Repetitive Research: A Conceptual Space and Terminology of Replication, Reproduction, Re-Implementation, Re-Analysis, and Re-Use in Computational Literary Studies

Christof Schöch (schoech@uni-trier.de)
University of Trier

Research Article

Keywords: Replication, Reproduction, Re-Implementation, Re-Analysis, Re-Use, Terminology, Computational Literary Studies, Open Science

Posted Date: March 8th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2657846/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at International Journal of Digital Humanities on November 6th, 2023. See the published version at https://doi.org/10.1007/s42803-023-00073-y.
Repetitive Research: A Conceptual Space and Terminology of Replication, Reproduction, Re-Implementation, Re-Analysis, and Re-Use in Computational Literary Studies

Christof Schöch$^{1*}$

$^{1*}$Trier Center for Digital Humanities & Department for Computational Linguistics and Digital Humanities, University of Trier, Am Universitätsring 15, Trier, 54296, Germany, https://orcid.org/0000-0002-4557-2753.

Corresponding author(s). E-mail(s): schoech@uni-trier.de;

Abstract

This article is motivated by the ‘reproducibility crisis’ that is being discussed intensely in fields such as Psychology or Biology but is also becoming increasingly relevant to Natural Language Processing and Digital Humanities. Using the phrase ‘repetitive research’ as an umbrella term, and with the objective to provide clarity and help establish best practices in this area, this article focuses on two issues: First, the conceptual space of repetitive research is described across five key dimensions, namely those of the research question or hypothesis, the dataset, the method of analysis, the team involved, and the results. Second, building on this new description of the conceptual space and on earlier terminological work, an easily understandable terminology for recurring scenarios of repetitive research is proposed. For each scenario, its typical purpose and added value in the research process are discussed, the requirements for enabling it are described, and it is illustrated using examples from the domain of Computational Literary Studies. The key contribution of this article, therefore, is a proposal for a transparent terminology underpinned by a systematic model of the conceptual space of repetitive research.

Keywords: Replication, Reproduction, Re-Implementation, Re-Analysis, Re-Use, Terminology, Computational Literary Studies, Open Science
1 Introduction


The aim of this contribution is to systematically describe, using a multi-dimensional conceptual space and a set of descriptive terms for recurring scenarios within that conceptual space, a specific research practice in the growing subfield within the Digital Humanities called Computational Literary Studies (CLS), a mode that, broadly defined, could be be termed repetitive research (RR).

Firstly, the mode of research I’d like to describe is one that repeats, in the sense that studies following this mode actively seek to align their research questions or hypotheses, their datasets and/or their methods of analysis, with research practiced and published earlier. This is done with the explicit aim to approximate an earlier study, but conscious also of the fact that perfectly identical repetition is virtually impossible to achieve. In many cases – and this might be a particularity of this kind of research in the humanities, where influential research may remain relevant for many decades –, this also means that the earlier research that is to be repeated has been practiced within the non-computational, or at least in the non-digital, paradigm. Any attempt at exact repetition can in fact only, in such a scenario, result in a reenactment and an approximation of earlier research.

Secondly, this mode of research is repeatable, in the sense that it (typically) makes all the efforts it can to provide the data, code, and explanatory information that make it possible for others, at a later point in time, to perform the same (or very similar) research again. Of course, an earlier study that aimed to be repeatable will be more amenable to a later study repeating it than one that did not consider this to be an important goal. In that sense, not only does repeatability foster repetition, but also the other way around; both issues are in fact two sides of the same coin. Actually, most research that repeated earlier research also aims to be repeatable, because this kind of repeatability that also involves transparency and sustainability, is something that researchers come to value when they practice the hardships of repeating earlier research.

In this way, this kind of research is located between past and future: a (never identical) reenactment of past research, and an invitation for (never identical) further reenactments in the future. This mode of research is practiced with the conviction, or at least in the hope, that this cycle of repetitions is not a futile treading in the same place, but a productive, insightful upwards spiral. Indeed, RR has important functions in the research process and makes essential contributions to the production of knowledge that are closely related...
to core values of the scholarly enterprise such as reliability, trustworthiness, transparency and sustainability, as discussed further below.

In terms of the scope of this contribution’s argument, the examples and use cases employed in the present contribution for illustration and clarification pertain to CLS. Most of the arguments with respect to the role of data and methods in RR are valid primarily for research that operates with datasets that represent the domain being investigated as well as with code-based implementations of the method of analysis. This means that the basic ideas are likely to be applicable only to other fields within the Digital Humanities that use algorithmic methods applied to evidence available in the form of digital data, sometimes collectively called Computational Humanities. Such studies are most amenable to RR but clearly do not represent all or even most of research in the Digital Humanities.

The remainder of this paper is structured as follows. In section 2, a trailblazing example of RR is provided as a first way of approaching the issue. Then, section 3 briefly motivates the paper, mostly by way of a defense of RR as a useful and indeed necessary answer to the lack of transparency, trustworthiness and corroboration that hinders much of present research. In section 4, earlier work on the issue is discussed, with a focus on the terminological issues surrounding RR. In section 5, a solution to this situation is proposed, based on a systematic description of the issue as a multi-dimensional semantic space. From this space, one can derive not only an understanding of the structure of RR, but also a clear and well-motivated terminology for recurring scenarios within RR. In section 6, several such scenarios are defined, discussed and illustrated. By way of a conclusion, in section 7, some of the benefits and limitations of the proposed terminology, but also of RR in CLS more generally, are discussed.

2 Example

In 2015, Geoffrey Rockwell gave a talk at the University of Würzburg titled “Replication as a way of knowing” (Rockwell, 2015). In this talk, he presented work he had done together with Stéfan Sinclair on reenacting a classic, very early, stylometric study by Thomas C. Mendenhall. This talk is an early example of the idea of repeating research, not just as a practice, but as a programmatic research principle, in the domain of CLS.

The story starts in 1887, when Thomas C. Mendenhall published an article in Science titled: “The Characteristic Curves of Composition” (Mendenhall, 1887). His fundamental idea was that it was possible to identify the author of a text by the characteristic distribution of word lengths in his or her texts. For example, Figure 1 shows the word length distribution plot that Mendenhall obtained for the first 1000 words of the novel Oliver Twist by British 19-century author Charles Dickens.

