

Reproducibility in Information Retrieval

Maria Maistro¹, Timo Breuer², and Nicola Ferro³

June 30, 2026

¹University of Copenhagen, mm@di.ku.dk

²TH Köln - University of Applied Sciences, timobreuer@acm.org

³University of Padua, nicola.ferro@unipd.it

Abstract

The reproducibility of earlier findings is fundamental to ensure the accountability and trustworthiness of the empirical sciences. Several reviews in different research fields showed that many earlier findings are not reproducible. To this end, the objective of this survey is to help setting a common ground and understanding of what reproducibility in Information Retrieval (IR) is and what can be done for reproducibility, also creating connections to similar issues in neighbouring fields, with the ultimate goal of fostering the adoption of good reproducibility practices and easing a more widespread and systematic production of reproducible papers and research in the field.

The survey focuses on reproducibility in IR, discusses the related terminology, challenges and barriers to reproducibility, ways of measuring and quantifying the achieved degree of reproducibility, and various reproducibility initiatives in the field. We also discuss and review best practices for achieving and improving reproducibility, concerning how to share data and code, how to document experiments and how to better adhere to open science principles; we also present the corresponding tools that can ease the day-by-day operation of such practices. We broaden the perspective on reproducibility to neighbouring fields, namely Recommender Systems (RS), Natural Language Processing (NLP), and others. Readers will be provided with a summary of current state of the art, the theoretical background, and actionable recommendations about how to validate the quality of their reimplementations, and how to prepare their experiments for reuse by others.

Chapter 1

Introduction

Whatever your stance about what science is [Kuhn, 1996, Popper, 2002, Quine, 1998], there is wide agreement that *experimentation* is at the core of the modern scientific method [Galilei, 1954] and that the possibility of carrying out the experiments more than once is central to science in terms of *validation* and *confirmation* of its findings. Broadly speaking, *reproducibility* (or *replicability* – a more precise discussion on terminology will come later on) is concerned with this need of running an experiment again and again to confirm, or even generalize, its results. Or, as better expressed by Popper [2002, p. 66]: “non-reproducible single occurrences are of no significance to science”.

Recently, the fourth paradigm [Hey et al., 2009] opened a completely new way of making scientific discoveries, based on computational and data-intensive approaches, where experimentation rooted in computation is getting even more prominent. This new paradigm, together with the possibility of collecting and processing bigger and bigger amounts of data, is giving rise to a whole new set of challenges for reproducibility.

Researchers in all areas of science are confronted with the so-called *reproducibility crisis* that has become apparent in more and more disciplines in recent years [Open Science Collaboration, 2015]. While the discussion on reproducibility started in Psychological Science [Pashler and Wagenmakers, 2012], it quickly became clear that researchers in almost every scientific field were struggling to reproduce previous experiments and findings, both their own and others’ work. As reported by Baker [2016], approximately 70% of researchers in physics and engineering are not able to reproduce someone else’s experiments, and roughly 50% fail to reproduce even their own experiments. Computational and data-intensive sciences are no exception [Freire et al., 2016, NASEM, 2019], as pointed out specifically for *Artificial Intelligence (AI)* and *Machine Learning (ML)* [Gibney, 2020, Pineau et al., 2021].

Information Retrieval (IR) is not exempt from its own reproducibility crisis, as well. In fact, reproducibility is especially relevant to IR, which is deeply rooted in experimentation and empirical evidence [Harman, 2011, Sanderson, 2010, Spärck Jones, 1981] and increasingly relies on machine learning approaches [Lu-

cic et al., 2022a]. In the last decades, the well-known Cranfield paradigm [Cleverdon, 1962, 1967] has helped IR researchers to live in a state of felt comfort and security because test collections have allowed a relatively high level of standardization and the adoption of best practices for evaluation. However, despite the Cranfield paradigm, IR is still far from achieving widespread reproducibility.

A survey by Ferro and Kelly [2018] among members of the SIGIR community¹ revealed that more than 75% of the respondents would have welcomed the introduction of policies for awarding badges for the reproducibility of papers, as a way to improve the scientific approach adopted by the field and the overall robustness of the reported findings. However, at the same time, about 40% of these respondents stated that reproducibility badges would probably or definitely not change their way of doing research, since they felt that they already conduct and report research in a way that makes it reproducible. This latter belief contradicts the everyday experience of the same researchers in the field, who often have severe issues in reproducing someone else experiments, with failure rates similar to those reported by Baker [2016].

Reproducibility is also a core part of the *open science* [Chubin, 1985] vision. Indeed, as pointed out by the *United Nations Educational, Scientific and Cultural Organization (UNESCO)* in its recommendation about open science [UNESCO, 2021, p. 7]: “open science sets a new paradigm that integrates into the scientific enterprise practices for reproducibility, transparency, sharing and collaboration resulting from the increased opening of scientific contents, tools and processes”. In this respect, reproducibility constitutes one of the *principles* of open science, where “increased openness leads to increased *transparency* and *trust* in scientific information and reinforces the fundamental feature of science as a distinct form of knowledge based on evidence and tested against reality, logic and the *scrutiny* of scientific peers” [UNESCO, 2021, p. 18]. Even if many practices in IR comply with and somewhat embody open science principles, the field still lacks a systematic discussion and a coordinated approach to open science and reproducibility is one cause of this still fragmented puzzle.

The objective of this survey is to help establish a common ground and understanding of what reproducibility in IR is and what the current best practices for reproducibility are. This survey also aims at creating connections to similar issues in neighbouring fields, such as *Recommender Systems (RS)*. Ultimately, the goal is to promote the adoption of good reproducibility practices and facilitate a more widespread and systematic production of reproducible papers and research.

