerusalem in 1099.

5 For an overview, see the collective review of current publications on Digital History by Mareike König, Die digitale Transformation als reflexiver *turn*. Einführende Literatur zur digitalen Geschichte im Überblick, in: Neue Politische Literatur 66. 2021, pp. 37 – 60, https://doi.org/10.1007/s42520–020–00322–2. See also Philippe Rygiel, Historien à l'âge numérique, Lyon 2017, https://books.openedition.org/pressesenssib/6303, as well as the recent white papers by the Roy Rosenzweig Center: Arguing with Digital History Working Group, Digital History and Argument, White Paper, Roy Rosenzweig Center for History and New Media, 13. 11. 2017, https://rrchnm.org/argument-white-paper/, and the German Association for Digital Humanities, Thesen. Digital Humanities 2020, https://dig-hum.de/thesen-digital-humanities-2020.

6 Bob Nicholson, The Digital Turn. Exploring the Methodological Possibilities of Digital Newspaper Archives, in: Media History 19. 2013, pp. 59 – 73; Wolfgang Schmale, Digitale Vernunft, in: Historische Mitteilungen 26. 2013, pp. 94 – 101, p. 94; Bachmann-Medick, Cultural Turns, p. 289; or recently also C. Annemieke Romein et al., State of the Field. Digital History, in: History 105. 2020, pp. 291 – 312, here p. 292.

7 Torsten Hiltmann addressed this issue in his opening keynote at the conference "Digital History. Konzepte, Methoden und Kritiken digitaler Geschichtswissenschaften" (1 – 3 March 2021, Göttingen), see Torsten Hiltmann, Abstract: Vom Medienwandel zum Methodenwandel. Die Digitalisierung der Geschichtswissenschaften in (medien-)historischer Perspektive, in: Digital History. Konzepte, Methoden und Kritiken digitaler Geschichtswissenschaften, 27. 12. 2019, https://digitalhist.hypotheses.org/1035. A publication of the paper is in preparation.

8 See Marco Büchler et al., Towards a Historical Text Re-Use Detection, in: Chris Biemann and Alexander Mehler (eds.), Text Mining. From Ontology Learning to Automated Text Processing Applications, Cham 2014, pp. 221 – 238.

I. The Conquest of Jerusalem in Crusade Chronicles

"Si verum dicimus, fidem excedimus"[9] ("If we tell the truth about it, we exceed the limits of what is credible") – this is how the chronicler and eyewitness Raymond d'Aguilers comments on the events after the capture of Jerusalem on 15 July 1099. Referring to the immense bloodshed that the troops of the first crusade had caused in the streets and houses of the city, he seemed to have reached the limits of verbal expression. Nevertheless, he continued his remarks with the words: "So it is sufficient to relate that in the Temple of Solomon and the portico crusaders rode in blood to the knees and bridles of their horses."[10] This exalted verbal image attracted the attention of later generations in a special way and led to different interpretations.[11] Were Raymond's words "évidemment une hyperbole, et prouvent que les historiens latins exagéraient les choses qu'ils auraient dû atténuer ou cacher," as the French Crusade historian Joseph François Michaud conjectured at the beginning of the nineteenth century?[12] This remark is undoubtedly based on standards of value that were closer to the thinking of the Enlightenment than to the horizon of understanding of contemporaries in the eleventh century.[13] But, on the 900th anniversary of the events, Michaud's thesis underwent an astonishing revival:[14]

9 John Hugh Hill and Laurita Lyttleton Hill (eds.), Le "Liber" de Raymond d'Aguilers, Paris 1969, p. 150.

10 Raymond d'Aguilers, Historia Francorum Qui Ceperunt Iherusalem, translated, introduced and annotated by John Hugh Hill and Laurita Lyttleton Hill, Philadelphia 1968, pp. 127 f.; Hill and Hill, Le "Liber," p. 150: "Si verum dicimus, fidem excedimus. Sed tantum sufficiat, quod in templo et porticu Salomonis equitabatur in sanguine ad genua, et usque ad frenos equorum."

11 A brief overview is provided by David Crispin, "Ihr Gott kämpft jeden Tag für sie." Krieg, Gewalt und religiöse Vorstellungen in der Frühzeit der Kreuzzüge (1095 – 1187), Paderborn 2019, pp. 16 – 18. Raymond's statement has recently led to an astonishing attempt to falsify the accuracy of the statement by exact calculations, cf. Thomas F. Madden, Rivers of Blood. An Analysis of One Aspect of the Crusader Conquest of Jerusalem in 1099, in: Revista Chilena de Estudios Medievales 1. 2012, pp. 25 – 37, here p. 35: "Therefore, in order to fill al-Aqsa Mosque's square meters to ankle level would require the blood of 92,960 people." The motivation that leads serious researchers to engage in such number games is beyond the reach of the authors.

12 Joseph-François Michaud, Histoire de croisades, vol. 1, Paris 1812, p. 411.

13 All the more remarkable: Jean Flori, Chroniqueurs et propagandistes. Introduction critique aux sources de la première croisade, Geneva 2010, p. 293: "se laisse aller à l'hyperbole en écrivant."

14 Alongside the reported position of Kaspar Elm, Die Eroberung Jerusalems im Jahre 1099. Ihre Darstellung, Beurteilung und Deutung in den Quellen zur Geschichte des Ersten Kreuzzugs, in: Dieter R. Bauer et al. (eds.), Jerusalem im Hoch- und Spätmittelalter. Konflikte und Konfliktbewältigung – Vorstellungen und Vergegenwärtigungen, Frankfurt 2001, pp. 31 – 54, a similar stance can be observed at the turn of the millennium for David Hay, Gender Bias and Religious Intolerance in Accounts of the "Massacres" of the First Crusade, in: Michael Gervers and James M. Powell (eds.), Tolerance and Intolerance. Social Conflict in the Age of the Crusades, Syracuse 2001, pp. 3 – 10. Kristin Skottki, Christen, Muslime und der Erste Kreuzzug. Die Macht der

"Why do both the chroniclers and the artists exaggerate, why do they, who were Christians after all, revel in the depiction of cruelties, murder and slaughter without explicitly and unmistakably stating regret, compassion or even shame, as one would expect?" asked the renowned medievalist Kaspar Elm.[15] The accusation of literary exaggeration of the events is based on a syllogistic conclusion: since the acts of violence during the capture of the Holy City allegedly corresponded to the customary wartime practice of the time, whereas the chroniclers painted them in the brightest colours, there must obviously be an exaggeration here. Although this conclusion must appear problematic both in its premise and in its partial circularity,[16] Elm makes it the starting point for a further attempt at exoneration:

> The basic argumentation here is that the authors concerned deliberately exaggerated the actual extent of the bloodshed perpetrated by the crusaders by drawing on biblical verbal images, precisely in order to place retrospectively the success of the campaign in analogy with biblical events.[17]

Not only the military measures, but also their historiographical presentation therefore allegedly follows old conventions. The verbal images used by the chroniclers admittedly originate from a different cultural sphere and must therefore not be confused with the medieval reality. Reference is made here to the descriptions of Roman punitive measures in Judea in the historical work of Flavius Josephus, but above all to the adaptation of the Old Testament and apocalyptic motifs for violence.[18] Since foreign verbal images are used, consequently there could be "no question here of personal observation."[19] The report is thus exposed as a literary collage in which empirical knowledge recedes behind what is borrowed. Raymond d'Aguilers, as quoted at the beginning, is suddenly transformed from an authentic observer of the events to a principal witness for Elm's fabrication-thesis, since his statement about

Beschreibung in der mittelalterlichen und modernen Historiographie, Münster 2015, pp. 117 f., sees this "Verharmlosung" as a "Reaktion auf christentumskritische Positionen." This covers only part of the issue, the virulent discourses on the relationship of the "West" to the Islamic world would have to be taken into account, cf. Michael Evans, Explaining or Excusing? The Crusades, Historical Objectivity, and the "War on Terror," in: Gwendolyn Morgan (ed.), The Year's Work in Medievalism XXI / XXII, 2005 and 2006, Eugene 2007, pp. 108 – 116.

15 Elm, Eroberung Jerusalems, p. 46 [all transl. by T. H. et al.].

16 John France, Victory in the East. A Military History of the First Crusade, Cambridge 1994, pp. 355 f., briefly presented a view that has been firmly rejected in research, cf. Benjamin Z. Kedar, The Jerusalem Massacre of July 1099 in the Western Historiography of the Crusades, in: Crusades 3. 2004, pp. 15 – 75, here pp. 67 f. Moreover, it stands in stark contrast to the then current research on high medieval conflict management, cf. Gerd Althoff, Spielregeln der Politik im Mittelalter. Kommunikation in Frieden und Fehde, Darmstadt 1997; the argumentation is circular because it draws its verdict of exaggeration from the identical sources on which France bases his findings.

17 Crispin, Krieg, Gewalt und religiöse Vorstellungen, p. 107.

18 Elm, Eroberung Jerusalems, pp. 49 – 52.

19 Ibid., p. 50.

the blood of the slain reaching to the reins of the horses ("usque ad frenos equorum") could be proven as a literal borrowing from the Latin text of the Revelation of John (14:20).[20]

The essentially apologetic position was positively received and expanded by parts of the research community, especially in the German-speaking area: drawing on Elm's statements, church historian Arnold Angenendt, for example, once again accentuated the contrast between supposedly conventional wartime routine and its historiographical processing, which had used a special "Blutsprache" in a deliberate pictorial exaggeration: "The conquest of Jerusalem, which was carried out in the usual cruelty of the period, is exaggerated along blood-religious lines."[21]

It may be registered as alarming that such "attenuations, explanations and apologies" of the mass killings of 1099 are in their quintessence meanwhile to a large extent "prevailing doctrine […], which are part of handbook knowledge."[22] On the other hand, one will have to point out the vehement opposition that Elm's theses have provoked in other parts of the research community. Not only was it repeatedly emphasized by way of convincing arguments that the acts of violence during and after the storming of Jerusalem "must surely have exceeded the usual extent in war."[23] In addition, the formula of the "empty topos" on which Elm's argument for exoneration is based, served as a point of attack. According to Gerd Althoff, it has often had "its merits in downplaying the event," but has been refuted by scholars in many cases.[24] Instead, the recourse to the authority of the biblical texts is to be understood precisely as a means of accentuation and authentication that further increases the claim to

20 This discovery goes back to the edition of John Hugh Hill and Laurita L. Hill; see Hill and Hill, Le "Liber," p. 150; respectively John France, A Critical Edition of the Historia Francorum Qui Ceperunt Iherusalem of Raymond of Aguilers, PhD thesis University of Nottingham 1967, http://eprints.nottingham.ac.uk/50918/, p. xcix.