In their repeating study, Stéfan Sinclair and Geoffrey Rockwell started out with the idea to implement Mendenhall’s study once more, but using digital
texts and some simple algorithms (Sinclair and Rockwell, 2015). When they did so, they obtained nearly identical results (Figure 2). Then, however, they repeated the same analysis for the entire novel and now obtained a distribution that, despite being recognizably related, clearly deviated from Mendenhall’s results as an effect of the much longer text being analysed (Figure 3).

Many of the typical properties of repetitive research are already present in this seemingly simple study. The starting point is a more or less famous example of early quantitative (in this case of course non-digital) research. We find a close (but most likely not perfect) alignment in terms of the data: the digital edition Rockwell and Sinclair used (from Project Gutenberg) is probably even based on a print edition that is roughly contemporary, and therefore quite similar, to the one Mendenhall used. Also, given that this repetition involves crossing the analog-digital domain, it makes use of course of a functionally similar, but certainly not identical implementation of the method of determining the number of words of any given length: while Mendenhall must have tabulated and visualized this by hand, Sinclair and Rockwell have of course implemented the same process algorithmically, in Python. To summarize it along the lines to be defined below: while Sinclair and Rockwell’s research...
question is identical to Mendenhall’s, the dataset and the method can best be described neither as identical, nor as entirely different, but as similar or closely related. The team performing the repetition is clearly entirely unrelated to the author of the original study and the immediate numerical results are again similar, but not identical.

In addition, we can notice efforts to approximate the earlier results, but also an attention for the slight and inevitable deviations. Note in Figure 2, for example, that there seem to be a few words with 13 letters that Rockwell and Sinclair found but Mendenhall did not; an effect of differences in the underlying text, of tokenisation, or of an error in how Mendenhall established the word lengths? It becomes clear that the authors employ this research strategy not so much as a way of checking the work of Mendenhall for flaws or errors (as would be the case in a strict replication study primarily designed as a verification of quality and integrity, but in order to better understand his approach and, in the process, also think about their own computational methodology. One can also see a display of the advantages of the digital paradigm in terms of scale: once the code works with 1000 words, to which Mendenhall limited his investigation for obvious pragmatic reasons, it is trivial to expand the analysis to entire novels. This second, modified step in their approach, relative to Mendenhall’s, places Sinclair and Rockwell’s study even firmer away from a strict replication. Given that replicating was not the goal, and that Sinclair and Rockwell’s results were quite similar to those of Mendenhall, one may still say that they confirmed Mendenhall’s findings.
What Rockwell and Sinclair do not do, however, is follow Mendenhall’s further claims and actually try to perform authorship attribution with this kind of data. C.B. Williams had, in fact, in 1975, determined that this does not work very well. The primary reason for this is that word length appears to be at least as strongly affected by form (notably, by the distinction between texts in verse and in prose) as it is by authorship (Williams, 1975). So while Sinclair and Rockwell confirm that Mendenhall had worked with considerable accuracy, when re-implementing his method, Williams had shown that the more general methodological conclusions regarding authorship attribution that Mendenhall had derived from his results do not hold up to scrutiny.

Finally, Sinclair and Rockwell understood very well the link between repeating earlier research and making one’s research repeatable. Therefore, they made their own research processes easily repeatable. In this case, their strategy is to provide a Jupyter notebook that not only contains the code and the required data, but also some explanatory prose that helps others understand how the study works and to easily run the code themselves.\footnote{See: https://github.com/sgsinclair/epistemologica.}

### 3 Motivation

One may ask, however, why a more systematic approach to RR, in terms both of practice and theory, appears to be particularly useful today. The first motivation is the “reproducibility crisis” in a range of academic fields, as discussed...
The term has appeared around the year 2015 and has served to highlight findings from a study conducted by the Center for Open Science in Charlottesville (Open Science Collaboration, 2015). In this meta-study, the authors attempted to reproduce the results of 100 papers from several key journals in psychology, all first published in 2008. They were able to do so in just 40 percent of the cases. There are many good reasons for this: Not only is it hard to avoid any and all so-called “questionable research practices” threatening reproducibility. The studies were also actually gathering relevant empirical data once more from scratch, not just checking the integrity of datasets and code provided. Finally, there are arguments that a certain level of non-reproducibility is to be expected as a matter of course (Bird, 2021). Still, a fundamental principle of the scientific method appears questioned by these results. The argument here is not that research as practiced today is not connected to the past and is not open to the future. Of course, we all read, learn from and quote earlier research extensively and hope that our own research will be read and quoted by others in the future. Research as practiced today is also not all unsound. In fact, it is an open question whether a field like CLS can and should even be held to the kinds of standards relevant in the natural sciences (see, e.g., Peels and Bouter, 2018; Peels, 2019; Penders et al, 2019). Different answers probably apply to different areas within the Digital Humanities. But I believe that RR can be an important aspect to make sure our research is well-integrated into the research tradition and will be useful, trusted and well-received in the future.

The second motivation, more specific to the Digital Humanities, are fears of a disconnect between the Digital Humanities and the established Humanities. There is an ongoing discourse expressing fears of a disconnect or a mismatch between the contextualizing, interpretive, detail-oriented modes of research typical of disciplines in the Humanities, on the one hand, and formalizing, quantitative approaches in the Digital Humanities, on the other (e.g. Marche, 2012; Eyers, 2013; Da, 2019). One of the main reasons for this fear is the assumption that, for instance, literary texts, historical documents or cultural artefacts and computational methods are simply not a suitable for elucidation through formal modeling, algorithms and computational analysis, because they are complex, semiotic, non-deterministic, contextual artefacts. Inversely, this would means that CLS and other similar subfields with the Digital Humanities that ultimately rely on counting surface features, can only ask and answer a specific set of questions (like authorship attribution) but are unable to make meaningful contributions to established, qualitative research in these fields. This is of course not an undisputed position and the paradigm of repetitive research, in particular, with its explicit focus on modeling and operationalisation across an earlier and a later study, can be a useful bridge between established, qualitative research on the one hand, and computational or quantitative research, on the other.