To this end, the survey is organized as follows: Section 2 presents an introduction to the Cranfield paradigm and can be skipped by readers already familiar with that. Section 3 focuses on reproducibility in IR, discussing the related terminology, challenges and barriers to reproducibility, ways of measuring and quantifying the achieved degree of reproducibility, and various reproducibility initiatives in the field. Section 4 discusses best practices for achieving

¹SIGIR is the ACM Special Interest Group on Information Retrieval. More information at: <https://sigir.org/>.

and improving reproducibility, concerning how to share data and code, how to document experiments and how to better adhere to open science principles; it also presents some tools which can ease the day-by-day operation of such practices. Section 5 broadens the perspective on reproducibility to neighbouring fields, namely *Recommender Systems (RS)*, *Natural Language Processing (NLP)*, *Machine Learning (ML)*, and computer vision. Finally, Section 6 draws some conclusions and outlines future research challenges and perspectives for reproducibility.

Chapter 2

Brief Introduction to IR Offline Evaluation

This Section presents a brief overview of IR offline evaluation following the well-known Cranfield paradigm. This is needed because test collections, evaluation campaigns, and offline evaluation are strictly connected to reproducibility. For example, evaluation campaigns play an important role in making datasets and systems comparisons publicly available (see Section 3.2). The reader already familiar with evaluation can simply skip this section. Ferro and Maistro [2024], Sanderson [2010] offer more comprehensive surveys of IR evaluation.

There are two main paradigms for IR evaluation: offline, or system-oriented evaluation, and online, or user-oriented evaluation. As the names suggest, offline evaluation happens offline and uses annotated or logged datasets, while online evaluation exploits interactions from users online. The two paradigms are not independent or detached, Kelly [2009] presents the evaluation spectrum as a continuum from the system focus to the user focus, and de Rijke [2018] defines “Retrieval as Interaction”, a two stage process where systems are first evaluated offline and then continuously adapted through online evaluation with user interactions. Next, we describe offline evaluation, the reader interested in online evaluation might consider: Ingwersen and Järvelin [2005], Ruthven and Kelly [2011], White [2016].

2.1 Cranfield Paradigm

IR offline evaluation is based on the Cranfield paradigm [Cleverdon, 1967, 1962], which uses *experimental collections*. There are three components that constitute an experimental collection $C = (\mathcal{D}, \mathcal{T}, \mathcal{A})$: $\mathcal{D}$ is the set of documents or *corpus*; $\mathcal{T}$ is the set of *topics* or queries, i.e., a representation of the information need; and $\mathcal{A}$ is the set of *annotations*, or ground-truth, i.e., given a pair with a document and a topic (d, t) , it is the level or grade of relevance of the document d to the topic t . Documents can be annotated on a binary scale,

e.g., non-relevant and relevant, or on a multi-graded scale, e.g., non-relevant, partially relevant, fairly relevant, and highly relevant. An IR system will take as input the corpus $\mathcal{D}$ and the set of topics $\mathcal{T}$, then, for each topic $t \in \mathcal{T}$, it will return a *run* $r_t = (d_1, \dots, d_n)$, which is a ranked list of documents sorted with decreasing relevance score, as estimated by the system.

Figure 2.1 illustrates the workflow of offline evaluation. Given a topic t , the IR system produces the run r_t . Each document in the run is mapped to its relevance label using the annotations in $\mathcal{A}$. The result is a ranked list of relevance grades, called the assessed run. The assessed run is then converted to a weighted run by replacing each relevance grade with the corresponding weight, i.e., a natural number, with non-relevant documents usually associated to a zero weight. For binary relevance, it is common to use 0 for non-relevant and 1 for relevant documents. For multi-graded relevance, the same measure can result in different instantiations by changing the weights, e.g., *Normalized Discounted Cumulated Gain (nDCG)* in [Järvelin and Kekäläinen, 2002] is instantiated with weights 0, 1, 2, 3 and 0, 1, 10, 100 for the relevance grades: non-relevant, partially relevant, fairly relevant, and highly relevant. The weighted run is then processed by standard IR evaluation measures, which return a score, usually in $[0, 1]$, representing the effectiveness of the ranked list for the given topic. Finally, because experimental collections include multiple topics, the measure scores are aggregated across topics and the average is reported.

There is one caveat to the workflow described above. It is unfeasible to collect relevance annotations for all topic and document pairs. For example, assume to have a very small collection with 100 000 documents and 50 topics, it will require 5 million annotations to have a complete set, i.e., one annotation for each document-topic pair. Let us assume, optimistically, that each annotator needs 30 seconds on average to annotate one topic-document pair, then, we will need roughly 4.76 years, which corresponds to 416 667 USD, with a hourly salary of 10 USD. Considering that the modern search engines index hundreds of billions of documents and process billions of queries every day, it is not possible to have complete annotations even for a representative subset. Therefore, the set of annotations $\mathcal{A}$ not only is incomplete, i.e., there are missing annotations for some document-topic pairs, but it is also a small percentage of all possible pairs. Documents without annotations are treated differently, depending on the measure and the experimental set-up. The easiest solution is to treat unjudged documents as non-relevant documents, to ensure that measures scores are not overestimated. Indeed, if an unjudged document that was considered as non relevant, is relevant instead, most measures scores will increase or stay equal.