21 Arnold Angenendt, Toleranz und Gewalt. Das Christentum zwischen Bibel und Schwert, Münster 2008[4], pp. 424 – 427; Angenendt, Die Kreuzzüge. Aufruf zum "gerechten" oder zum "heiligen" Krieg?, in: Andreas Holzem (ed.), Krieg und Christentum. Religiöse Gewalttheorien in der Kriegserfahrung des Westens, Paderborn 2009, pp. 341 – 368, here p. 358.

22 Gerd Althoff, Papst Urban II. und das Massaker von Jerusalem. Zur Legitimation der Gewalt gegen "Ungläubige," in: Regina Grundmann and Assaad Elias Kattan (eds.), Jenseits der Tradition? Tradition und Traditionskritik in Judentum, Christentum und Islam, Berlin 2015, pp. 129 – 152, here p. 135. Here reference is made to the restrained adaptation in Nikolas Jaspert, Die Kreuzzüge, Darmstadt 2010[5], p. 42.

23 Peter Thorau, Die Kreuzzüge, München 2007[3], p. 73.

24 Althoff, Papst Urban II., p. 134. This tendency is also objected to in clear words by Alan V. Murray, The Demographics of Urban Space in Crusade-Period Jerusalem (1099 – 1187), in: Albrecht Classen (ed.), Urban Space in the Middle Ages and the Early Modern Age, Berlin 2009, pp. 205 – 224, here p. 210: "slaughter remains slaughter, even if it is described in apocalyptic terms." Cf. also Hans-Henning Kortüm, Krieg im Mittelalter. Der Blick auf die Kinder, in: Alexander Denzler et al. (eds.), Kinder und Krieg. Von der Antike bis zur Gegenwart, Berlin 2016, pp. 201 – 218, p. 213.

truth of what is described. According to Philip Buc, one would possibly fall short if one were to attribute only the "verbal excesses," but not also the "actual and factual violent deeds," of the crusaders to the biblical motifs.[25] For Benjamin Z. Kedar, too, the evidence of an action in older testimonies does not seem to be suitable at all for a general denial of the reference of a source to reality.[26] As an example, he deals with Elm's remark that the "practice of smashing the heads of infants […] can be traced back to no one other than Flavius Josephus and his account of the atrocities committed at the hands of the legionaries of Titus and Vespasian."[27] Not only would Psalm 137:9 be a biblical reference point.[28] Source criticism must also seek to understand each testimony in its own temporal context: "Bringing Josephus Flavius into the discussion, then, tends to obscure rather than enhance our understanding of the events at hand."[29]

However, Kedar overlooks the fact that the parallel of the text evoked by Elm is neither referenced in the annotations nor found in the work of the ancient Jewish historian. The same applies to the alleged "other images from the apocalypse" in the "Historia Francorum" of Raymond d'Aguilers – here, too, the author lacks concrete proof.[30] Even if one considers it to be the current consensus of scholars that "Elm is right in assuming that the chroniclers may have used literary models,"[31] views on which passages of text actually formed the basis of the descriptions of Jerusalem's capture clearly differ.

Arnold Angenendt, for example, accentuates the relationship between the chronicles of the crusades and the "Blutsprache of the books of Maccabees."[32] With their meticulous depiction of violence, the chroniclers had endeavored

25 Philippe Buc, Holy War, Martyrdom, and Terror. Christianity, Violence, and the West, Philadelphia 2015, p. 271.

26 Kedar, Jerusalem Massacre, p. 72: "Yet the existence of a description in earlier literature surely does not preclude the possibility that its recurrence is based on actual observation."

27 Elm, Eroberung Jerusalems, p. 51.

28 Kedar, Jerusalem Massacre, p. 72. Cf. on the passage Albert of Aachen, Historia Ierosolimitana. History of the Journey to Jerusalem, ed. by Susan B. Edgington, Oxford 2007, VI 23, p. 432. Cf., however, Christian Hofreiter, Making Sense of Old Testament Genocide. Christian Interpretation of *Herem* Passages, Oxford 2018, p. 178: "It should also be noted that there are no exact verbal parallels between Albert's description and the Vulgate's wording of the psalm." Nahum 3:10 could also be considered. See also Sini Kangas, The Slaughter of the Innocents and the Depiction of Children in Twelfth- and Thirteenth-Century Sources of the Crusades, in: Elizabeth Lapina and Nicholas Morton (eds.), The Uses of the Bible in Crusader Sources, Leiden 2017, pp. 74 – 101, here p. 92.

29 Kedar, Jerusalem Massacre, p. 72. When Kedar refers in this respect to "the very many accounts about Germans killing Jewish infants in this way during World War II," he undoubtedly provides a serious argument, cf., e. g., Saul Friedländer, Das Dritte Reich und die Juden. Die Jahre der Verfolgung 1933 – 1939. Die Jahre der Vernichtung 1939 – 1945, München 2007, p. 700.

30 Elm, Eroberung Jerusalems, p. 51.

31 Kedar, Jerusalem Massacre, p. 71.

32 Angenendt, Die Kreuzzüge, p. 354.

"to evoke Old Testament parallels of the Maccabean wars, even to surpass them."[33] Such references can indeed be explicitly proven in the chronicles of the first crusade in a prominent place – admittedly not in connection with the capture of Jerusalem, but for instance in the prologue of Fulcher of Chartres, on the occasion of a battle before Antioch by Raymond d'Aguilers or in the sermon of Pope Urban II, in Clermont by Guibert of Nogent.[34] In the alternative version of the crusade call from the pen of Balderic of Dol, however, there is also a reference to Psalm 79, which Gerd Althoff places at the centre of his interpretation of the massacre.[35] The proximity to the Old Testament motif of the ban *(herem),* specifically the fall of the city of Jericho, postulated several times,[36] has been recently subjected to an extensive analysis by Christian Hofreiter.[37] His sobering conclusion: apart from an explanatory addition in the account of Bartolf de Nangis, reworking Fulcher here, "no other direct reference is made to any of the OT extermination commands or narratives,"[38] and also the retrospective justification of this chronicler – that the Christian warriors had shied away from milder measures of violence for fear of divine wrath – does "not provide evidence that any of the crusaders were in fact motivated by this (or any other) *herem* text."[39]

Against this background, it seems rather daring when Luigi Russo, in his 2017 study on the historiography of the massacre, emphasizes time and again the continuous borrowing of expressions and words from the holy text by the chroniclers, "strictly bound to the Bible."[40] The Holy Scriptures are to be understood "as a key text, helping chroniclers to rethink all of the events leading up to and including the conquest of the Holy City."[41] The first two test cases that he has selected produce a clear result: "Neither the anonymous author of the *Gesta* nor Peter Tudebode employ the Bible to convey a new meaning to the events."[42] As concrete evidence of a reference to the Bible by

33 Angenendt, Toleranz, p. 426.
34 Cf. Christoph Auffarth, Die Makkabäer als Modell für die Kreuzfahrer. Usurpationen und Brüche in der Tradition eines jüdischen Heiligenideals. Ein religionswissenschaftlicher Versuch zur Kreuzzugseschatologie, in: Christoph Elsas et al. (eds.), Tradition und Translation. Zum Problem der interkulturellen Übersetzbarkeit religiöser Phänomene, Berlin 1994, pp. 362 – 390; Elizabeth Lapina, The Maccabees and the Battle of Antioch, in: Gabriela Signori (ed.), Dying for the Faith, Killing for the Faith. Old-Testament Faith Warriors (Maccabees 1 and 2) in Cultural Perspective, Leiden 2012, pp. 147 – 159.
35 Althoff, Papst Urban II., p. 136.
36 See Jaspert, Kreuzzüge, p. 42; Kedar, Jerusalem Massacre, p. 71.
37 Hofreiter, Making Sense, pp. 176 – 183.
38 Ibid., p. 181.
39 Ibid., p. 180. The fact that the author strongly accentuates these few findings seems accordingly irritating.
40 Luigi Russo, The Sack of Jerusalem in 1099 and Crusader Violence Viewed by Contemporary Chroniclers, in: Lapina and Morton, Uses of the Bible in Crusader Sources, pp. 63 – 73, here p. 71 and p. 73.
41 Ibid., p. 71.
42 Ibid., p. 68.

contemporary reporters of the capture of the city in 1099, he can once again cite only the testimony of Raymond d'Aguilers. It was perhaps on the basis of such observations that Jean Flori formulated the justified question whether the chroniclers would not have suppressed rather than accentuated the eschatological interpretation of the events in retrospect.[43]
These findings are not intended to create a false impression: the Sacred Scripture is an important point of reference for Crusade chroniclers. The question is, though, in which form and to what extent? The exemplary look at the research on the fall of Jerusalem in 1099 has revealed that on the one hand, the discovery of biblical phrases and references among the chroniclers of the First Crusade has an enormous influence on the scholarly reading of these texts. But, on the other, researchers have so far failed to systematically compare the texts of the chronicles with the books of the Bible and thus to thoroughly prove their assertions. So there is a strong need to know more about how exactly the Bible was used in these texts. Investigating these textual links, however, entails great challenges.

II. Different Ways to Approach the Identification of the Re-Use of the Bible

One of the biggest challenges lies in the subject itself, as has often been pointed out in research: the linguistic reference to the biblical text is often "fairly oblique" and "not specific enough for us to be certain that a deliberate allusion was intended."[44] This begs the question of how this has been addressed in research to date and how digital methods can be used to answer the questions raised above.

1. Analogous Approach

So if we were to examine systematically the use of the Bible in the chronicles of the first Crusades, how would we usually proceed? As we have seen, current research has subjected only a part of the chronicles to closer examination,[45] or the analysis has been limited to certain segments or a specific theme of the Scriptures.[46] To extend this endeavor to a detailed comparison of the entire Bible with one or several chronicles would be very demanding. In order to be able to recognize the various uses of Scripture, one must be very well versed in

43 Jean Flori, Pierre l'Ermite et la première croisade, Paris 1999, pp. 419 – 423, here p. 420: "Il me semble, pour ma part, que l'interprétation ultérieure des chroniqueurs ne va pas dans le sens eschatologique. Elle s'en éloigne au contraire et cherche à l'évacuer."
44 Thomas Lecaque, Reading Raymond. The Bible of Le Puy, the Cathedral Library and the Literary Background of the Liber of Raymond of Aguilers, in: Lapina and Morton, Uses of the Bible, pp. 105 – 132, here p. 115; Hofreiter, Making Sense, p. 177.
45 See, e. g., Russo, The Sack of Jerusalem.
46 Hofreiter, Making Sense, pp. 176 – 183.

the Bible. Robert B.C. Huygens, for instance, expressed his high expectations with regard to editors who are dealing with texts that might have cited the Scripture as follows:

> And if you do not know the Bible, don't edit any text at all without first having thoroughly familiarized yourself with the typical flavour of biblical language. If you only know the Bible in your own language, start reading it more than once in Latin.[47]

This is only possible with an appropriate background and sufficient time, something that certainly not everyone can acquire for this purpose, and especially when it is only a matter of solving one of many research questions.