---

2Please note that the terminology varies widely, both between and within fields, and a systematic terminology will only be introduced in section 6.
The third motivating factor is the most closely connected to CLS: it is the 2019 paper by Nan Z. Da in the influential journal *Critical Inquiry*, with the title “The Computational Case Against Computational Literary Studies” (Da, 2019). The author basically argues that computational, quantitative approaches are fundamentally unsuited for investigations into literary texts. And she argues that the selection of studies she looked at either had statistically solid results that were meaningless, or seemingly meaningful results that were not statistically sound. This paper is highly problematic and has been commented on and criticized extensively, for a large number of good reasons. However, it also points to some serious and relevant challenges for the field of CLS: notably, the difficulty to reproduce work in this field, starting with issues of access to data and code, but also concerning questions of lacking reporting standards, limited scholarly recognition, and missing community commitment and capacity that would all be needed to foster a culture of RR in CLS and beyond.

4 Earlier work

Moving on from an understanding of the general relevance of RR, it appears useful to now consider more closely two key issues in selected contributions to the rather large body of existing conceptual work on the issue: first, regarding terminological discussions around the definitions of repeatability, reproducibility, reproduction, replication, re-analysis and others; and second, regarding the purposes, functions and epistemological added value of RR in the research process.

The disciplinary focus of the present paper notwithstanding, the principle of RR is of course not limited to this field, quite the contrary. The “reproducibility crisis” (Baker, 2016) has first become very visible in fields such as social psychology, biology and medicine (e.g. Open Science Collaboration, 2015; Freedman and Inglese, 2014; Hunter, 2017) and has begun to become a major issue in Artificial Intelligence, Natural Language Processing and Linguistics (e.g. Hutson, 2018; Cohen et al, 2018; Belz et al, 2021; Berez-Kroeker et al, 2022), three fields closely related to CLS. In recent years, the issue has also begun to be discussed in the (Digital) Humanities (e.g. KNAW, 2018; Peels and Bouter, 2018; Schöch et al, 2020), with the most prominent and controversial work applying replication in CLS certainly being the one, already mentioned, by Da (2019). However, the debate about the definition of relevant terms has occurred mostly outside the domain of CLS.

The terminological situation, which can certainly be described as complex and confusing, is in itself another motivating factor for this paper, but earlier efforts to get to grips with the conceptual space and the terminology also present a substantial learning opportunity. The fact that the terminology has developed over time and often in parallel in different fields, is part of the reason why it has become so unwieldy. Researchers have variously noted this,
for example Goodman and colleagues, who note that “the language and conceptual framework of ‘research reproducibility’ are nonstandard and unsettled across the sciences” (Goodman et al., 2016, 1). Cohen and colleagues remark that, in addition to using terms with varying definitions, “it is not uncommon to see reproducibility and replicability or repeatability used interchangeably in the same paper” (Cohen et al., 2018, 156). Hans Plesser describes his contribution as “a history of a confused terminology” (Plesser, 2018, 1). As Cohen and colleagues also note, “[t]his lack of consensus [in] definitions related to reproducibility is a problem because without them, we cannot compare studies of reproducibility (Cohen et al., 2018, 156).

Hans Plesser has summarized the “Claerbout terminology”, in which reproducing means the exact repetition of earlier research (where data, implementation and results should be identical), whereas replicating means using a new implementation and reaching similar conclusions (Plesser, 2018). Plesser also explains that this is at odds with terminology in the sciences, for instance in Chemistry, where repeatability is understood as exact repetition within a lab under identical conditions (to measure within-run precision), and reproducibility means repetition of the same experiment, but performed by a different person in a different lab with different conditions (to measure between-run precision, so robustness). Since 2020, the ACM uses repeatability for the exact repetition of the same experimental setup by the same team, reproducibility for repetition of the same experimental setup by a different team, and replicability for repetition of a different experimental setup by a different team, swapping the latter two terms relative to earlier ACM usage (ACM, 2020). The ACM terminology, however, does not differentiate between several distinct aspects of the experimental setup, notably concerning the dataset and the method of analysis.

The contribution by Prasad Patil and colleagues is helpful not so much for the clarification of the terminology, but rather for a clear understanding of the conceptual space around reproducibility and replication (Patil et al., 2016). They address this issue by identifying a relatively large number of features (around a dozen) that can be used to describe the relationship between an earlier, original study and a later, repeating study. However, they do not attempt to directly relate the terms they propose to these descriptive dimensions.

Another contribution that explicitly aims to achieve a clarification of the terminological confusion is by Goodman et al. (2016). Their contribution is quite systematic in the way it considers key dimensions of RR, such as data, method and results. They propose research reproducibility as the cover term and note that an essential difference between different kinds of such research practices is whether or not they use of new evidence, that is different or additional data. They also note that the terms reproducibility and replication are frequently used to mark this difference, but without consistency as to which of the two terms is used for which scenario. Their proposal is to only use the term reproducibility, but to characterize it further depending on the specific

---

3Other authors following this understanding are Branco et al (2020); Peng (2015).
scenario, in order to distinguish *methods reproducibility* (the exact repetition of earlier research), *results reproducibility* (using additional data to corroborate earlier results) and *inferential reproducibility* (where the methods of analysis and the conclusions drawn from results may be different). As far as I can see, this terminology has not been adopted widely.

Another terminological tradition, proposed for example by Drummond (2009) and explained very clearly by Huber and Çöltekin (2020) in the context of Natural Language Processing (NLP), is the following: “We use the term replication to refer to the activity of running the same code on the same dataset with the aim of producing the same (or sufficiently similar) measurements presented in the original paper. We use the term reproduction to refer to the activity of verifying the claims with experimental settings that are different from the ones in the original paper. For NLP experiments, this typically means re-implementation of the method(s) and the use of different datasets and/or languages” (Huber and Çöltekin, 2020, 5604). In another contribution to the terminological debate within NLP, Cohen et al (2018) follow this terminological tradition, but use *reproducibility* also as a relatively broad cover term.4