2.2 Evaluation Campaigns

Collecting relevance annotations, even for a small portion of an experimental collection, is a time-consuming and expensive task. Shared evaluation campaigns have emerged as efficient solutions to collect good quality annotations while sharing the effort to build experimental collections. They are public in-

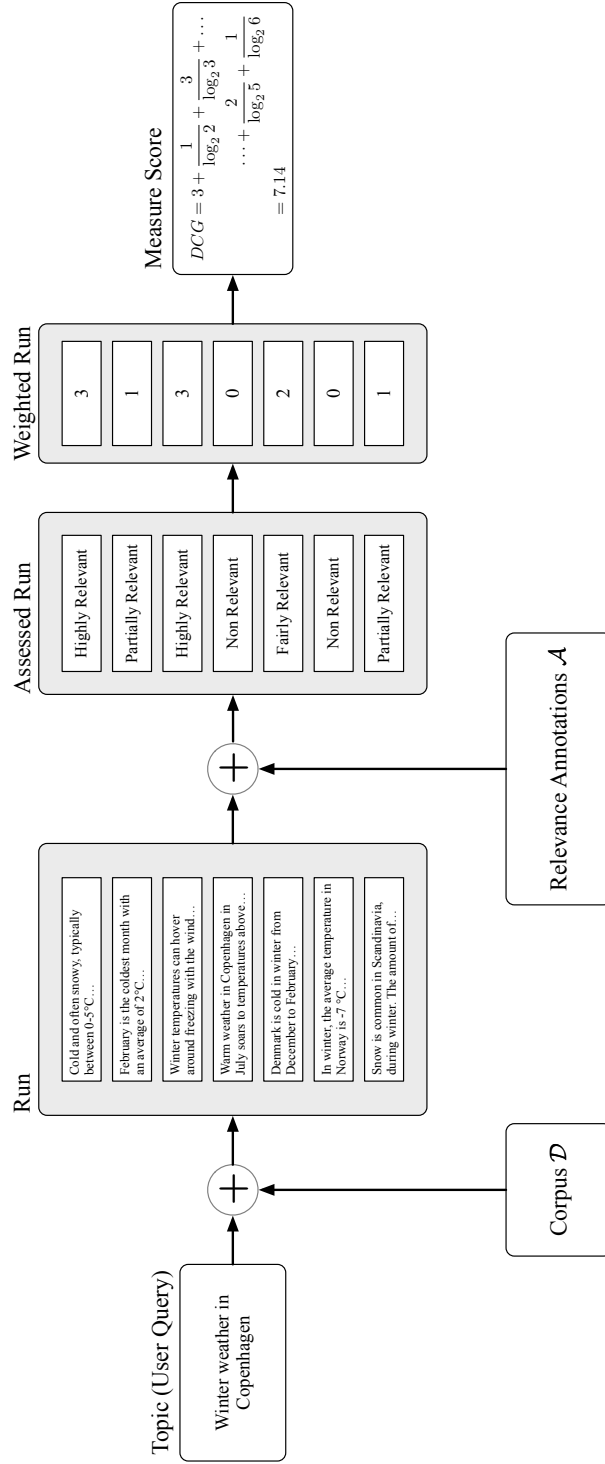


Figure 2.1: Workflow of offline evaluation. Give a topic and the set of documents (corpus), the IR system returns the run, i.e., the ranked list of documents sorted by estimated relevance. The documents in the run are mapped to the corresponding relevance annotations (assessed run) and relevance weights (weighted run). Finally, an IR effectiveness measure, e.g., DCG, takes as input the list of relevance weights and returns the effectiveness score.

ternational initiatives, or challenges, that involve different actors: organizers, assessors, and participants. Usually, the organizers choose and/or define the task, and prepare the corpus and the topics. The assessors are the annotators who will produce the relevance annotations, i.e., will read document-topic pairs and assign the relevance labels. Sometimes, assessors will work together with the organizers to prepare the topics. This is because relevance is highly subjective, therefore, it is important for an assessor to have a genuine interest in the information need and associated topic. Assessors can be common users, e.g., hired through crowdsourcing platforms, or domain experts, e.g., trained in medicine if the task requires annotating a corpus of medical documents. Finally, participants are groups or individual researchers and practitioners, who develop IR systems to retrieve and rank documents, given the corpus and topics provided by the organizers.

The typical cycle of an evaluation campaign works as follows. First, the organizers define the task and prepare an appropriate corpus. For example, if the task is Web search, the organizers can collect or reuse a crawl of Web pages. Then, the organizers build the topics, usually with the help of the annotators. In the example of Web search, topics can be sampled from real user queries submitted to a search engine. Once ready, the corpus and topics are released to the task participants. Participants will run their IR systems and submit runs to the organizers. At this point, the organizers need to prioritize the document-topic pairs to be annotated, as this can be done only for a sample, whose size depends on the available budget. This process is usually called pooling (see Section 2.3). Once annotators have judged the pooled documents, the relevance labels are returned to the organizers, who will proceed with the evaluation part. Usually, organizers will compute several evaluation scores for all the submitted runs and return all the resources, i.e., relevance annotations and measure scores, to the participants. To conclude the cycle, the organizers will perform further data analyses, e.g., statistical tests, to compare different systems and runs, and publish an overview paper, summarizing the main findings and lessons learned from the task.

Several evaluation campaigns have been running for more than 30 years. The *Text REtrieval Conference (TREC)*¹ [Harman and Voorhees, 2005] started in 1992 and has, since then, run every year in the USA. The *NII Testbeds and Community for Information access Research (NTCIR)*² [Sakai et al., 2021] is based in Japan and has been running every 18 months since 1999. In Europe, there is the *Conference and Labs of the Evaluation Forum (CLEF)*³ [Ferro and Peters, 2019a], with yearly editions since 2000. The most recent is the *Forum for Information Retrieval Evaluation (FIRE)*⁴, which started in 2008 and takes place every year in India.