Of course, one could also limit one's annotations to those passages where the medieval author himself indicated his use of the Scriptures, as was done in the still authoritative text edition of the "Historia Hierosolymitana" by Fulcher of Chartres, which was procured by Heinrich Hagenmeyer: Bible passages are primarily indicated here where the author himself explicitly marked his borrowings.[48]

In order to recognize also those uses of the Scripture which have not been indicated by the author himself, one has to rely on one's own experience and instinct. The performance of a scholar depends here on how acquainted he or she is with high-medieval chronicles, the Scriptures and their use within these kinds of texts, and whether he or she is able to identify passages in the texts that do not quite fit and could thus be related to another source or even to the Holy Bible. The scholar's gut feeling plays a particularly important role here.

We find this confirmed when we look at the references to the biblical passages in other editions of the chronicles. Often, the direct connection with the given passage in the Bible is not immediately evident. Especially since in those scholarly editions there is usually no explanation as to why a certain passage was chosen as a reference to the Scriptures and not another similar one. In John and Laurita Hill's 1969 edition of Raymond d'Aguiler's "Liber," for instance, the editors give for the passage on the conquest of the city and Raymond's subsequent commentary four different Bible references, and their translation of the edition adds a fifth.[49] If we take a closer look, this fifth reference is quite vague since it seems to be based only on the term "mirabilia."[50] In the Latin original, this passage reads: "Sed cum iam nostri menibus potirentur civitatis et turribus, tunc videres mirabilia."[51]

47 Robert B.C. Huygens, Ars Edendi. A Practical Introduction to Editing Medieval Latin Texts, Turnhout 2000, p. 11.
48 Cf. Fulcheri Carnotensis Historia Hierosolymitana (1095 – 1127), ed. by Heinrich Hagenmeyer, Heidelberg 1913.
49 Cf. Hill and Hill, Le "Liber," pp. 150 – 152.
50 D'Aguilers, Historia Francorum, p. 127, note 21.
51 Hill and Hill, Le "Liber," p. 150.

The translators explain in their footnote that Raymond draws the beginning of his description of "the marvelous works" done during the conquest "from Psalms 25:7; 39:6." Psalm 25:7 reads "ut clara voce praedicem laudem et narrem omnia mirabilia tua," while Psalm 39:6 says: "multa fecisti tu Domine Deus meus mirabilia tua et cogitationes tuas pro nobis non invenio ordinem coram te si narrare voluero et numerare plura sunt quam ut narrari queant."

One could certainly claim that this reference is highly speculative and more an interpretation of Raymond's intention than an actual quote or intended allusion. The starting-point for that presumed reference seems to be, in particular, the word "mirabilia." However, this seems rather vague, as the term "mirabilia" appears 62 times in the version of the Vulgata used in the context of this paper.

To give another example taken directly from the edition of the Latin text, describing the rejoicing after the capture of the Holy City: for the "nova dies" in the phrase "Nova dies, novum gaudiums, nova et perpetua leticia laboris atquet devotionis consummatio, nova verba nova cantica, ab universis exigebat"[52] the authors refer to:

(i) Isaiah 65:17 f.: "ecce enim ego creo caelos novos et terram novam et non erunt in memoria priora et non ascendent super cor sed gaudebitis et exultabitis usque in sempiternum in his quae ego creo quia ecce ego creo Hierusalem exultationem et populum eius gaudium,"

(ii) 2 Peter 3:13: "novos vero caelos et novam terram et promissa ipsius expectamus in quibus iustitia habitat,"

(iii) Apocalypse, 21:5: "et dixit qui sedebat in throno ecce nova facio omnia et dicit scribe quia haec verba fidelissima sunt et vera."

The very fact that, for these three words, a total of three different references have been given instead of a single one may underline the observation that these references are offered more associatively than in order to indicate a clear re-use of a biblical quotation. And it certainly would also have been possible to quote other passages from the Scriptures here. What we can see here is already an interpretation of Raymond's text.

The Bible references given in this edition appear to a certain extent vague and arbitrary. They clearly do not reflect the full range of possible or of exact references.[53] The same can be shown for the references given in another

52 Ibid., p. 151.

53 For example, the verbatim parallel between Raymond's "per vicos et plateas" (ibid.) and Song of Songs 3,2 has not been identified – perhaps because the search of the loving bride hardly seems to be in harmony with the place where severed body parts are deposited. For further possible quotes from the Scriptures, see James F. Brundage's review of the translation in 1969, which cites other potential references: James E. Brundage, Review of: Raimond d'Aguilers, Historia Francorum qui ceperunt Iherusalem, transl. with an introduction and notes by John Hugh Hill and Laurita L. Hill, in: Speculum 44. 1969, pp. 541 f.

scholarly edition of Raymond's chronicle by John France.[54] However, they can give us some clues as to how the detection of potential Bible quotations works here: namely, on the level of ideas and concepts. In order to recognize a quote from or allusion to the Bible, a scholar works in most cases on the level of general meaning, as he or she remembers the text or its content. The exact words of a quote, one would suppose, play rather a minor role here. If a possible quote has been found, the scholar would start to compare the passages word-for-word. Only then would she or he try to evaluate and categorize the reference. The decision as to why he or she thinks that this is indeed an (intentional) quote from the Bible is hardly ever disclosed. This also means that not every decision may be made according to the same set of criteria. This is important if we want to compare the analogue with the digital method.

2. Digital Approach with Text Re-Use Analysis

Prerequisites in Approaching Text Digitally

While the traditional approach is characterized by the fact that the contextual knowledge directly flows into the source work and grows with it, it has the disadvantage of being based mostly on intuitive selection. In this respect, the computer is able to do something that humans cannot: namely, systematically compare two large volumes of text and display their similarities. What may sound simple does in fact entail a number of challenges, especially when working with historical data.

The fundamental difference is that computers cannot "understand" texts or interpret them in terms of their meaning: the machine is "semantically blind," as Silke Schwandt puts it.[55] Instead, they merely process them in a way that has to be precisely defined by the user.[56] The computer does not make a qualitative distinction on its own between alphanumeric or special characters like a whitespace. Such differences, i. e. linguistic or contextual information, must be made explicit at the data level if they are to be taken into account in computational processing. Hence, in contrast to human text processing, which infers the level of meaning through the aggregated world knowledge, the

54 France, Critical Edition of the Historia Francorum, pp. 344 – 350. The edition identifies a total of three biblical references for the capture of Jerusalem, one of which is not indicated by Hill and Hill, Le "Liber," pp. 150 – 152. The words "cantantes canticum novum" are shown as referring to Rev 5 : 3, while Hill and Hill list three possible psalms at this point.

55 Silke Schwandt, Digitale Methoden für die Historische Semantik. Auf den Spuren von Begriffen in digitalen Korpora, in: GG 44. 2018, pp. 107 – 134, here p. 108 and p. 133.

56 Cf. Fotis Jannidis, Methoden der computergestützten Textanalyse, in: Vera Nünning and Ansgar Nünning (eds.), Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze – Grundlagen – Modellanalysen, Weimar 2010, pp. 109 – 132, here p. 110.

computer identifies and counts phenomena on the surface of the text.[57] That means that "text" for the computer is still primarily a sequence of characters, or a set of segments that can be used to make certain calculations and provide insights into the ways in which the lexical items of a document or collection are related: How often does each character string appear in the text? Which character strings appear together more often? How does their composition differ from text to text?

Basing the results on a certain rule-based sequence of processing steps and calculations, i. e. on an algorithm, makes it theoretically always possible to trace how exactly a certain result is achieved.[58] While the same algorithm can be applied to large amounts of texts or data, the evaluation criteria laid down in the algorithm remain the same throughout the whole process. This means that every text passage is systematically reviewed in the same way, no matter how large the text is. However, the big challenge is to operationalize the respective research question algorithmically. One needs to have a very good prior understanding of the research subject as such in order to be able to do this in an adequate way. In the then step-by-step and often iterative application of the algorithm, one becomes increasingly better acquainted not only with the methods, but also with the research object itself.

Text Re-Use Analysis as a Methodological Framework

In order to identify the re-use of parts of texts in another text, i. e. to determine where, to what extent, and in what way the Bible was used in the Crusade Chronicles, we draw on the method of text re-use analysis – a method that has already been used for centuries. As a distinct field of research, it encompasses both the research traditions of humanities scholars, especially philologists, who have identified and systematized traces of intertextuality in historical sources through meticulous studies and textual criticism, and more recent approaches in computer science.[59]

As the name indicates, it is basically about the repeated use of text. This raises two questions: first, what do we mean specifically when we speak of re-use? Second, what do we mean when we speak of text? In general, we understand text as a document composed of various linguistic units such as words, phrases, sentences, and paragraphs. Text re-use occurs when parts, ideas or

57 Cf. ibid.

58 This applies at least to the algorithms that are rule-based. However, if the algorithm also includes the application of machine learning methods, this is almost impossible at the moment.

59 See Marco Büchler, Informationstechnische Aspekte des *Historical Text Re-Use*, PhD thesis Universität Leipzig 2013, https://ul.qucosa.de/api/qucosa%3 A11877/attachment/ATT-0/, p. iv, pp. 26 f., pp. 31 – 38, pp. 41 f., pp. 46 f., pp. 50 – 52 and pp. 57 – 62; Paul Clough and Rob Gaizauskas, Corpora and Text Re-Use, in: Anke Lüdeling and Merja Kytö (eds.), Corpus Linguistics. An International Handbook, Berlin 2009, pp. 1249 – 1271, here pp. 1262 – 1265.

concepts of already existing documents are extracted and used in a new document and context.[60] This can happen, as we have already seen, in very different ways. Text re-use can be assessed according to different categories, such as the formal relationship, i. e. the degree of literal or textual concordance, between the original and re-use, or the degree of intentionality, i. e. whether re-use has taken place consciously or unconsciously.[61] One of the most obvious forms of re-use is certainly the literal quotation, in which content is deliberately taken word for word from one text to be integrated into another. Marco Büchler classifies verbatim quotations as "syntactic re-use," since the order of words or the syntactic context is largely preserved. This field also includes winged words, common phrases, and multi-word units.[62] We can distinguish this syntactic re-use from those text re-uses where the content has been reworded to fit into a new context or altered by using, say, synonyms or hyponyms. This category is referred to as "semantic text re-use." It includes allusions, paraphrases and analogies, but also applies to translations and summaries.[63]

One question that can hardly be solved with the digital approach concerns the level of intention of the text re-use or, especially, its absence. The different types of text re-use can convey information and knowledge, but they can also simply indicate stylistic devices and certain forms of expression.[64] The question of whether this happened intentionally or unintentionally in each case is detached from this. Stylistic devices and expressions can make relationships or influences visible even on an unintended level. More importantly, however, this demonstrates that text re-use is more than the intentional adoption of a formulation, idea or concept through quotations or paraphrases. It also includes non-intentional use of idiomatic linguistic expressions.[65]

In his dissertation, Marco Büchler drew attention to the difference between short "Text Re-use" and "Language Re-use": How can we distinguish between an actual, meaningful reprise of text and simple idiomatic co-occurrences, i. e. common expressions that are widely used?[66] As mentioned earlier, this distinction can be made based on experience when applying the analogous approach, as has been done for centuries in text-based humanities and especially philologies. To the machine, the decision whether a textual correspondence is meaningful or not has to be built into the algorithm separately (as far as this is possible), since it processes the text only as a sequence of characters without automatically including context or semantics.