What also becomes clear from the research literature is that there are not only different kinds of RR, but that they also have different functions in the research process or make different contributions to the production of knowledge. Indeed, these functions and contributions are closely related to core values of the scholarly enterprise such as reliability, trustworthiness, transparency, sustainability and progress. As formulated by Freedman and Inglese: “Research advances build upon the validity and reproducibility of previously published data and findings” (Freedman and Inglese, 2014). As a consequence, functions of RR mentioned in the research literature include the “verification of a previously observed finding” (Gomez et al, 2010); “enhanc[ing] the reliability of research” (Gomez et al, 2010); the “corroboration” of earlier evidence, claims or findings (Babin et al, 2021; Goodman et al, 2016); converting “tentative belief to accepted knowledge” (Berthon et al, 2002), also in a Bayesian perspective of accumulating evidence (Sprenger, 2019); establishing the degree of robustness and/or generalizability of results Goodman et al (2016); and verifying the quality and integrity of research, for example with respect to identifying “(i) fraud, falsification, and plagiarism, (ii) questionable research practices, partly due to unhealthy research systems with perverse publication incentives, (iii) human error, (iv) changes in conditions and circumstances, (v) lack of effective peer review, and (vi) lack of rigor” (Peels, 2019, 1). In summary, one may say with the authors of the Open Science Collaboration report: “Replication can increase certainty when findings are reproduced and promote innovation when they are not” (Open Science Collaboration, 2015, 943). Being aware of these important functions may also help address what has been called a ”publication bias” in replication studies, where (among other effects)

4Earlier papers from outside the NLP domain and also using this terminology are Berthon et al (2002) and Gomez et al (2010), the latter using *re-analysis* as their third term with a definition that is similar to the one proposed below.
research corroborating earlier work has a lesser chance of being submitted and published than work revising earlier conclusions (Berinsky et al., 2021).

Several conclusions can be drawn from this brief discussion: First of all, given that NLP is a field closely related to CLS, and for further reasons explained below, I propose to accept the terminology in the tradition of Drummond, where replication refers to the exact repetition of earlier research, adding further terms as necessary to describe research scenarios that differ, in some defined way, from replication. Second, it appears useful to use a limited number of relevant dimensions to describe the conceptual space of RR, notably research question, dataset, method of analysis, team, and results. And third, the discussion of the dimensions of the conceptual space and the distinct scenarios of RR should not only describe them in terms of their position in the conceptual space, but also in terms of their functions or purpose in the research process.

5 The conceptual space of RR

In this section, I would like to propose a typology of and terminology for RR that is based on a simplifying but useful multi-dimensional description of the conceptual space of RR. Describing the conceptual space in a systematic manner helps clarify which aspects of RR are relevant and what the meaning is of any term proposed in the terminology, because its meaning can be described by specifying the subspace in the overall conceptual space that is covered by the term. It also shows the relationships, distances and differences between the meaning of specific pairs of terms and, of course, the relationships between studies that can be described using those terms.

The fundamental assumption behind my description of the conceptual space is that any research can be described, not fully of course, but fundamentally and usefully in the context of RR, by five dimensions:

- (Q) The key research question being studied (or the key hypothesis or claim to be verified)
- (D) The dataset used (or more generally, the empirical basis of enquiry)
- (M) The research method employed (including its implementation in a code-based algorithm or tool)
- (T) The team performing the research (including, of course, the case of a one-person team)
- (R) The result of the research (and the claims or conclusions supported by the results)

In addition, this model of the conceptual space assumes that the relationship between an earlier study that is being repeated and a later study that repeats it can, for any of these five aspects, be described as corresponding to one out of four simplifying categories:

- (1) Identical
- (2) Similar (closely related, but not literally identical)
• (3) Dissimilar (not closely related, but not entirely unrelated)
• (4) Unrelated, contradictory or not applicable

It is true, of course, that assuming just five dimensions and four possible values for each dimension is a simplification. It is also true that it may not always be possible to clearly distinguish between a scenario, for instance, where the dataset is similar and one where the dataset is dissimilar to that of an earlier study, as these are terms expressing a gradient rather than a category. However, considering that this conceptual space defines no less than $4^5 = 1024$ theoretically possible positions, I would argue that it provides more than enough differentiation and clarity to support the definition of a clear descriptive vocabulary that allows to characterize any repeating study with relatively little ambiguity. For some recurrent scenarios, we can define labels that function as shortcuts. For others, we may prefer to describe their exact position in the conceptual space. This gives us three ways to describe an instance of RR: We can use one of the terms proposed below, if one is applicable, for recurring scenarios of RR. We can describe a particular instance using a verbal description containing the five dimensions and four values. And we can express such a verbal description in the condensed form of a vector-like representation.

By way of an example, we could describe a study repeating earlier research in the following way: it pursues an identical research question or hypothesis (Q=1) using dissimilar but not unrelated data (D=3) and a similar but not entirely identical method, such as a re-implemented algorithm (D=2), was performed by an entirely unrelated team, for example at an institution in another country (T=4) and obtained similar, though not entirely identical, results (R=2). The vector-like representation of this scenario would therefore be $RR_{(Q,D,M,T,R)} = (1,3,2,4,2)$. With respect to the set of terms proposed below, this scenario would fall into the relatively broad category of follow-up research. One could further deduce from the description that the repeating study likely corroborated the earlier study because with additional relevant data, it came to a very similar conclusion to the same research question. The more different the data, the stronger the corroboration, because the results can be understood to be more robust and supported by more evidence.

6 A terminology of RR

The question that arises from the description of the multi-dimensional conceptual space is how some of the recurrent scenarios within this semantic space can be delimited and labeled. This is likely to always remain controversial, but once the conceptual space is clearly defined, the terminology in fact becomes somewhat less of a matter, because it is always possible to define terms with respect to the conceptual space, or to describe a specific scenario independently of a given term. In this perspective, the terms become convenient shortcuts that have their usefulness and importance, but that are ultimately not essential.

As mentioned, $4^5 = 1024$ possible combinations or different scenarios of RR are clearly more than we would care to qualify using individual terms. One
part of the terminological complexity in the field stems from the fact that this conceptual space can be divided up in multiple ways; another stems from the fact that the inventory of semantically suitable terms for different areas of the conceptual space contains many terms that are very similar to each other and, as a consequence, have been used interchangeably at different times and in different fields (see section 4).