¹<https://trec.nist.gov/>

²<https://research.nii.ac.jp/ntcir/index-en.html>

³<https://www.clef-initiative.eu/>

⁴<https://fire.irsi.org.in/>

2.3 Pool Creation for Document Annotation

As discussed in previous sections, collecting annotations for all document-topic pairs is unfeasible and organizers of evaluation campaigns need to select a representative subset of document-topic pairs to be annotated. Selecting a random sample of document-topic pairs would not work well because relevant documents are not uniformly distributed. It is indeed the opposite, relevant documents usually represents a very small percentage of the corpus, therefore, a random sample would result in most of the documents being labeled as non-relevant and consequently a waste of the annotation budget. *Pooling* represents a more efficient way to select such a sample considering the available budget.

Several pooling approaches have been introduced during evaluation campaigns, the easiest and most intuitive is top- k pooling. This pooling technique takes as input the corpus and (a subset of) all the runs for a given topic. Then, the first k documents of each run are kept and become part of the pool, while the remaining documents are discarded. The duplicated documents within the top k are removed. These are the documents that are retrieved by two or more runs, possibly at different rank positions. This process is repeated for each topic to create the pool, i.e., the set of document-topic pairs that will be annotated. The threshold k is decided based on the available budget, it is around 100 for large evaluation campaigns.

Top- k pooling is based on two intuitions. First, if the participants' IR systems are good enough, it will be much more likely to find relevant documents within the documents ranked in the top k positions, rather than uniformly sampling the corpus. Second, when the same document is retrieved within the top k by multiple and diverse systems, it might be because that document is relevant. Therefore, a good quality pool, i.e., a pool that contains many relevant documents, requires a good selection of runs, produced by the participants' IR systems. These systems needs to be diverse, i.e., if they have a small overlap, there are more chances that they retrieve diverse relevant documents and therefore it is more likely to retrieve almost all relevant documents for a given topic.

The quality of a pool is also defined in terms of its robustness, i.e., a pool that is not biased towards the systems that created that pool. This is a fundamental concept related to the reusability of experimental collections. For example, consider a researcher who develops a new IR system and decides to evaluate it on an experimental collection made available through an evaluation campaign. The new system did not contribute to the pool, therefore, it might retrieve relevant documents that are considered non-relevant because unjudged. The problem arises when this system is compared against other systems that contributed to the pool. This new system might underperform compared to those systems, simply because it retrieves relevant documents that were not part of the original pool and are not annotated. Zobel [1998] shows that top- k pooling is indeed a robust technique, i.e., it leads to fair comparison between systems that did and did not contribute to the pool, when the threshold k is sufficiently large. More recently, Voorhees et al. [2022a] show that top- k pooling is still robust

even when comparing against new IR models based on deep learning. Yilmaz et al. [2020a] show that this is not the case if the pool is not sufficiently deep, i.e., when the threshold k is set to 10.

2.4 Evaluation Measures

Next, we review some of the commonly used measures to evaluate the effectiveness of IR systems. All these measures take as input the weighted run, i.e., the ranked list where documents are replaced by their corresponding relevance weight (see Figure 2.1), and return as output a score, usually in $[0, 1]$. This score represents the effectiveness of the run, i.e., an estimation of how well the ranked list of documents satisfies the user’s information need.

Precision and Recall *Precision* (P) and *Recall* (R) are two fundamental measures in IR. Precision is the proportion of retrieved documents that are relevant, and Recall is the proportion of relevant documents that are retrieved, see Equations (2.1) and (2.2) respectively.

$$P = \frac{\# \text{ retrieved relevant documents}}{\# \text{ retrieved documents}} \quad (2.1)$$

$$R = \frac{\# \text{ retrieved relevant documents}}{\# \text{ relevant documents}} \quad (2.2)$$

Precision and Recall are often considered together because they measure the ability of an IR system to retrieve all relevant documents ($R = 1$), while at the same time non relevant documents are discarded ($P = 1$).

The intended use of Precision and Recall definitions was with sets of documents, not ranked lists. This is because early IR systems, e.g., systems based on Boolean operators, returned sets of documents instead of rankings. To compute Precision and Recall with rankings, we first define a cut-off threshold, then consider all the documents ranked above the threshold as a set. More precisely, given a cut-off k , Precision and Recall are computed as follows:

$$P@k(r) = \frac{1}{k} \sum_{i=1}^k \hat{r}[i] \quad \text{and} \quad R@k(r) = \frac{1}{RB} \sum_{i=1}^k \hat{r}[i] \quad (2.3)$$

where $\hat{r}[i]$ is the relevance weight for the document at rank position i and the *Recall Base* (RB) is the total number of relevant documents. Note that Precision and Recall are binary measures, i.e., the relevance weight $\hat{r}[\cdot]$ is either 0 or 1. In case of multi-graded relevance, all relevant documents are assigned a weight of 1 independently of their grade, e.g., a highly relevant and a partially relevant document will both have a weight equal to 1.

The choice of the cut-off value depends on the task and the domain. For example, for Web search it is very often set to 10 because the first page of a Web search engine usually displays ten results. For recommendations, it is common

to consider also larger values, e.g., $k = 20$, because users might receive longer recommendation lists. For specialized domains, such as legal search, where it is crucial not to miss any relevant documents, the cut-off can be even larger, e.g., $k = 100$.

Finally, note that Recall depends on RB, the total number of relevant documents. To compute RB exactly, we would need to annotate all document-topic pairs. However, as discussed in Section 2.1, this is not feasible. Therefore, Recall is based on an estimate of the RB, obtained through the pooling technique (see Section 2.3). This is a further argument in favour of having pools of documents that are sufficiently deep and robust because the deeper and the better the pool construction, the better the estimate of RB.