60 Cf. ibid., p. 1249.
61 Cf. ibid., p. 1251.
62 See Büchler, Informationstechnische Aspekte, p. 40 and p. 77.
63 Cf. ibid.
64 See ibid., p. 28 and p. 30.
65 See ibid., p. 24.
66 See ibid., p. 43 f.

The only decisive factor here is that several words are used in approximately the same way in both texts, regardless of whether they are common linguistic phrases or quotations. This means that one risks obtaining a great deal of noise in the form of unwanted results, which needs to be reduced. This is perhaps the most important challenge for the digital approach described here. Whatever the result, it always requires careful manual examination, filtering and interpretation. After all, and this is important to underline, the final assessment of the individual results in the context of the respective research question can only be done by the scholar him or herself.

While the distinction between the different formal types of text re-use is secondary for the analogous approach, their more detailed classification is of the highest importance for the digital approach. In using digital methods, we must clearly define from the start what exactly we are looking for, since each type of text re-use is characterized by different features that we have to search for, often with a different algorithm. For example, detection of syntactic re-use can be done by calculating whether certain words occur together, perhaps even in a particular order. This involves dividing the two texts to be compared into segments, such as chapters, paragraphs, sentences, or sub-sentences. For these, we can then either calculate which words occur individually or in conjunction with certain other lexical units. That is, which two-word groups (bi-grams) or three-word groups (tri-grams) are contained in each segment. Then the individual segments of the different texts can be compared with each other and analyzed for similarity with regard to the number of these features.

Additional lists can be added for the words used in those texts, making the comparison more robust against changes in spelling. For example, one can normalize the different spellings in the texts (for example, "analysis" and "analyzis"), which is especially important for texts on earlier levels of historical language. Furthermore, we can trace with the help of a lemmatizer the respective words back to their basic form (lemma), so that grammatical changes in the respective text passages in terms of number, tense, case, et cetera can be neutralized. If the ultimate aim is to detect not only syntactic but also semantic text re-use, one can add further lists that map the lemmas used in the texts to lists with synonyms and hyponyms, i. e. with terms of alternative and superordinate meaning. In this way, it is theoretically also possible to find those instances of text re-use where individual words have been replaced by other, similar words.

This identification procedure is based on workflows and calculational approaches that are also widely used in other text analysis and processing techniques. For instance, a common strategy to identify resemblance between re-use and potential root is to apply some measures of similarity.[67] These

67 Cf. Clough and Gaizauskas, Corpora and Text Re-Use, p. 1249.

established methods all make more or less use of the statistical distribution of lexical units, such as words. For example, in stylometry, different lexical characteristics are used to describe the stylistic similarities that a text shares with other texts, which is primarily based on the word distributions in the texts to be compared. These similarities can be used, among other things, to clarify questions of authorship of texts or to identify genre-specific features.[68] Another example, practically the counterpart to text re-use, is the computational collation in the course of establishing digital editions. Here, the objective is to find not the same or similar passages in the different text versions, but deviations in order to assign them to one another.[69] Regardless of whether it is a matter of investigating authorship, text re-use or collation, the techniques used to calculate similarities and differences, such as sequence alignment, overlap of n-grams or vector space models, are often based on established procedures in for example bioinformatics or the broad research field of Information Retrieval;[70] in other words, fields of research that we would not normally associate with the humanities.

Lastly, it should be noted that text re-use analysis, as we apply it in this paper, refers essentially to the same concepts as those used in conventional plagiarism software.[71] Nonetheless, instead of applying this technology to contemporary texts like doctoral theses, the aim here is to test it with regard to its applicability to historical language corpora. And the computational detection of textual borrowings in historical corpora is considerably more complex compared to contemporary texts. This is due to the fact that historical texts are much less standardized. On the one hand, we have to deal with a temporally and regionally determined wealth of variations in spelling. On the other, the evolution of language plays a role, in the same way that the semantic shift of a term can affect the form of re-use.[72] Finally, and equally importantly, there are many more resources and useful tools available for contemporary languages than for their historically earlier stages. This wealth of resources

68 Cf. Jannidis, Methoden der computergestützten Textanalysen, pp. 112 – 114.

69 Cf. Ronald Haentjens Dekker et al., Computer-Supported Collation of Modern Manuscripts. CollateX and the Beckett Digital Manuscript Project, in: Digital Scholarship in the Humanities 30. 2015, pp. 452 – 470.

70 For a brief overview, see Clough and Gaizauskas, Corpora and Text Re-Use, pp. 1253 – 1262.

71 For an overview and evaluation of procedures for identifying plagiarism, see Martin Potthast et al., Overview of the 5th International Competition on Plagiarism Detection, in: Pamela Forner et al. (eds.), Working Notes for CLEF 2013 Conference, Valencia 2013, http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-PotthastEt2013.pdf; Potthast et al., Overview of the 6th International Competition on Plagiarism Detection, in: Linda Cappellato et al. (eds.), Working Notes for CLEF 2014 Conference, Sheffield 2014, pp. 845 – 876. Better known is the "VroniPlag" project, in which, among others, doctoral theses i. a. by politicians were checked (mainly manually) for plagiarism; see the collaborative VroniPlag Wiki, https://vroniplag.wikia.org/de/wiki/Home.

72 See Büchler, Informationstechnische Aspekte, p. 26 and p. 45.

allows, as for contemporary English, the use of other methods and techniques that are not available for historical languages like medieval Latin.
In the context of this paper, we decided to proceed exemplarily on the basis of a tool called "Tracer."[73] This command line tool uses a number of different procedures that eventually lead to a comparison between different texts, searching for the parts where they are identical or at least similar in some way. However, what the tool actually does is considerably more complex. Tracer has been developed by Marco Büchler to provide a tool that brings together different approaches of text re-use detection in one single framework. The architecture of the Tracer tool consists of roughly 700 algorithms that can be arranged modularly.[74] In other research projects, the tool is used for the detection of motifs in folkloristic literature,[75] or as part of authorship attribution in historical texts.[76] A small community producing regular scholarly output has formed itself around its utilization.[77]

III. The Concrete Implementation of Text Re-Use Analysis with the Tool Tracer

We will now demonstrate what it means in concrete terms to use the process described in practice, using the tool Tracer as an example. As for the text, we have chosen the "Historia Francorum qui ceperunt Ihrusalem" by Raymond

73 Marco Büchler, TRACER. A Text Reuse Detection Machine, https://doi.org/21.11101/0000–0007-C9CA-3. Other alternatives would be "Passim," a popular open-source software designed for the detection of re-use in both historical and contemporary corpora, which detects significant overlaps and aligns these repetitions into larger clusters; see David A. Smith, passim, https://github.com/dasmiq/passim; Smith et al., Detecting and Modeling Local Text Reuse, in: 2014 IEEE / ACMN Joint Conference on Digital Libraries (JCDL), London 2014, https://doi.org/10.1109/JCDL.2014.6970166, pp. 183 – 192; BLAST, a software which was originally developed for fast matching of biological sequences against large sequence databases and is especially useful when dealing with error-prone OCR data, see Aleksi Vesanto, Text Reuse Detection with BLAST, https://github.com/avjves/textreuse-blast; Vesanto et al., Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771 – 1910, in: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Gothenburg 2017, pp. 54 – 58; or the "textreuse" package in "Cran," see Lincoln Mullen, textreuse. Detect Text Reuse and Document Similarity. R package version 0.1.5, 2020, https://CRAN.R-project.org/package=textreuse.
74 Cf. Greta Franzini et al., TRACER Text Reuse Detection Machine. The User Manual, https://tracer.gitbook.io/-manual/.
75 Digital Breadcrumbs of Brothers Grimm, http://www.etrap.eu/digital-breadcrumbs-of-brothers-grimm/.
76 Tracing Authorship in Noise, http://www.etrap.eu/research/tracing-authorship-in-noise-train/.
77 Historical Text Reuse Zotero Group, https://www.zotero.org/groups/500373/historical_text_reuse.

d'Aguilers as a smaller test case, which we will examine for traces of references to the Bible.

1. The Mis-En-Place

The source code of Tracer is published under the Academic Free License 3.0 and is therefore freely available.[78] The basic instructions are explained in a written documentation.[79] For installation and running, only Apache Ant and Java 8 are required.[80]

Since our method is based on the analysis of the verbatim of the text and not on related ideas and associations, the question of the text basis is much more important here than in the analogue method. John and Laurita Hill's historical-critical edition from 1969 is currently the authoritative version for Raymond d'Aguilers' text of the "Historia Francorum."[81] Furthermore, there is also a second edition from about the same time (1967) by John France.[82] However, we only dispose of both in printed form, which would mean having to scan the text, to OCR it, and to post-process the OCR results, in order to establish a usable digital version of the text, which however does not exclude any errors that may occur in the process. On the other hand, the "Computational Historical Semantics Project" provides access to more than 4,000 texts in medieval Latin,[83] including the "Historia Francorum." These texts are for the most part printed historical editions that have been verified, annotated word- and sentence-wise, and heavily enriched with linguistic information like the lemma of each word and its word-category, et cetera. For the "Historia Francorum," the text is drawn from Jacque-Paul Migne's "Patrologia Latina," which for the most part dates back to the seventeenth century.[84] Although it is much older than the printed editions mentioned above, the fact that this version has already been converted into data, checked and annotated makes it

78 A copy is available on GitLab, https://vcs.etrap.eu/users/sign_in.
79 Franzini, TRACER User Manual.
80 Ibid.
81 Hill and Hill, Le "Liber."
82 France, Critical Edition.
83 Roberta Cimino et al., Digital Approaches to Historical Semantics. New Research Directions at Frankfurt University, in: Storicamente 11. 2015, no. 7, pp. 1 – 16, here p. 5.
84 Raymond d'Aguilers, Historia Francorum qui ceperunt Jerusalem, in: Jacques-Paul Migne (ed.), Patrologia Latina, vol. 155, Paris 1854, cols. 591 – 668. Jacques-Paul Migne (ed.), Patrologiae cursus completus. Sive biblioteca universalis, integra, uniformis, commoda, oeconomica, omnium SS. Patrum, doctorum scriptorumque eccelesiasticorum qui ab aevo apostolico ad usque Innocentii III tempora floruerunt, vol. 155, Paris 1854, cols. 591 – 668. It is based on the text by Jacques Bongars (ed.), Gesta Dei per Francos, vol. 1, Hannover 1611, pp. 139 – 183. It should be noted that even the text offered by Hill and Hill is not beyond all doubt; see Conor Kostick, The Social Structure of the First Crusade, Leiden 2008, p. 27: "The editors seem to have been unaware of France's work, which has established that among the surviving manuscripts, MS 14378 is relatively far removed from the archetype." The generally successful edition by John France, however, must be partially addressed as strongly reconstructive.

much more valuable for the time being for our project and outweighs possible inadequacies in the textual basis.