However, if we set aside for a moment the dimensions four and five (team and results) – important descriptive aspects that do not require, however, to be included in distinctions between recurring scenarios of RR –, we obtain a three-dimensional conceptual space that can be visualized as a cube and, with its four possible values in each dimension, distinguishes just $4^3 = 64$ scenarios. Within this cube, we can define specific sub-spaces and provide terms for them. I propose to distinguish the following terms, as shown in Table 1 with a short definition and their vector shortcut, to provide an overview before describing them in more detail in the following sections:

<table>
<thead>
<tr>
<th>Term</th>
<th>Verbal description</th>
<th>Vector (Q,D,M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication (of research)</td>
<td>Exact repetition, with identical question, data and method. As for all scenarios, team and results may vary.</td>
<td>(1,1,1)</td>
</tr>
<tr>
<td>Re-implementation (of the method)</td>
<td>Approximate repetition, with identical question and identical data, but using a similar method.</td>
<td>(1,1,2)</td>
</tr>
<tr>
<td>Reproduction (of results)</td>
<td>Approximate repetition, with identical question and method, but applied to similar or dissimilar data.</td>
<td>(1,2-3,1)</td>
</tr>
<tr>
<td>Re-interpretation (of results)</td>
<td>Approximate repetition, with identical data and method, but used for a similar or dissimilar question.</td>
<td>(2-3,1,1)</td>
</tr>
<tr>
<td>Re-analysis (of data)</td>
<td>Related research, with identical data, but same or similar question and similar or dissimilar method</td>
<td>(1-2,1,2-3)</td>
</tr>
<tr>
<td>Re-use (of data)</td>
<td>Related research with identical data, but employing a different or unrelated method for a different or unrelated question.</td>
<td>(3-4,1,3-4)</td>
</tr>
<tr>
<td>Re-use (of the method)</td>
<td>Related research, with similar, dissimilar or unrelated question; similar, dissimilar or unrelated data, but identical or similar method or code.</td>
<td>(1-3,2-4,1-2)</td>
</tr>
<tr>
<td>Follow-up research</td>
<td>Related research in a relatively broad sense, where question, dataset or method can be either identical, similar, different or unrelated, but not all three can be all identical, similar or unrelated.</td>
<td>(1-4,1-4,1-4)</td>
</tr>
</tbody>
</table>

Table 1 Overview of terms proposed for specific subspaces of the conceptual space of repeating and repeatable research. The vector concerns only research question (Q), dataset (D) and method of analysis (M), whereas the issues of team (T) and results (R) are left to be specified in addition to each term.
These scenarios fall into three groups: only replication (of research) implies an exact repetition of earlier research, whereas re-implementation (of the method), reproduction (of results) and re-interpretation (of results), in which two dimensions are identical and the third deviates, can be understood to be approximate repetitions, due to the clear and close relationship that they establish between an original study and a repeating study. The remaining scenarios, re-analysis (of data), re-use (of data), re-use (of method) and follow-up research describe scenarios that can still be called related research, but are further removed from the original study than the previous group, with only one or even no identical dimension. Finally, in the scenarios where all three primary dimensions are dissimilar or unrelated, there is no clear relationship at all anymore between an earlier and a later study, so that they fall outside the scope of RR and are omitted here entirely. The exact subspace each term covers may be subject to debate, of course, a debate that is, however, supported by the explicit conceptual space that underpins the terminology.

Note that these terms, as defined in Table 1, are meant to describe the relationship between an earlier, original study and a later, repeating study. However, if the concern is to characterize a given study with respect to the degree to which it enables repetition, then the corresponding terms expressing can be used, that is replicability (of research), reproducibility (of results), re-interpretability (of results), re-analysability (of data) and re-useability (of data or code). More than one term may be used to describe a given study, in this case, for example to say that the re-usability of its data is high (e.g. because the data is encoded in a widespread data format and is well-documented), but the replicability of the research as a whole is expected to be low (e.g. because the code has not been provided).

The following discussions of the scenario describe each scenario’s location in the conceptual space, argue what constitutes a successful repetition in that scenario, comment on the functions or purpose of this scenario in the research process, name the requirements for enabling this scenario and provide one or several examples of this scenario from CLS.

6.1 Replication (of research)

Within the terminology proposed here, the term replication (of research) designates practices of RR in which the research question, the dataset and the method of analysis of the repeating study are all identical to the original study. The term can be used irrespective of the team and the results. The term replication is preferred for this configuration of strict repetition without any modification because it is etymologically related to Italian ‘replica’ (copy, repetition) and Latin ‘replicare’ (to duplicate, to repeat) and because this is the accepted meaning of the term in NLP, a field closely related to CLS (as discussed in section 4).

Only if the results of the repeating study are strictly identical to the original study, can such a replication be said to be successful. All other outcomes would signal some flaw in the data and/or the code and would mean the
replication was unsuccessful. In this sense, the purpose of replication is to check the integrity and correctness of the code when applied to the data. Generally speaking, then, such a replication does not add new knowledge about the domain or research question but rather serves as a quality check. If the team performing the replication is identical or very similar, for example coming from the same research group, the function of this replication would amount to an internal check of the research as a whole, ensuring for example the completeness of the dataset, the functionality of the implementation and the correct match between data and code. If the team is different or unrelated, for example when a replication is performed as part of a peer-review process, the function of this replication would be quite similar, but would likely also be helpful in order to ensure the completeness of documentation that allows people outside the team to run the code on the dataset; in addition, it would serve to ensure the integrity of the research process in the sense of verifying that no manual manipulation has occurred in the original study that would affect reported numerical results or data visualisations.

Compared to other scenarios, strict replication, then, is more important for quality assurance than for advancing knowledge. The greatest epistemological value lies not in replication, but in subtle and controlled departures from the original study design, as described in the following scenarios. However, research could also purposefully enable such insight through enabling paths for deviation, for example by making it easy, in the code, to explore alternative choices with respect to selection of data, parameters for preprocessing, or use of specific measures or calculations.

The requirements for enabling replication (of research) are quite high, as the complete dataset and all code need to be available in a form that allows running the code without modifications or re-implementations. If data and code are available and can be run, then performing the replication itself is comparatively trivial. This usually requires some degree of documentation, although at minimum or in a first step, a replication can treat the code and data as a black box, focusing on the comparison of the results reported in the original study and those obtained by running the code on the data once more. However, as soon as something does not quite work as expected, documentation and insight into the code and data becomes crucial, of course. Also, sustainability is an issue here, when underlying libraries and environments become incompatible over time. More generally, see the reservations and challenges discussed by Arvan et al (2022) in NLP, broadly applicable also to CLS.