Average Precision *Average Precision (AP)* [Buckley and Voorhees, 2005] can be viewed as a way of combining Precision and Recall into a single score, while also accounting for the rank positions of relevant documents. Indeed, one of the main limitations of Precision and Recall is their set-based nature, i.e., they do not reward a retrieval model if relevant documents are close to the beginning of the ranking. Instead, AP is a top-heavy measure, meaning that the closer a relevant document is placed to the top of the ranking, the higher its contribution to the measure score.

Average precision is defined as the average of Precision values, computed at cut-offs corresponding to the retrieved relevant documents. Its formulation is defined below:

$$AP(r) = \frac{1}{RB} \sum_{i=1}^n \hat{r}[i] \cdot P@k(r) = \frac{1}{RB} \sum_{i \in \mathcal{R}} P@k(r) \quad (2.4)$$

where RB is the recall base, n is the length of the run, $\hat{r}[i]$ is the relevance weight of the document at position i , and $\mathcal{R}$ is the set of rank positions where a relevant document is retrieved. AP is a binary measure, therefore, as for Precision and Recall, the relevance weights are either 0 or 1. AP ranges in $[0, 1]$ and the higher the score the better. When $AP = 1$, it means that all relevant documents are retrieved and they are placed continuously from the top of the ranking, i.e., all the first RB rank positions, and only those, are filled with relevant documents. The average of AP scores across topics is called *Mean Average Precision (MAP)*.

Discounted Cumulated Gain DCG is a top-heavy measure that accommodates multi-graded relevance [Järvelin and Kekäläinen, 2002]. DCG assumes a user who scans documents in the ranked list sequentially, from the top to the bottom of the list. Relevance is discounted, depending on the rank position, and then “cumulated” while scanning the rank positions from the top towards the end of the run (a sum of discounted relevance weights). Up to rank position b , there is no discount, and the measure is essentially set-based. Therefore, the relevance weights of all documents ranked in the first b positions are simply summed. After rank position b , the discount function is applied to reduce

the contribution of relevant documents to the summation. The discount is computed as the inverse of the logarithm of the rank position of a relevant document. Therefore, the further a relevant document is from the top of the ranking, the higher the discount, and the lower its contribution to the cumulated relevance. The exact formulation of DCG is as follows:

$$\begin{aligned} DCG@k(r) &= \begin{cases} \sum_{i=1}^k \hat{r}[i] & \text{if } k < b \\ \sum_{i=1}^{b-1} \hat{r}[i] + \sum_{i=b}^k \frac{\hat{r}[i]}{\log_b(i)} & \text{if } k \geq b \end{cases} \\ &= \sum_{i=1}^k \frac{\hat{r}[i]}{\max(1, \log_b(i))} \end{aligned} \quad (2.5)$$

where $\hat{r}[i]$ is the relevance weight of the document at rank position i . The switching condition is based on the parameter b , which represents the patience of the user. When $b = 10$, DCG assumes a patient user who examines all documents up to position 10 (the first *Search Engine Result Page (SERP)*) independently of their rank positions. Conversely, when $b = 2$, DCG assumes a very impatient user who examines the first document with probability equal to 1, but, after that, the following documents are examined with decreasing probability, modeled by the discount function.

DCG has two limitations that make it hard to interpret its scores, especially across different topics. First, it is not bounded in $[0, 1]$. As shown in the example computation in Figure 2.1, the final score of DCG can be greater than 1, depending on the relevance weights, the number of relevant documents retrieved, and their rank positions. Second, DCG does not account for the recall base, i.e., DCG score does not provide information on how many relevant documents are missed by the IR system. *Normalized Discounted Cumulated Gain (nDCG)* addresses these limitations by normalizing DCG with the score of the ideal run. This is the run that retrieves all relevant documents and sorts them by decreasing relevance label, e.g., all highly relevant documents, followed by all fairly relevant documents, then all partially relevant documents, and finally non-relevant documents. nDCG is defined as follows:

$$nDCG@kr = \frac{DCG@k(r)}{DCG@k(i)} \quad (2.6)$$

where i is the ideal run.

Other Measures Precision, Recall, AP, and nDCG are some of the most popular measures in IR evaluation. However, there exist several other measures; we briefly mention some of them next. *Reciprocal Rank (RR)* [Singhal et al., 1997] is a binary measure computed as the reciprocal of the rank position of the first relevant document. *Binary Preference (bpref)* [Yilmaz and Aslam, 2006] is an extension of AP which is expected to be robust to missing relevance annotations. *Rank-Biased Precision (RBP)* [Moffat and Zobel, 2008] is a binary measure that assumes a user who scans the documents from the top to the bottom of the run and decides to examine the next document with probability p

or abandon the search with probability $1 - p$. *Markov Precision (MP)* [Ferrante et al., 2014] is a binary measure that instead assumes a user who can skip documents and re-examine documents multiple times. It uses Markov chains to model the user behavior. *Expected Reciprocal Rank (ERR)* [Chapelle et al., 2009] can be seen as a multi graded extension of RR. Again, it assumes a user who scans the run from the top to the bottom and at each rank position decides whether to examine the next document or abandon the search, with a probability that depends on the rank position and on the relevance of the previously examined documents.