The same reasoning applies to the version of the Bible that we used for our study. When analyzing the use of the Scriptures in the chronicle of Raymond d'Aguilers on the basis of the verbatim text, one must first consider the tradition of the textual versions of the Bible available to him. The Latin biblical text in the version established by Jerome in the fourth century became the most widespread biblical text until the High Middle Ages, which is why it is also called the Vulgate.[85] The Vulgate thus gradually replaced the earlier Latin translations of the Old and New Testament, commonly known as "Vetus Latina."[86] However, Jerome did not leave an authorized edition of his translations. Since it was distributed as handwritten copies throughout the centuries, it was inevitable that variants in the text would occur during the process. As older versions of the Bible and the Gospels were still in use, especially until the eighth and ninth centuries, and remained present in particular in the liturgy, they also influenced the process of textual transmission. As a result, other, older readings were also integrated into versions of Jerome's text, which makes the original wording unknown to us in detail.[87] From the "Carolingian Renaissance" onwards, the Bible revision of Alcuin of York, which he presented to Charlemagne on the occasion of his coronation as emperor of the Romans, functions in a sense as a first "standardized" version of the Bible and also influenced several other Bible versions in the centuries to come.[88]

Which of the various transcriptions were available to Raymond d'Aguilers cannot be determined with certainty.[89] What is decisive is the fact that the

85 However, as Pierre-Maurice Bogaert points out, this term is rather inappropriate for large parts of the Middle Ages and the Jerome Bible as such; see Pierre-Maurice Bogaert, The Latin Bible, c. 600 to 900, in: Richard Marsden and E. Ann Matter (eds.), The New Cambridge History of the Bible, vol. 2: From 600 to 1450, Cambridge 2012, pp. 69 – 92, here p. 69.

86 See Bogaert, The Latin Bible, in: James Carleton Paget and Joachim Schaper (eds.), The New Cambridge History of the Bible, vol. 1: From the Beginnings to 600, Cambridge 2013, pp. 505 – 526, passim.

87 Ibid., p. 508 and p. 518.

88 Cf. David Ganz, Carolingian Bibles, in: Richard Marsden and E. Ann Matter (eds.), The New Cambridge History of the Bible, vol. 2: 600 to 1450, Cambridge 2012, pp. 325 – 337, here pp. 330 – 334; see also the genealogy of different Bible versions in: Raphael Loewe, The Medieval History of the Latin Vulgate, in: G. W. H. Lampe (ed.), The Cambridge History of the Bible, vol. 2: The West from the Fathers to the Reformation, Cambridge 1969, pp. 102 – 154 , here pp. 103 – 105.

89 However, Thomas Lecaque speculates that Raymond d'Aguilers had access to and was influenced by another complete but non-standardized Carolingian Bible (Biblia Aniciensis, Paris, Bibliothèque Nationale de France, Latin 4, 1 – 2) originating in the ninth century and housed in the library of the cathedral of Notre Dame du Puy. This version is not based on the Alcuinian texts; see Lecaque, Reading Raymond, pp. 105 f., p. 108, pp. 110 f. and pp. 113 f.

comparative material, the Bible, is rich in variations and that this affects the computational analysis. That is to say, as long as we do not know the version of the Bible that Raymond was trained in or that was at his disposal, we cannot really rely on the results of our analysis on the level of verbatim quotations, either – in contrast to what might be done, for instance, for the analysis of contemporary texts. Therefore, we have to allow for a certain flexibility from the beginning when it comes to the actual wording of the texts. A word-for-word analysis is only possible with reservations. For this reason, we have used in this case study the "Vulgate Clementina." Although its publication date of 1592 differs significantly from the period under study, it is a distillation of earlier versions of the Bible.[90] It comes closest to a standardized basis for comparison. The very version we used is derived from the "Clementine Vulgate Project,"[91] an online edition of the "Clementine Vulgate" which itself is based mostly on an edition of 1946.[92] This version is also accessible via the "Computational Historical Semantics Project" as thoroughly checked and annotated data.

The data are provided in the TEI format.[93] To transform the heavily annotated TEI files of the "Vulgate" and the chronicle into a format that Tracer can process, we wrote a Python script. This transformation also includes the breakdown of the texts into smaller chunks, the segments. These segments can have any size, depending on the type of re-use one wants to study.[94] Since, in our case, we are aiming to detect short expressions made up of no more than a few words, we segment the texts into short parts of sentences.

The end result of the prepared and segmented textual data has to be a simple text file with each line being a segment of the corpus. In addition, it needs to store a unique numerical identifier for each segment, the date of creation of the text file, and the name of the text to which the respective segment belongs as tab-separated values.[95] All texts to be compared must be put in this way and combined in a single file. The result looks like this:

1100000001	apocalypsis b	2020–04–30	Apocalypsis_B_Ioannis_Apostoli
1100000002	ioannis apostoli	2020–04–30	Apocalypsis_B_Ioannis_Apostoli

90 On the so-called Sixto-Clementine Vulgate, see Bruce Gordon and Euan Cameron, Latin Bibles in the Early Modern Period, in: Cameron (ed.), The New Cambridge History of the Bible, vol. 3: From 1450 to 1750, Cambridge 2016, pp. 187 – 216, here pp. 211 – 215.

91 Clementine Vulgate Project. The Full Text of the Clementine Vulgate, Freely Available Online, http://vulsearch.sourceforge.net/index.html.

92 Turrado Colunga et al., Biblia Sacra Iuxta Vulgatam Clementinam, Madrid 1946.

93 TEI. Text Encoding Initiative, https://tei-c.org/.

94 See Büchler, Informationstechnische Aspekte, p. 91.

95 Franzini, TRACER User Manual.

1100000003	apocalypsis iesu christi, dedit illi deus palam facere servis suis, quae oportet fieri cito 2020–04–30 Apocalypsis_B_Ioannis_Apostoli	2020–04–30	Apocalypsis_B_Ioannis_Apostoli
1100000004	significavit, mittens angelum suum servo suo ioanni, qui testimonium perhibuit verbo dei, testimonium iesu christi, quaecumque vidit	2020–04–30	Apocalypsis_B_Ioannis_Apostoli
1100000005	beatus qui legit, audit verba prophetiae huius, servat ea, quae ea scripta sunt	2020–04–30	Apocalypsis_B_Ioannis_Apostoli
1200000001	raimundi agiles canonici podiensis historia francorum qui ceperunt jerusalem. monitum.	2020–04–30	historia-francorum
1200000002	raimundus agiles agilaeus, canonicus podiensis, episcopi sui, qui comes tolosanus erat, capellanus, quo an	2020–04–30	historia-francorum
1200000003	1095 terram sanctam abiit, res quinquiennium gestas testis descripsit	2020–04–30	historia-francorum

These data can already be used for text re-use detection with Tracer. But to detect not only verbatim quotes, we need to improve comparability between the texts. This is done, as described above, by lemmatization. Tracer can perform this with a mapping table that maps all the inflected forms of a given word appearing in the corpus to the corresponding lemmata. Additionally, the part-of-speech tag (indicating the word-category) of the word has to be included. The mapping table for the lemmatization has to be structured as follows:[96]

abiectam abiectus a
abiecti abicio v
abiectio abiectio n

Both the Bible and the chronicle have been lemmatized and provided with POS-tags by the "Computational Historical Semantics Project."[97] We can extract this information from the TEI files that the project provides. By drawing on this, we have at our disposal 318,082 out of 473,164 words (tokens)

96 Ibid.
97 Cimino, Digital Approaches to Historical Semantics, p. 9.

in lemmatized form across our whole corpus. This includes all books of the "Vulgate" as well as the chronicle.
Since we are comparing two texts whose date of creation is separated by several hundred years and which were apparently also lemmatized separately, some lemmata differ in spelling between our two texts (for example, "charitas" and "caritas"). To account for such variations in spelling, we added alternative written representations – for each lemma in our corpus – to our lemma file. The data were obtained from the Knowledge Base "LiLa: Linking Latin."[98] This project integrates Latin language data from multiple sources – including spelling variations – and publishes them as Linked Open Data. The data can be extracted through a SPARQL endpoint.[99]
This illustrates an important issue of research data in historical studies in general: namely, high data quality (with regard to the sources and historical context) does not necessarily guarantee its comparability and interoperability with other datasets. If this interoperability is not ensured, a lot of effort has to be made to make these datasets processable in a common framework. Otherwise, any subsequent analysis would be subject to errors. In our case, we had to extend the process of data preparation by including additional, normalizing data in order to build a bridge between the two texts.