The study by Nan Z. Da mentioned above can be understood, in part, as a replication study, although it is not quite clear from the paper and the relevant Github repositories to what extent individual studies were actually replicated in their entirety. It appears that for the most part, either data or code, or both, were not available and a full replication of the research was therefore not possible. In any case, such complete replications have not been documented online.\footnote{In addition to the paper (Da, 2019), see the Github repositories at https://github.com/nan-da.} As mentioned, another scenario where strict replication
may happen is as a project-internal quality-assurance measure or as part of
the peer-review process. However, there would likely be no publicly available
record of this process (although this could be the case in an open peer-review
scenario). In any case, among the journals relevant to CLS, only the *Journal of
Computational Literary Studies* routinely expects code and data to be made
available when papers are submitted but does not, by default, include a full
replication step in the peer-review process (see JCLS, 2023).

6.2 Re-implementation (of the method)

This scenario involves an identical research question and an identical dataset,
but the use of method of analysis that is only similar to the original one.
This scenario can be termed *re-implementation (of the method)*. Again, this is
irrespective of the team or the results. The idea here is to use a similar method
to investigate the same question using the same data, as a way of verifying
whether or not it is possible to arrive at the same conclusions using a new,
but functionally-identical, implementation of the method of inquiry.

The requirements for enabling such a *re-implementation* of an earlier
method are similar to those for *re-analysis (of data)* described below, namely
that the dataset be made available. In addition, it is essential in this sce-
nario that the research question is defined with considerable precision, so that
exactly the same issue can be investigated. In case the method used in the
original study was not implemented algorithmically or the code is not made
available, a very detailed, step-by-step description of the method is required
for a close re-implementation.

The purpose of a *re-implementation* of the method using the same data
is to investigate the robustness of earlier results. If a non-identical but simi-
ar method (for example involving a related, but different statistical measure)
investigating the same features of the same data nevertheless produces the
same findings, then the results of the earlier research are corroborated. Simi-
larly, if the re-implemented method involves the selection of different features
in the same dataset, and still comes to the same results, this is again a
corroboration of the earlier results.

An example of a repetition that is clearly an example of a *re-
implementation (of the method)* is a review of Nicholas D. Paige’s book
*Technologies of the Novel* published by myself (Paige, 2020; Schöch, 2023).6
The author of the book has made the full dataset used for the book publicly
available and describes his analyses in detail in the book. However, he has not
provided the code used to analyse and visualize the data. Most of the repe-
titions that I performed therefore were attempts to reconstruct specific plots
contained in the book by developing code that would functionally approxi-
mate the analysis that the author must have performed to obtain the plots, in
order to verify that the plots are really based precisely on the data provided.
This is probably as close to a *replication* as one can get when the code is not

---

6 See, in particular, the Github repository accompanying the review at https://github.com/
christofs/paige.
provided, although it was rather a process of reverse-engineering the plots and hence already a re-implementation.

6.3 Reproduction (of results)

Another recurring configuration occurs when research question and method of analysis remain identical, with respect to the original study, but the dataset, instead of also being identical, may be similar or different, but not unrelated or irrelevant to the research question. Again, this is independent of the team and the results. The term that appears best suited to describe this configuration is reproduction (of results), because this is a frequent scenario and the term is well-established, albeit not always in this precise meaning. In addition, one may motivate it by its etymology, in the sense that it is derived from biological reproduction which, at least in the case of sexual reproduction, does not imply the production of organisms that are identical to their parents, but rather the production of somewhat modified organisms.

In this scenario, the function or purpose is to verify whether or not, with another set of similar or different, but relevant data, the results of the original study can be confirmed. If the reproduction of the results is successful, the results of the original study is corroborated in the sense of that the initial results, being valid across more than one dataset, are in this way shown to be valid more generally or more broadly than if they hold true only for one particular dataset. Such a success is also a confirmation that the method of analysis is robust across multiple datasets or that a theory or claim holds up even when it is tested on a different dataset (on theory testing, see Brendel et al (2021)).

The requirements for enabling reproduction (of results) are relatively high, in the sense that the implementation of the method needs to be available and sufficiently-well documented in order for the reproduction be able to re-use without a re-implementation. In addition, because it is important to be able to determine the exact degree of similarity between the dataset of the original study and that of the repeating study, in order to correctly interpret any differences in results, the dataset needs to be available or at least described in sufficient detail.

This scenario appears to be frequent in the sciences, for example in the famous case of the Open Science Collaboration cited above (Open Science Collaboration, 2015), where new but functionally equivalent data was obtained empirically in order to verify that the results obtained in the original studies could be obtained once more. In CLS, where we have very different conditions for obtaining new but functionally-equivalent data, studies that can be described as a reproduction (of results) occur especially in the context of the development and evaluation of methods. For example, in stylometric authorship attribution, new measures have typically been proposed using English-language corpora (classic cases being Burrows’ Delta and Zeta, see Burrows (2002, 2007)). Studies implementing these measures as closely as possible (given that Burrows described the measures, but did not provide a
code-based implementation), proposing new variants of the measures and/or evaluating them using additional datasets containing different literary genres and/or different languages, include Rybicki and Eder (2011); Evert et al (2017) (for Delta) and Craig and Kinney (2009); Schöch et al (2018) for Zeta. Outside of the development of measures and methods, my own study on sentence length in books by Belgian writer Georges Simenon, and in books by contemporary authors, can be understood as an attempt to reproduce the results obtained earlier using a non-digital but clearly quantitative approach by Richaudeau (1982): with very similar data and a very similar method at first, then with an extended but still very closely related dataset (Schöch, 2016).

6.4 Re-interpretation (of results)

The scenario labeled re-interpretation (of results) is somewhat particular in the sense that the core of the research (data and method) remains identical between earlier and later study, so that we could almost speak of replication, but the question pursued is different, in the sense that the same results may be seen in a different light or re-purposed for a different research question. In this sense, the success of such a re-interpretation does not lie so much in reaching the same conclusions, but on the contrary, this research can be said to be successful if new conclusions or claims can be obtained.