Chapter 3

Reproducibility in Information Retrieval

This section offers a comprehensive examination of the terminology and definitions surrounding reproducibility, replicability, generalizability, and related concepts (cf. Section 3.1). Focusing specifically on *Information Retrieval (IR)*, we explore the unique challenges and barriers to reproducibility that are prevalent in this field, many of which are also relevant to the broader domains of computational and data-driven sciences (cf. Section 3.2). We then discuss evaluation measures that effectively quantify the levels of reproducibility and replicability in reimplemented IR experiments (cf. Section 3.3 and 3.4). The section concludes with an overview of current initiatives and efforts within the IR community aimed at enhancing and evaluating reproducibility practices (cf. Section 3.5).

3.1 Terminology and Definitions

The following addresses the ACM terminology adopted by SIGIR and reflects the debates regarding the correct usage of the terminology across different scientific fields. Eventually, these definitions are aligned with the PRIMAD model, which offers a distinct yet compatible perspective on reproduced IR experiments.

3.1.1 The SIGIR Community

The SIGIR community is aligned with the definitions introduced by the *Association for Computing Machinery (ACM)* for its policies on artifact review and badging,¹ version 1.1 as of 24 August 2020², and reported below.

¹<https://www.acm.org/publications/policies/artifact-review-and-badging-current>

²This is the most recent version at the time of writing, November 2025

Repeatability (Same team, same experimental setup) The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a team of researchers can reliably repeat their own computation.

Reproducibility (Different team, same experimental setup) The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author’s own artifacts.

Replicability (Different team, different experimental setup) The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

These definitions are consistent with the recommendations and definitions provided by the US National Information Standards Organization [NISO, 2021] on reproducibility and badging, definitions that, in turn, are based on the definitions of reproducibility and replicability given by the US National Academies of Sciences, Engineering, and Medicine [NASEM, 2019]. In particular, according to NASEM, “*reproducibility* is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis” while “*replicability* is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data”. Note that NASEM’s definition of reproducibility is synonymous with *computational reproducibility*, thus being particularly well-suited for the IR field, and closely matches the one adopted by ACM. On the other hand, the definition of replicability might be slightly broader than the ACM one, focusing more on the use of new data for confirmation and not just on a new artifact/system. This notion of replicability somehow overlap with another concept frequently used in science, i.e., **generalizability** that both Bollen et al. [2015] and NASEM [2019] define as “the extent that the results of a study apply in other contexts or populations that differ from the original one”.

Figure 3.1 shows what we could call the *terminology nautilus* and highlights in bold what changes and adds up from one definition to another and how these definitions entail increasingly complex and broader implications.

We begin with the original *experiment* and the first step is *repeating* it, so the same group is carrying it out again on the same data/setup, using the same software (they developed) and for the same purpose, i.e., same task/goal. As reported by Baker [2016], already this step is not trivial, since roughly 50%

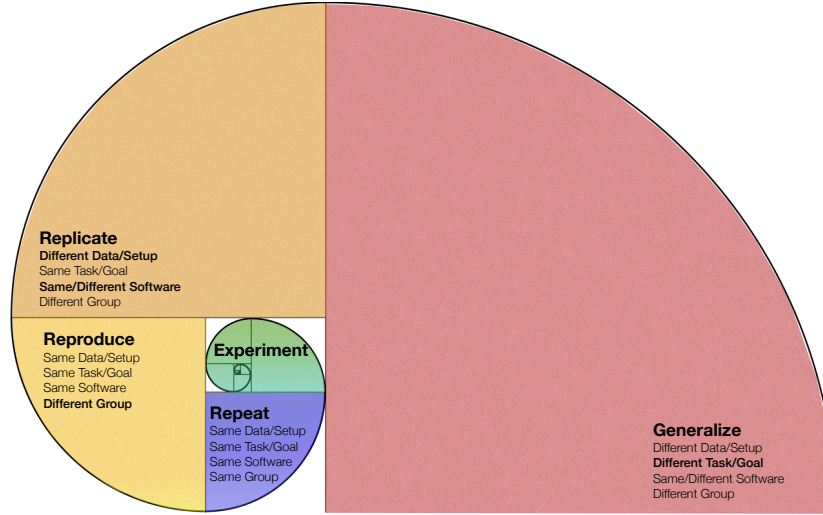


Figure 3.1: The terminology nautilus.

of researchers fail to repeat their own experiments. Even though the study by Baker [2016] does not include a specific category for Computer Science (the most related category is Physics and Engineering), the warning is still worth heeding.

The second step is to *reproduce* an experiment where, ideally, what changes with respect to the previous step is that now a different group is conducting the experiment; we say “ideally” because in practice both the software and the data may need some adjustments or modifications to run the experiment, even if the authors of the original experiment shared them. Also this step is far from trivial, with the failure rate being around 70% according to Baker [2016].

The third step is *replicating* an experiment which, in addition to the changes in the previous steps, requires changing the experimental setup, i.e., using different data and/or software. In this step, we find many of the issues which led to the *reproducibility crisis* in many areas of science [Baker, 2015, 2016, Open Science Collaboration, 2015] and led to lack of confirmation and validation of previously published findings.

Finally, the fourth and last step is *generalization*, where researchers explore if a method can be generalized and/or re-purposed for a different task/goal. In this last step, which is the most challenging and somewhat “uncertain”, we are departing more and more from reproducibility and replicability in a strict sense and we open up to wider scientific inquiries, seeking for other areas or fields of application of a methodology.

3.1.2 Terminological Debate

While the terms repeatability, replicability, and reproducibility are widely used in most areas of science, their definition and use often changes from area to area, leading to a quite fragmented picture [Plesser, 2018].

Even the definitions introduced above are the results of a quite recent “standardization” process in the area of data intensive sciences but they were understood quite differently until a few years ago. Indeed, the ACM policy for artifact review and badging was firstly approved on 8 June 2016, and then revised in April 2018 and in July 2020, and in this version³ – now labelled Version 1.0 (not current) – the definitions of replicability and reproducibility were essentially swapped.