2. The Process

Having prepared the data in the way described above, we can now set up the text analysis itself. As already mentioned, Tracer consists of six different steps that compose its architecture, which are pre-processing, training, selection, linking, scoring, and post-processing. Each comes with a variety of implementations for different algorithms, each pertaining to a particular part of its processing architecture. Which algorithm one selects for each step strongly depends on the data that are analyzed (for example, are there several texts to be compared or one? To what extent do these texts differ in length?), as well as on what one wants to detect in this way (for example, what is the overall research question? Which re-use types does one want to find?). The execution of each step results in the production of interim results that are stored as part of the research data. This has the advantage not only of making the whole analysis transparent, but also of reducing overall computing time when results of previous steps can be re-used.
The first step, pre-processing, prepares each word for the subsequent analysis. The goal is to standardize the words, for example by changing all uppercase

98 Marco Carlo Passarotti et al., The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin, in: Thierry Declerck and John P. McCrae (eds.), Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK-PS 2019), Leipzig 2019, pp. 6 – 11.
99 LiLa: Linking Latin. Triplestore, https://lila-erc.eu/sparql/.

letters to lowercase letters and creating references to their respective lemmata. The requirements for this part have already been described above.
The next step, training, divides the text into defined sets of characteristics or features that Tracer can compare. This is about describing the text in the segments and extracting the characteristics that they carry, i. e. the different words that they contain or the particular order in which the words are found here. In order to do so, we can choose, for instance, between a word-based and an n-gram-based approach. Since the first takes into account all the words contained in a segment equally and thus ignores the concrete order of the words, it is better suited to detecting more liberal text re-use like memes and allusions. N-grams, on the other hand, are better suited to detecting full verbatim re-use.[100] It is therefore a prerequisite for the digital approach to be clear from the beginning as to which specific kind of text re-use we want to detect in a given analysis run. This choice requires a firm understanding of the different categories of text re-use in the first place, which would be more grounded if we had a better understanding of the actual ways that the text of the Bible could have been used and re-used by Raymond d'Aguilers and his contemporaries. That means that we cannot presume to find all kinds of text re-use at the same time in a single run; rather, we have to proceed step-by-step and, most importantly, to explore the different methods and how they work on our material.
The most crucial decision, however, lies in the next step, which is to choose a specific selection algorithm that filters the features that are most likely to indicate a re-use. Here, Tracer draws on a method called fingerprinting. Its goal is to create a mathematical representation of the characteristics of a feature (in our case, a word) in its textual context. For this step, a high number of variables have to be taken into account: Again, which types of text re-use should be detected? Which information shall be considered during the calculations: should they be restricted to information embedded in every word of the text (for example, their part-of-speech tags) or should other measures like distribution frequencies over the whole text be taken into account as well?[101] The selection process reduces the number of available features of the segments to a set that is manageable and yet significant. For our purpose, initial tests have shown that, at the moment, the best results can be achieved by an algorithm called "Global Entropy Selection." This algorithm combines several concepts of Information Theory to filter the features, including entropy (a measure describing the density of a piece of information in relation to the characters used to communicate it)[102] and redundancy.[103]

100 See Büchler, Informationstechnische Aspekte, p. 101.
101 See ibid., p. 106.
102 Claude E. Shannon, A Mathematical Theory of Communication, in: Bell System Technical Journal 27. 1948, pp. 623 – 656.
103 See Büchler, Informationstechnische Aspekte, p. 110.

The next two steps in the Tracer architecture are linking and scoring. Their aim is to take the features of the segments, selected in the previous step, and calculate the similarity between each pair of features.[104] The algorithms for linking differ depending on whether the aim is to compare two versions of the same text or two completely different texts. Accordingly, the selection of a scoring technique depends on the extent to which the compared texts are identical or not. Comparing two different texts, we use an implementation that is able to link features from different sources. Since our purpose is to detect elements of one text (the Vulgate) in another text ("Historia Francorum"), using a containment scoring strategy is required.[105] Thus, we used an algorithm called "Selected Feature Containment Similarity."

The last step, post-processing, serves as a bridge between the calculated results and their visualization in a human readable way. At the end, the scholar is presented with a representation of the two texts and references to the segments that may be connected by similar features. Although Tracer is able to detect text passages in which re-use is alleged to have taken place, this is only based on a statistical probability and, as mentioned above, does not indicate in any way whether those similarities are based on an intended text re-use or not. What we are provided with is merely a list of possible candidates for text re-use. The interpretation and thus the evaluation of these results is then entirely up to us. To decide which of these possible intertextual relationships are actually meaningful and in what way requires a deep understanding of the subject, the texts and their contexts as such. At the same time, to recognize and explain possible patterns and distortions in the results also requires a sound understanding of the methods used during the process – so that, if needed, the selected algorithms can be refined at the end in order to further optimize the results of the next run. This is part of the knowledge discovery process at various levels: it requires that we constantly engage with the texts, the results of our statistical calculations, and the methods that we use to obtain them within the context of our specific subject, and thus simultaneously expands our knowledge in all these areas.

3. Testing and Improving

In our given case, starting from our initial situation, we had to try multiple algorithms in different combinations and evaluate our results to find the best solution. In order to keep things manageable, we have limited ourselves in this paper to the detection of syntactic re-use.

104 See ibid., p. 90.

105 The alternative strategy would be called resemblance. Andrei Z. Broder, On the Resemblance and Containment of Documents, in: Bruno Carpentieri et al (eds.), Proceedings. Compression and Complexity of SEQUENCES 1997, Los Alamitos, CA 1998, pp. 21 – 29.

To assess the performance of our various configurations, we first established a kind of preliminary ground truth.[106] In other words, we collected the known cases of syntactic text re-use of passages from the Bible in the "Historia," as they are mentioned in the different critical editions of the text. With every run of our analysis, we could check our results against this list to see if they are included.

This gives us a baseline to evaluate different configurations of Tracer by calculating different quality measurements called precision, recall and F_1 Score – common measures used in Information Retrieval to assess the quality of automatically obtained results.[107] Precision describes the ratio of the desired results to the total number of results, while recall represents the ability of the system to detect all the desired results (true positives).[108] The F_1 Score is a normalized combination of these two measures, the aim being to make different detection methods comparable as a whole.

Although one wants ideally to detect as many elements of the predefined ground truth as possible (recall), while avoiding a large number of false positives (precision), combining these two at the same time is hardly ever possible. In practice, one has to decide the direction in which one wants to maximize the results.[109] To ensure that Tracer detects a high number of possible candidates and to avoid having any good results already sorted out during the detection process, we have attempted to optimize our configuration for a higher recall. This inevitably led to a very low level of precision, i. e. a large number of false results in our result sets.

As might be expected, the first results were very mediocre. When we tested the different configurations, the results varied greatly both in quantity and quality – largely depending on our choice of a selection strategy.[110] While the recall was improvable, the precision was very poor. It was noticeable that Tracer quite often detected linguistic expressions as re-uses, for example, "… et … et …" or recurring biblical expressions, like "dixit dominus." Although these are in principle also types of re-use, they are mostly recurring (linguistic) structures and not the kind of intended re-use that we were aiming for. To tackle this

106 This ground truth is preliminary, since we do not know even for a certain part of the chronicle nor for a certain book of the Scriptures what re-use can actually be found. Only then would this be a real ground truth. This knowledge, however, can only be established exploratively.

107 Thomas Mandl and Christa Womser-Hacker, Information Retrieval, in: Mehdi Khosrow-Pour (ed.), Encyclopedia of Information Science and Technology, Hershey 2015, pp. 3923 – 3931.

108 Precision is calculated by dividing the number of expected results in the ground truth that were actually detected by the number of all detected results. Recall is calculated similarly, but by dividing the number of expected results that were detected by the number of all expected results; see ibid., p. 3926.

109 Martin Potthast, Technologies for Reusing Text from the Web, Weimar 2011, https://doi.org/10.25643/bauhaus-universitaet.1566, pp. 20 f.

110 At the beginning, the number of results amounted to 30,000 in some cases.

issue, we redefined our data preparation procedures by removing all stop words before analyzing the corpus with Tracer.[111] This significantly improved the quality of our results.

In order to improve our recall, we experimented with multiple variations in the length of segments and the use of n-grams instead of word-based features. Furthermore, instead of comparing the whole text of the Bible to the "Historia," we decided to compare them book by book, in order to keep the particularities of the different books of the Bible within the analysis, since in the selection process the context of the different words in the respective text is also taken into account. Besides that, it helped to reduce the computing time for the different runs. Thus, we limited our analysis to the comparison with the Book of Psalms and the Book of Revelation, from which stemmed most of our approximative ground truth. Furthermore, these texts were the most likely to yield interesting results, since previous researchers have suspected that these writings were the most influential for Raymond d'Aguilers.[112] These different strategies enabled us to improve our recall from 33 to 68 percent. At the same time, we learned a lot more about the tool, the different algorithms, and our data.

The various tests and their evaluation have shown that a recall of nearly 100 percent is certainly impossible in this way. Remaining tasks, like the reduction of noise, can be tackled with existing mathematical models and algorithms, implemented in the tool. Here, the challenge lies in the complexity of these models not only in themselves, but also in their mutual dependencies and their concrete application to the respective data set. As we have discussed above, the digital approach – unlike the analogue approach – is based entirely on the formal characteristics of the quotations to be found. This has to be taken into account on every level. Therefore, we first need to understand better both those formal features and the data that shape them. This is the crucial prerequisite for deciding which configuration of algorithms and parameters – or, rather, which combination of several different configurations – we need to apply in order to detect the type(s) of re-use that we are interested in, and to do this in the specific environment of Latin biblical texts and medieval historiography. For, even if we stay on the level of syntactic re-use detection, we are still dealing with several types of text re-use, each showing different formal characteristics. Especially the selection step in Tracer's architecture needs a different adaptation for each type we are looking for. Besides that, there are also other strategies to explore in the further process, such as including part-of-

111 These are functional words such as "et," "in" or "ad" that, although very common in a language, do not have much intrinsic meaning.

112 See chapter I.

speech tagging to the analysis or radius retrieval,[113] in order to improve our recall and to reduce the noise, too.

4. First Results

Despite the challenges and range of possible improvements described above, we nevertheless discovered already during our tests some new re-uses in the "Historia," re-uses that have remained unknown or at least unmentioned so far. The examples that we found in our preliminary tests can only be proof of concepts. A few of them shall be presented here.

Our first reference concerns a star that hovered over the city of Antioch, fell into three parts, and descended behind the enemy lines: "Eo tempore contigerunt nobis plurime revelationes per fratres nostros, et signum in celo mirabile vidimus."[114] John and Laurita Hill interpreted this as a re-use from the Book of Jeremiah. At least, they marked this passage with a reference to Jeremiah 10 : 2: "haec dicit Dominus iuxta vias gentium nolite discere et a signis caeli nolite metuere quae timent gentes."[115] But there are at least two other candidates from the Book of Revelation with the same degree of resemblance. These passages can be found in 12 : 3 and 15 : 1 respectively:

Rev. 12 : 3: "Et visum est aliud signum in cælo et ecce draco magnus rufus habens capita septem, et cornua decem : et in capitibus ejus diademata septem,"

Rev. 15 : 1: "Et vidi aliud signum in cælo magnum et mirabile, angelos septem, habentes plagas septem novissimas : quoniam in illis consummata est ira Dei."

Whether this linguistic similarity was also intentional and refers to an intended reference remains to be discussed.[116] It can also simply be a matter here of a general biblical language, which refers in this way to heavenly signs.[117]

More interesting therefore is our second example. It is a passage from the "Historia" where a discussion breaks out among the conquerors about the

113 Amihood Amir et al., Consensus Optimizing Both Distance Sum and Radius, in: Jussi Karlgren et al. (eds.), String Processing and Information Retrieval, Berlin 2009, pp. 234 – 242.