The requirements for this kind of repeating research are quite high, however, at least when the research is in fact repeated. However, a re-interpretation of results may also happen without running the analysis again, but based simply on the existing data and reported results.

Examples of this scenario are hard to identify and one might even question whether this is a recurring, important scenario of RR at all. It is included in the present typology rather for reasons of structural symmetry.

6.5 Re-analysis (of data)

The scenario I propose to call re-analysis (of data) differs from re-implementation (of the method) in that in addition to the method of analysis the repeating study not being identical to the original study’s method, the question being investigated can also diverge to some degree. In fact, a changing method of analysis may of course induce the research question to shift to some extent, whether intended or not by the researchers conducting the re-analysis. The term appears fitting because the data remains the same and the overall perspective is still closely related to the earlier study. In contrast to this scenario, I propose the term re-use (of data) (described in subsection 6.6) for a scenario where question and method are more clearly distinct from the original context.

The function of re-analysis (of data), where the research question remains the same, is primarily to examine the same question from a different methodological angle. If the re-analysis is successful in the sense of producing results supporting conclusions that are identical to those of the original study, then these results are corroborated and their robustness is confirmed. If the method
used in such a re-analysis differs from the original method only in the sense that it is functionally equivalent but implemented in a new way, then the same results are to be expected. If the results turn out to be different, the re-analysis can be said to have been unsuccessful, pointing, however, to a potential flaw not only in the original study, but possibly also in the re-implementation of the study. Only once this can be ruled out would a differing result point to a potential flaw in the original study’s implementation of the analysis.

The requirements for enabling re-analysis (of data) are clearly lower than for replication (of research), because strictly speaking, only the dataset needs to be available in identical form, whereas the other aspects of the research will deviate from the original study in any case. It can be noted that providing data (and code) with publications is becoming increasing common, and sometimes even expected, in CLS, as evidenced for example in the submission guidelines of the Journal of Computational Literary Studies.

The review of Paige’s Technologies of the Novel, mentioned in subsection 6.2 as an example of reimplementation (of the method), also contains examples of a re-analysis (of data): In some cases, instead of attempting to produce essentially identical plots, I also aimed to develop plots that would more accurately make the limitations of or uncertainties in the data visible, or added statistical tests of significance for the same purpose. Even though the code and therefore the results deviated purposefully and functionally from the original study, the issue investigated remained basically the same.

6.6 Re-use (of data)

For the scenario where the dataset used in a later study is identical to an earlier study, but the research question and method of analysis are different or even unrelated, I propose the term re-use (of data). Re-use of data is different from the previously described re-analysis (of data) in the sense that, while the dataset remains identical in both cases, the research question and method of analysis are not similar anymore in this scenario, but different or even unrelated. Note, however, the potential overlap between the two terms.

The requirements for enabling re-use (of data) are comparatively simple: the dataset needs to be publicly available. However, as in all other cases where the dataset is required, this seemingly simple condition hides considerable complexity. Not only does the dataset need to be available, but it also needs to be understandable, interoperable, sufficiently well-documented and suitable for an inquiry into the new research question. As a consequence, in many cases where re-use of data is the goal, considerable effort is first invested into augmenting, cleaning, annotating or otherwise enhancing the dataset.

These modifications are legitimate, in this scenario, because we are here in the domain of related research, where the function of re-using the data is not to

---

7 The sustainability and interoperability of data in the Digital Humanities is a well-researched topic of its own that is beyond the scope of this paper; standardized data models such as XML-TEI (for text) or RDF (for Linked Open Data) have an important role to play here; see e.g. Rehm and Witt (2008); García et al (2016).
check its quality or the quality of the study it was used in, but simply to save
time and effort by re-using an existing dataset rather than creating a new one.

With the publication of datasets becoming increasingly common in CLS,
there are no doubt many cases of re-use of data. While project-specific,
research-driven corpora are sometimes not easily re-used, others, in particular
curation-driven ones that are large in size, contain reliable texts, use standard-
ized encoding, include rich metadata and provide detailed documentation, are
routinely being re-used. Examples for such corpora relevant to CLS include
the *Oxford Text Archive*, the *Deutsches Textarchiv*, *DraCor* or the *European
Literary Text Collection (ELTeC)*.\footnote{On these corpora, see: *Morrison (1999); Haaf et al (2022); Fischer et al (2019); Schöch et al (2021).*} Examples of re-uses of these corpora in
CLS include a study of the chorus in a Spanish-language drama corpus that is
included in *DraCor* (Dabrowsa and Fernández, 2020) or a study of the titles
of novels across multiple languages included in ELTeC (Patras et al, 2021).
Re-use of data is not limited to analysis, of course, but can also be performed
when one or several existing corpora are used to create a new corpus, as in the
case of the KOLIMO corpus which was created using texts from the *TextGrid
Digital Library*, the *Deutsches Textarchiv* and *Gutenberg-DE* (Herrmann and
Lauer, 2018).

6.7 Re-use (of method)

When the research question is similar, different or unrelated and the dataset
is different or unrelated with respect to earlier research, and only the method
or code is used in identical or very similar form, then we may speak of re-use
(of method) (or of code). Depending on how similar the dataset is and on how
flexibly the code can be used, the research question may of necessity be more
or less closely related to that of the original study.

Again, this scenario does not fall into the realm of exact or approximate
repetition of earlier research, but it is nevertheless a specific mode of repetitive
research. And it does have a specific function, which is to save implementation
time and increase the reliability and robustness of an implementation. In a
sense, any use of a tools or software packages developed by others is re-use
in this sense.

The main requirements for enabling efficient re-use (of method) are the
availability of the code, software package or tool with a minimum of (financial
or technical) hurdles as well as a detailed and understandable documentation.
Using existing tools for one’s own purposes is, of course, a normal practice in
Digital Humanities and CLS, so mentioning specific examples does not appear
particularly useful in this case.

6.8 Follow-up research

If the research question remains identical or very similar, but the dataset used
as well as the method of analysis are different or even unrelated, then I propose
to use the term *follow-up research*. This scenario is clearly in the domain of related research, rather than exact or approximate repetition, because there is a comparatively distant relationship between earlier studies and the later study on the same question.