In 2019, the report by the US the *National Academies of Sciences, Engineering, and Medicine (NASEM)* Committee on “Reproducibility and Replicability in Science” was released [NASEM, 2019] and this report informed the work of the US *National Information Standards Organization (NISO)* Standing Committee⁴ on “Taxonomy, Definitions, and Recognition Badging Scheme” which then adopted them in their recommendation [NISO, 2021]. The NASEM and NISO vision was motivated by a thorough survey by Barba [2018] who analysed the terminology adopted in the “reproducibility” domain across different areas of science and noted that the notion of replicability and reproducibility introduced in the previous section was the most commonly used and understood. A report by Heroux et al. [2018] further discussed the misalignment between the “reproducibility” definitions and pledged for a more unified view. As a consequence NISO sent a letter, delivered in July 2019 to the ACM Publications Board, asking ACM to reconsider their definitions [Barba, 2021]. Therefore, ACM decided to update its own definitions for artifact review and badging to be aligned with the (then forthcoming) NISO definitions;⁵ this, eventually, led in August 2020 to the swap in the definition of reproducibility and replicability between version 1.0 and version 1.1 of the artifact review and badging policy.

3.1.3 The PRIMAD Model

The *Platform, Research goal, Implementation, Method, Actor, and Data (PRIMAD)* (pronounced “primed”) model [Ferro et al., 2016a, Rauber et al., 2016a] takes a different stance on reproducibility and, instead of focusing on the definition of the different reproducibility-related terms, acts as a framework to distinguish the major components describing an experiment in computer science (and related fields). It was developed during the Dagstuhl Seminar on “Reproducibility of Data-Oriented Experiments in e-Science” [Freire et al., 2016] and covers approaches to reproducibility across different areas of Computer Science (e.g., IR, Databases, Computer Systems, Data Visualization, etc.) in order to come to a common ground across disciplines, leverage best-of-field approaches, and

³<https://www.acm.org/publications/policies/artifact-review-badging>

⁴<https://www.niso.org/standards-committees/reproducibility-badging>

⁵<https://www.acm.org/publications/badging-terms>

provide a unifying vision on reproducibility.

PRIMAD consists of the following components:

Research goal (R) characterizes the purpose of a study;

Method (M) is the specific approach proposed or considered by the researcher;

Implementation (I) refers to the actual implementation of the method, usually in some programming language;

Platform (P) describes the underlying hardware and software, like the operating system and the computer used;

Data (D) consists of two parts, namely the input data and the specific parameters chosen to carry out the method;

Actor (A) refers to the experimenter.

The term PRIMAD is derived from the first letters of the component names, but in a different order: **P**latform – **R**esearch goal – **I**mplementation – **M**ethod – **A**ctor – **D**ata.

In the following, we analyze possible changes for each of these components in order to characterize specific forms of reproducibility:

A → **A'** : When another researcher tries to reproduce an experiment, if she tries to rerun the experiment without changing anything else, then we have another actor, that is, A is changed to A', in other terms, the actor is “primed”. If successful, this experiment would demonstrate that the original researcher has supplied enough information to ensure its reproducibility. In case the results of the experiment are the same, then the original findings have been successfully reproduced and thus confirmed.

P → **P'** : In most cases, independent researchers do not have access to the platform used in the original experiment. Even different versions of system libraries, or external resources such as dictionaries, might have subtle effects on the outcome of experiments.

I → **I'** : Here a researcher uses a different implementation, say Terrier⁶ instead of Lucene⁷, or their own reimplementation.

D → **D'** : Rerunning an experiment with different parameters might be useful for testing the robustness of a method. Applying the implementation to different input data (for example, test collections) aims at investigating the generality of the method.

M → **M'** : Most of the research in the field of IR deals with the investigation of alternative methods, e.g., retrieval models. This implies also a new implementation I', which often runs on a different platform P'. However, for performing comparisons, the (input) data D should be the same.

⁶<http://terrier.org/>

⁷<https://lucene.apache.org/>

$\mathbf{R} \rightarrow \mathbf{R}'$: When the research goal is changed, then we *repurpose* some of the components of the experiment for another research question (for example, performing interactive retrieval experiments). So method $\mathbf{M}'$ and implementation $\mathbf{I}'$ usually are also changed, and other components as well.

We can observe how the above changes can be aligned with the terminology nautilus in Figure 3.1:

- **Repeat**: all the components are kept the same, so nothing is “primed” in this case;
- **Reproduce**: this happens when another researcher is involved, i.e., $\mathbf{A} \rightarrow \mathbf{A}'$;
- **Replicate**: in this case, in addition to a possible change in $\mathbf{A} \rightarrow \mathbf{A}'$, we could have changes in the platform, $\mathbf{P} \rightarrow \mathbf{P}'$, in the implementation, $\mathbf{I} \rightarrow \mathbf{I}'$, or in the data, $\mathbf{D} \rightarrow \mathbf{D}'$;
- **Generalize**: in this case, in addition to all the previous changes, we could also change the method, $\mathbf{M} \rightarrow \mathbf{M}'$, or the research goal, $\mathbf{R} \rightarrow \mathbf{R}'$.

Thus, the PRIMAD model allows us to provide a more fine-grained description of what has been changed and modified as we progress in the terminology nautilus, still giving us the possibility of benefiting from a standardized terminology.

Breuer et al. [2019] leveraged the PRIMAD model to inform the design and development of *InfraSTRUCTurEs for Living LABs (STELLA)*, an infrastructure for running reproducible experiment in living labs.