114 Hill and Hill, Le "Liber," p. 74. Cf. Karin Fuchs, Zeichen und Wunder bei Guibert de Nogent. Kommunikation, Deutungen und Funktionalisierungen von Wundererzählungen im 12. Jahrhundert, München 2008, p. 192.

115 See Hill and Hill, Le "Liber," p. 74.

116 Contemporaries generally shied away from clear references, and the other chroniclers of the First Crusade also partly refused an interpretation of celestial phenomena; see Fulcheri Carnotensis Historia Hierosolymitana, I 14, pp. 203 – 205; Guibert de Nogent, Dei Gesta per Francos et cinq autres textes, ed. by Robert B. C. Huygens, Turnhout 1996, VII 35, pp. 333 f.

117 Thus, the parallel report in the Gesta Francorum et aliorum Hierosolymitanorum, ed. by Rosalind Hill, London 1962, p. 62: "Nocte quippe superueniente, ignis de caelo apparuit," could be traced back to two Kings 1 : 14: "Ecce descendit ignis de caelo."

possibility, once Jerusalem has been taken, of electing a king who shall protect the Holy Land. This idea, according to Raymond d'Aguilers, is firmly rejected by the bishops: "Non debere ibi eligere regem ubi Dominus passus et coronatus est. Quod si in corde suo diceret, sedeo super solium David et regnum eius obtineo, deneger a fide et virtute David."[118]

Neither John and Laurita Hill nor John France give any references for the expression "Quod si in corde suo diceret."[119] As a matter of fact, in Psalm 9, the expression "dixit enim in corde suo" is repeated three times, indicating a clear emphasis.[120] This Psalm of David speaks of the blasphemer and godless ruler who disregards God. Reference to this Psalm makes the meaning of this passage from Raymond's "Historia" much clearer. It paints the danger of installing a godless king as direct successor to David and ultimately to Christ. This very much makes sense since Raymond considered it illegitimate for Godefroy de Bouillon to accept the royal dignity, whereas his own count piously rejected it.[121] This new evidence thus clarifies the argument that Raymond is making here with a direct reference to the Bible.

The third reference that we found during our tests may also help make a given place in the chronicle more comprehensible. This passage ties in directly with the scene mentioned above, and describes the situation after these confrontations, before God finally sends them a sign as to how they could make up for this situation with him: "Tandem misericors et propicius Dominus propter nomen suum simul ne adversarii nostri legi eius insultarent, dicentes ubi est Deus eorum?"[122]

For this passage, John and Laurita Hill, as well as John France, offered reference to the Bible. John and Laurita Hill link the passage in question to the Book of Judith 7:21, where the Israelites are in a desperate situation, cited here with a little more context: "(20) Tu, quia pius es, miserere nostri, aut in tuo flagello vindica iniquitates nostras, et noli tradere confitentes te populo qui ignorat te, (21) ut non dicant inter gentes: 'Ubi est Deus eorum?'"

The passage from the Bible does not really fit for this part of the chronicle, since it does not create any hope, but foreshadows an act of surrender. John France, for his part, links this passage to Psalm 113:10, which is a better match:[123] "(9) Non nobis, Domine, non nobis, sed nomini tuo da gloriam: (10) super misericordia tua et veritate tua; nequando dicant gentes: Ubi est Deus eorum?"

118 Hill and Hill, Le "Liber," p. 143.
119 Ibid.; France, Critical Edition, p. 323.
120 Psalm 9:27, 9:32, 9:34.
121 Hill and Hill, Le "Liber," p. 152. See also Jay Rubenstein, Godfrey of Bouillon versus Raymond of Saint-Gilles. How Carolingian Kingship Trumped Millenarianism at the End of the First Crusade, in: Matthew Gabriele and Jace Stuckey (eds.), The Legend of Charlemagne in the Middle Ages. Power, Faith and Crusade, New York 2008, pp. 59 – 75.
122 Hill and Hill, Le "Liber," p. 143.
123 France, Critical Edition, p. 325.

With the help of Tracer, we can add another passage that seems even more relevant than the other two. Psalm 78:10 says, quoted here in its larger context:

(8) Ne memineris iniquitatum nostrarum antiquarum; cito anticipent nos misericordiæ tuæ, quia pauperes facti sumus nimis. (9) Adjuva nos, Deus salutaris noster, et propter gloriam nominis tui, Domine, libera nos: et propitius esto peccatis nostris, propter nomen tuum. (10) Ne forte dicant in gentibus: Ubi est Deus eorum? et innotescat in nationibus coram oculis nostris ultio sanguinis servorum tuorum qui effusus est.[124]

Here, the crucial postscript about bloody revenge can actually be linked to the similar bloody events during the capture of Jerusalem. Raymond d'Aguilers may have intentionally used this passage to accentuate the providential work of God in the deeds of the crusaders.

These examples illustrate that we are indeed able to recognize hitherto undiscovered Bible passages in the text of a chronicle in an automated way, and give an insight into the possibilities that are connected with this method. But it should also have become clear that the way there is not easy, and that further research is needed before the method can be applied to perhaps more texts.

In the next step of the project, we would therefore have to optimize our setup to improve our results further while suppressing the informational noise. Building on this, we could then improve our data preprocessing by integrating synonyms, hyponyms and co-hyponyms, so that we can ultimately recognize not only syntactic but also semantic re-use.

IV. Conclusions, or What Can Be Learned about the Specificities of the Digital Approach

The starting point of this paper was the observation that digital methods have apparently not yet fulfilled the often high expectations placed on them to achieve significant new insights. To understand why this is the case, one has to take a closer look at what digital methods actually are. For this purpose, it is important to understand how digital methods work in practice and how they differ from analogue methods. This was demonstrated by the example of the analysis of text re-use from the Bible in the chronicles of the Crusades, which is of central importance for the evaluation of these texts and their reports on the conquest of Jerusalem. We have examined how digital methods, in this case text re-use analysis in its concrete implementation with Tracer, can actually open up new perspectives on our discussion of well-known sources as well as on the construction of scholarly knowledge. It has become clear that the traditional analogue approach

124 Psalm 78:9 f. The importance of this psalm for the motivation of the crusade is explicitly pointed out by Althoff, Papst Urban II., pp. 147–149.

and the use of decidedly digital methods lead to different results, and are based above all on very different foundations.

In the universe of academic discourse, we are used to encountering texts, be they on paper or screen, at the layer of meanings. Formed into statements and arguments, these represent the real currency of debates in the humanities beyond the mere positivist findings. Michel Foucault rightly distinguishes them as "more than a mere collection of signs, which, in order to exist, need only a material base," as they necessarily require the "existence of an associated domain," since linguistic expression seems "always […] to be inhabited by the other, the elsewhere, the distant."[125] Aleida Assmann has characterized this "progression from the present to the absent" as a "thought-rapid, even somewhat automatic reflex" that does not allow for any lingering on the level of the signifiers:[126] according to such a reading, the letters and sequences of signs present on the paper serve only as a material bridge to advance as quickly as possible from the "realm of signal" to the "realm of meaning" that blossoms outside the textual form and is constructed from "cultural units" and that forms the actual arena of studies in the humanities.[127] In this way, textual work becomes a mental matter, in the hermeneutic sense based on the individual experiences and intuitions of the interpreter. This leads to the "anticipation of meaning," which, according to Hans-Georg Gadamer, is the basis of any (hermeneutic) understanding of texts in the humanities.[128]

How computers deal with texts, on the other hand, is different. They do not operate in the "realm of meaning," but in the "realm of signal" just mentioned. Text is processed here as "a sequence of signs."[129] The best way to explain this is by applying the theory of signs. While the traditional humanities focus on the dimension of meaning, the signified, digital methods concentrate on the expressive layer of the written characters, the signifier.[130] Prevalent are "atomistic

125 Michel Foucault, The Archaeology of Knowledge and the Discourse on Language, New York 1972, p. 96 and p. 111.

126 Aleida Assmann, Die Sprache der Dinge. Der lange Blick und die wilde Semiose, in: Hans Ulrich Gumbrecht and Karl Ludwig Pfeiffer (eds.), Materialität der Kommunikation, Frankfurt 1988, pp. 237 – 251, here p. 238. The de-materialization of the text expressed here can certainly be criticized in the wake of a material turn, but it corresponds to a widespread view.

127 Umberto Eco, Einführung in die Semiotik, München 1994[8], here p. 64 and p. 74.

128 Hans-Georg Gadamer, Truth and Method, London 1989[2], p. 293.

129 Paul Caton, On the Term "Text" in Digital Humanities, in: Literary and Linguistic Computing 28. 2013, pp. 209 – 220, here p. 212. Cf. generally Julia Nantke, Annäherungen an eine digitale Semiotik. Chancen und Grenzen computergestützter Untersuchungsmethoden für die semiotische Analyse literarischer Texte, in: Zeitschrift für Semiotik 1 / 2. 2017, pp. 83 – 108; Joris J. van Zundert, Screwmeneutics and Hermenumericals. The Computationality of Hermeneutics, in: Susan Schreibman et al. (eds.), A New Companion to Digital Humanities, Chichester 2016, pp. 331 – 347.

130 See, as a basis, Ferdinand de Saussure, Course in General Linguistics, New York 1959, pp. 65 – 67.

notions" of a signal sequence reduced to the discrete individual characters.[131] The text is represented as data that, in the end, are nothing but a sequence of ones and zeros, i. e. the binary or machine code. These ones and zeros can stand for the various characters and spaces in a text, but not for any meaning.[132] The machine can only work with the information that we provide, and even that only in the prescribed sequence of operations that we have previously defined as our algorithm in the computer code.[133] All supplementary information such as meaning, context or relationships to other words or concepts must first be defined and made explicit so that we can make it available to the machine as data and incorporate it into our algorithms. All our intuition, our prior understanding of the things and how they work, which guides us in our human way to process the data, is outside of the machine and has first to be encoded in an explicit, distinct and unambiguous way in order to be incorporated in our operations. In most cases, this will be a strictly restricted code, linking a signifier only to a specific signified.[134] But, when this is done, we can combine very different and extensive amounts of data and build complex algorithms to process them. Once established, we can process more and more data and include them in our studies, so that eventually our approach is only limited by the computing power and data storage of the machines that we have at our disposal (and the quality of our programming code).

In this way, no revolutionary leaps are to be expected. In fact, with the digital approach, we first have to develop, conduct research on, and acquire a completely new set of methods. This takes time and a lot of research. But, with media change and the ever growing number of digitally available sources, this approach is becoming increasingly relevant and, in some cases, even inevitable.[135]

131 Noah Bubenhöfer and Joachim Scharloth, Maschinelle Textanalyse im Zeichen von Big Data und Data-Driven Turn. Überblick und Desiderate, in: Zeitschrift für germanistische Linguistik 43. 2015, pp. 1 – 26, here p. 13.