In terms of functions of *follow-up research*, a later study on the same question that obtains similar results can certainly be understood as a corroboration of earlier studies on the topic, showing that these results are robust against variations in the dataset and the method used to elucidate the question. However, the relationship between the studies being much looser than in the RR scenarios strictly speaking, differing results based on a follow-up study do not necessarily indicate that anything is wrong in the earlier studies, given that the dataset used is different.

The requirements for enabling *follow-up research* are very low, as research question, data and method or code do not need to be and usually are not identical. Pushed too far, for example by focusing on very different or unrelated research questions, and this scenario breaks the chain of incremental increase in our knowledge about a domain.

My own study repeating a classic study by stylistician Leo Spitzer about French playwright Jean Racine is an example of such *follow-up research* (Spitzer, 1931, 1969; Schöch, to appear). The original intent was to perform a very close re-enactment of this non-computational and qualitative study using digital data and methods, but it became clear rather quickly that a strict replication was impossible and that even a re-implementation (of the method) was hardly feasible, given the analog-digital and the qualitative-quantitative divides as well as the lack of information about the exact editions used by Spitzer. As a consequence, this instance of repetitive research proceeded in multiple steps that became increasingly distant from the original study, ending up with a larger and more diverse corpus, a mixed-methods approach combining modeling of stylistic devices with statistical analysis that ended up shifting the research question from Racine himself to Racine’s position among the contemporary authors: clearly, *follow-up research* rather than any more specific type of repetitive research.

### 7 Conclusion

As a way to conclude, it appears useful to briefly reflect on the affordances and limitations of the proposed conceptual space and terminology of RR.

The primary affordance of the description of RR as structured into a five-dimensional conceptual space appears to be that this allows us describe, with considerable granularity, precision and transparency, and without necessarily using a given terminology, the relationship between an earlier study and a later study that is in some sense related to it. The main advantage of the set of terms proposed above is that they are clearly defined with respect to the conceptual space and that they provide us with convenient and distinct labels for a certain number of recurring scenarios within the conceptual space of RR. Several
studies that have been designated with the same term, within this terminology, can reasonably be expected to show a considerable amount of similarity. Taken together, especially when considering that the conceptual space and the set of terms also come with a description of functions and requirements, the hope is that this conceptual work can support the community of researchers in CLS and DH more broadly to more clearly understand, value more appropriately, and more frequently practice RR.

A possible limitation of the conceptual space is the choice of the five dimensions. One may argue that the concrete implementation of a method, for example as an executable algorithm implemented in a programming language, should have been separated out from the method understood broadly and turned into an additional dimension. A similar argument may be made for separating the claims or conclusions from the raw results, instead of treating them as one shared dimension of research. That is true but doing so would come at the cost of additional complexity not only of the conceptual space, but also of the terminology it supports. Another potential limitation of the terminology as it is proposed here is that the terms similar and dissimilar, while clear enough in everyday settings, are of course not categorically separate classes, but two rather fuzzy areas on a continuum. Specifications like functionally or strictly speaking similar or dissimilar may help clarify the usage in certain cases. Another possible limitation of the terminology is, of course, that despite efforts to minimize both gaps and overlap between the terms, some degree of both appears to be inevitable if the number of terms is to remain manageable and if terms are reserved for particularly frequent or important scenarios. Finally, people may disagree with the choice of terms themselves, in one or several cases. Fortunately, given that the terms and the conceptual space are defined quite explicitly, alternative ways of dividing up that space, alternative definitions of terms, or alternative terms for a given scenario can easily be proposed and exchanged.

With respect to the practice of RR itself, whether in CLS or in DH more broadly, there are some challenges, among them the considerable effort required to perform RR and the potential (or apparent) conflict with disciplinary values. Clearly, RR may be perceived to be at odds with the ways in which value is usually ascribed to research. Research usually needs to be original, innovative, ground-breaking, relevant, timely in order to be considered valuable. The kind of research I advocate for here, in contrast, is fundamentally concerned with repeating research that has been completed years or even decades ago. Can such research be said to be valuable in this sense, or to foster excellence in research? Of course it can, precisely because it serves such important and varied functions in the research process, whether it is quality assurance and building trust (as in strict replication), corroboration of results (as in several different scenarios), efficiency and sustainability (as in the re-use of data or code) or incrementally but methodically pushing the boundaries of knowledge.

\footnote{However, only once RR is practiced more frequently, will it become possible to observe which scenarios are indeed recurring, potentially leading to an adjustment of the terminology.}
Repetitive Research

(as in all follow-up research). In addition, practicing RR appears to be a learning opportunity to me, because one understands previous research much better when trying to re-implement or otherwise repeat it, including its strengths and limitations. More generally speaking, we also need RR as a way of guaranteeing the continuity, over time, of the disciplinary context of our work, especially in the Digital Humanities. Finally, and maybe most importantly, many of the functions of RR constitute or support best practices in the perspective of Open Science.

8 Acknowledgements

My thanks go to colleagues from the CLS and Distant Reading community who have provided invaluable feedback on some of the ideas proposed here on the occasion of several invited talks on the subject in Bern (2017), Montpellier (2019), Budapest (2019) and Munich (2020) that helped me sharpen my argument. All errors remain my own. – The epigraph comes from (Widdows, 2004, 2).

9 Declarations

• Funding: Part of this work was supported by the European Commission through the COST Action ‘Distant Reading for European Literary History’ (CA16204).
• Competing interests: None.
• Ethics approval: Not applicable.
• Consent to participate: Not applicable.
• Consent for publication: Not applicable.
• Availability of data and materials: Not applicable, but see data provided in Schöch (2016) and Schöch (to appear).
• Code availability: Not applicable, but see code provided in Schöch (2016) and Schöch (to appear).
• Authors’ contributions (following the CRediT taxonomy): CS is responsible for Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

References


Rockwell G (2015) Replication as a way of knowing in the Digital Humanities. In: Lectures in Digital Humanities, University of Würzburg


Spitzer L (1931) Die klassische Dämpfung bei Racine (1928). In: Romanische Stil- und Literaturstudien I. Elwert, Marburg, p 135–268