132 Cf. Fotis Jannidis, Zahlen und Zeichen, in: Jannidis et al. (eds.), Digital Humanities. Eine Einführung, Stuttgart 2017, pp. 59 – 67; Christof Schöch, Big? Smart? Clean? Messy? Data in the Humanities, in: Journal of Digital Humanities 2. 2013, no. 3, pp. 2 – 12, here p. 3, http://journalofdigitalhumanities.org/2–3/big-smart-clean-messy-data-in-the-humanities/.

133 On the functionality of algorithms, see for an introduction Thomas H. Cormen et al., Introduction to Algorithms, Cambridge, MA 2009[3], pp. 5 f. and pp. 11 – 14.

134 Cf. Eco, Einführung in die Semiotik, p. 49.

135 On the role of born-digital and reborn-digital sources for the study of history and their methodological and source critical implications, see, e. g., the contributions by Kiran Klaus Patel, Zeitgeschichte im digitalen Zeitalter. Neue und alte Herausforderungen, in: VfZ 59. 2011, pp. 331 – 351, as well as Nicola Wurthmann and Christoph Schmidt, Digitale Quellenkunde. Zukunftsaufgaben der Historischen Grundwissenschaften, in: ZF 17. 2020, no. 1, pp. 169 – 178, or studies on the Web as a historical source in Niels Brügger and Ralph Schroeder (eds.), The Web as History. Using Web Archives to Understand the Past and the Present, London 2017. Insofar as these sources are already available as data, their investigation with digital methods is evident, as recently shown by, for instance, Melanie Althage at the conference "Digital History. Konzepte,

This observation may also explain why, even though the potential of the method for the research question has already become clear from the initial results, our first experimental implementation of the text re-use method has not yet resulted in an efficient and easily accessible structure that would allow the given research questions to be answered quickly. The article should have made clear that this is hardly possible because analogue and digital approaches are far too different for this. Because of the differences in the analogue and digital processing of text described above, two completely different ways of thinking are necessary. Thus, there are two very different approaches at work here. Based on this, it has also become evident that, contrary to what is sometimes assumed, the analogue way of thinking and the questions based on it cannot simply be transferred to digital methods, either.

Instead of starting out from certain ideas and statements and drawing possible cross-references associatively, the digital method starts out from the individual characters and strings which compose words, word combinations and sequences, and the potential patterns that occur in them. These different text re-uses can take on very different forms. It may be that several words in a sequence coincide, or that these words are distributed over a larger segment interrupted by sense-expanding interpolations, or even that some terms have been replaced by synonyms or hyponyms. Each of these formally distinct patterns at the character level can require its own algorithms or at least its own parameters for detection, which is why they must first be clearly distinguished, described and categorized. The search can only be started when it is known what the desired results should look like and by which characteristics they can be identified. Thus, while the analogue approach is based on encyclopedic content-related knowledge and reading, the digital approach is based on precise observation and description of formal similarities in the desired results and their translation into specific search strategies.

Starting from the question at hand, the digital method must be researched, tested and explored step by step on the basis of the data and the desired results. This is indeed a longer, presuppositional and iterative process, but in turn, and this is important, one that contributes significantly to the gain in insight and that is part of the process of knowledge construction itself. It ultimately leads to an ever deeper understanding of the historical object (i. e. the texts, contents, contexts and contemporary practices), the data and the method. All three – subject, data, and method – must be examined in greater detail in order to be able to coordinate them as well as ultimately develop a viable procedure. In this way, with each cycle and the evaluation of the respective results, one gradually develops an increasing understanding – which is actually nothing other than a hermeneutic circle

Methoden und Kritiken digitaler Geschichtswissenschaften" (1 – 3 March 2021, Göttingen), see Althage, Abstract: Trends und Entwicklungen der historischen Fachkommunikation im Spiegel von H-Soz-Kult, in: Digital History. Konzepte, Methoden und Kritiken digitaler Geschichtswissenschaften, 2. 3. 2020, https://digitalhist.hypotheses.org/991. A publication of the paper is currently in the planning stage.

between data, method, individual results, and the larger picture. Each cycle and its results open up new perspectives that contribute to a better understanding of the whole and vice versa. That means that the development of digital methods for historical research is ultimately also genuine historical work.

This complex development process takes place only in small steps, partly because each step often breaks new ground and at the same time requires very different skills. But, as has been shown, issues of data availability, data quality, and the interoperability of data also play an important role. That said, this procedure does not automatically lead to any conclusions. After all, the algorithms only generate lists of potential candidates, which in turn must be evaluated. The assessment and evaluation of the results provided require the same expertise as has been described above for the analogue approach, in order to identify the significant results and, if necessary, to refine the algorithms. It is important to emphasize this again: the decision as to whether a specific result is relevant or not, and in our case whether it is a significant, intentional form of text re-use or not, is ultimately up to the human researcher. This is done, firstly, in how he or she has incorporated certain decisions into the algorithms (in our case, for example, at the selection stage), and, secondly, in his or her evaluation of the results. Either way, the process leads to results that still have to be interpreted and classified by the human being.

However, once established, we can scale these digital approaches to large amounts of sources, and reuse our algorithms to process a wide range of data over and over again. In doing so, we are also able to adjust them based on their results and to refine them with every application. In the process, we learn more about the data and about the historical information that they represent, as well as about the methods that we apply. Thus, the digital approach is quite different from analogous practices. The same applies to the results that we achieve in this way. But we are still only at the beginning when it comes to understanding how we can apply those digital methods to historical sources, how they may consolidate, change and enhance historical information during the process and thus contribute to the production of new insights into history.[136]

From this contrasting logic, it also follows that both approaches in principle lead to different types of results. This means that the results of digital methods cannot easily be equated or compared with the results of the analogue method. They are just different. While, in the given context, analogue results are characterized by similarities in content, the digital results are characterized by formal similarities between the texts, which can manifest themselves in very different ways, but are always systematically recorded for the entire body of text according to the same set

136 The question of the influence of digital methods on historical research and knowledge production compared to the analogue approach is also raised by, for instance, Jörg Wettlaufer, Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern, in: Zeitschrift für digitale Geschichtswissenschaften 1. 2016, http://www.zfdg.de/2016_011.

of rules. For the same reason, it is not really possible to simply transfer analogue methods and approaches into digital methods and, as our example has shown, to answer a question conceived in analogue terms ("How is the Bible used in the chronicle of the Crusades?") in the familiar way by means of digital methods.

But what does this mean for our question about the role that the Bible played in the chronicle of Raymond d'Aguilers, and how was it used here? A definitive answer to this question cannot be given here, as it depends on the interpretation of a wide range of different aspects. However, in order to be able to provide this interpretation, we first require a thorough study along with evidence that draws on more than individual exemplary passages, as has been the case so far. Although there seems to be a broad consensus in current scholarship about the significance of biblical textual references in the historiography of the First Crusade, there is a remarkable desideratum regarding the identification and documentation of concrete scriptural quotations beyond exemplary studies that take the Bible in its entirety into account.

This paper has shown that the digital method can be used to analyze comprehensive texts in their entirety and find new, previously unknown evidence. Alternative, more solid scriptural references could be offered for passages that have already been identified as Bible quotations. Other passages were identified as quotes from the Scripture for the first time. Two of the three examples discussed above add a new level of meaning to the respective passage, and bind the text – at least based on these results – even more clearly to the Psalms. But this may also be simply because the analysis has been limited so far to this Book and the Revelations of John. Once established, this method can also be applied to other books of the Bible and especially to other Crusade chronicles as well as to Middle Latin texts in general. This opens up new, more solid research perspectives for the study of the Crusade chronicles and the specific influence of the Bible on their accounts, as well as on the reception of the Bible in Middle Latin chronicles in general.

Aside from a methodological and data-critical point of view, the digital approach also raises new philological and cultural-historical questions: it has shown in particular that we need a much more precise understanding of what we mean by the use of the Bible in the context of Crusade chronicles as such. We need not only to clarify which contemporary versions of the Scriptures (or their inclusion in liturgy) were referenced by the chroniclers, but also to determine what exactly constitutes an intentional Bible quotation. The digital method forces us, considerably more than the analogue method, to investigate systematically the historical practice of re-using passages from the Bible. It also impels us to define clearly what we actually assume to be an intended use of the Bible – which we must also describe formally in order to operationalize it for the use of digital methods. To what extent was a priest like Raymond d'Aguilers able to quote large passages by heart, and how close was he in textual terms? To what extent were narrative forms, linguistic style or simply episodes adopted – without this being a syntactical re-use per se? This historiographical groundwork and a much clearer definition of the expected results are fundamental to using digital methods to

advance our understanding of the use of the Bible in the chronicles of the Crusades, which can also provide new research questions for more general historical research.

In conclusion, this paper has demonstrated how differently digital and analogue methods work, and how they differ epistemologically, which to a certain extent may also explain the discrepancy between expectation and reality in this matter. While the analogue is anchored in the "realm of meaning," the digital approach is based in the "realm of signals." This leads to very different approaches and ultimately to very different results. While the analogue method is based on the knowledge and intuition of the researcher, the digital method starts from the formal properties of the signifiers, on the exact basis of which, however, it can be applied to very large amounts of data at the same time. In the end, the digital method, in our case, does not replace the analogue method nor can it reproduce it; it complements it. By doing so, it can open up new perspectives in a given field of research and considerably advance its study. Already at an early stage of application, the case study on the influence of the Bible on the Chronicles of the Crusades has led to new discoveries that significantly exceeded the current state of knowledge, and revealed gaps that we have not yet seen, at least not in this clarity. It is obvious that the further adaptation of the method for this field of research will provide new foundations and new insights. It therefore seems highly advisable to us to continue along this admittedly challenging path.

Prof. Dr. Torsten Hiltmann, Humboldt-Universität zu Berlin, Institut für Geschichtswissenschaften, Friedrichstraße 191 – 193, 10117 Berlin
E-Mail: torsten.hiltmann@hu-berlin.de

Prof. Dr. Jan Keupp, Westfälische Wilhelms-Universität, Historisches Seminar, Domplatz 20 – 22, 48143 Münster
E-Mail: jan.keupp@uni-muenster.de

Melanie Althage, Humboldt-Universität zu Berlin, Institut für Geschichtswissenschaften, Friedrichstraße 191 – 193, 10117 Berlin
E-Mail: melanie.althage@hu-berlin.de

Philipp Schneider, Humboldt-Universität zu Berlin, Institut für Geschichtswissenschaften, Friedrichstraße 191 – 193, 10117 Berlin
E-Mail: philipp.schneider.1@hu-berlin.de