In the following, we provide two examples of application of PRIMAD to both system-oriented and user-oriented evaluation.

Application of PRIMAD to System-oriented Evaluation

In system-oriented evaluation for ad-hoc retrieval (Chapter 2 and [Sanderson, 2010, Ferro and Maistro, 2024]), the PRIMAD model could be instantiated as follows:

- **Research goal**, the question to be addressed, is in broad terms to produce a high-quality ranking of result documents, on average across a set of queries. There is a range of formulations of the hypothesis but the general principle is that the aim is to produce rankings where relevant documents are on average more highly ranked than in a benchmark system.
- **Method** is the mapping from the query to an ordering of the documents. In much IR research this is encapsulated as a scoring or similarity function for a document with respect to a given query, but this is not universal. For example, the behaviour of a function may change as documents are processed or may in some way depend on a document ordering (in the collection, or in a candidate list of answers).

- *Implementation* and *Platform* constitute the system being used for retrieval. A core element is (obviously) the mapping that constitutes the method, but there are also other significant elements that are influential on the outcome. These include the document parser, which might be as simple as a tokeniser but might be a full NLP tool; external resources such as dictionaries or reference collections; and the tools used post-retrieval for computing effectiveness from the document rankings.
- *Actor* is not an explicit element in system-oriented information retrieval. However, the actor is implicitly present as an agent who is undertaking the experiment and can influence the reported outcomes. For example, test collections can be used to both develop methods and to evaluate their effectiveness. The actor needs to ensure that the development process and the evaluation process are reasonably independent of each other, or the results will be contaminated by the possibility of having been tuned to the data.
- *Data* is the test collection being used in the experiments. This may be a shared collection, or private. It embodies numerous decisions: the domain of documents of interest; the kinds of queries that apply to the documents; and the basis on which a document (or a fragment of a document) can be judged relevant.

Application of PRIMAD to User-oriented Evaluation

In the case of user-oriented evaluation [Kelly, 2009], the PRIMAD model could be instantiated as follows:

- *Research goal* is the research question to be addressed. In most cases, this part should also include the hypotheses to be tested with the experiment described in the remainder of the research paper.
- *Method* relates here to the experimental settings, which are used for testing the hypotheses specified before. So, besides the type of study, the method includes also the relevant aspects of the settings related to the research objectives.
- *Implementation* and *Platform* correspond here to the environment in which the study was carried out. Besides the system used for the study, also the group of users participating in the study as well as the exact conditions under which they participated belong to this aspect.
- *Actor* is the experimenter. In cases where the experimenter has direct contact with the users, the actor might have influence on the results of the study. Thus, the actor should be kept constant throughout the study and carefully described.
- *Data* has a twofold meaning in user studies. First, there are the data that comprise the so-called testbed, like the document collection or the tasks

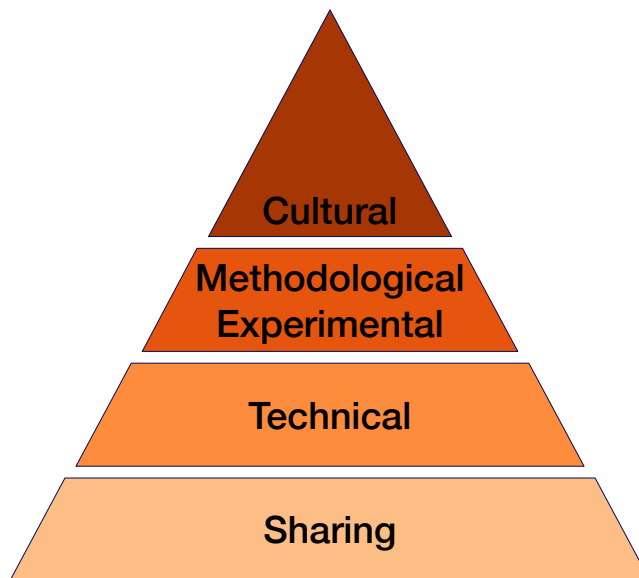


Figure 3.2: Challenges and barriers to reproducibility.

carried out by the users. Second, there are the observation data collected throughout the study (thus, the user is regarded here as a data generator).

3.2 Challenges and Barriers to Reproducibility

Reproducibility in IR shares challenges and barriers with many other disciplines, even if some of them have specificities due to the field.

Figure 3.2 summarizes the main categories of challenges and barriers to reproducibility that we discuss in the following. They are arranged in a pyramid-like fashion to highlight both their increasing complexity and their dependence on each other.

To give the reader a more concrete idea of the various forms these barriers can take, Table 3.1 at the end of this section offers an overview. Specifically, the table details the barriers identified in the proceedings of the ECIR Reproducibility Track from 2015 to 2024 and aligns them with the categories of Figure 3.2.

In the following, we outline the underlying issues associated with each category and provide additional examples from the literature to illustrate these points.

Sharing Challenges and Barriers

It is widely acknowledged that sharing *data* and *code* is vital to reproducibility [Freire et al., 2016, NASEM, 2019, Zobel et al., 2011], being at the core of the

definition of reproducibility itself (same experimental setup) and representing, if not a pre-requisite, at least an enabler and facilitator for replicability.

Geburu et al. [2021] highlight how, despite datasets playing a critical role in ML nowadays, there is currently no standardized process for documenting such datasets and this, in turn, hampers both reproducibility and also the possibility of documenting and tracking any biases in the datasets themselves. Given the importance of datasets and source code in IR, especially in the current era of neural IR methods [Tonellotto, 2024], all these issues should not be underestimated.

In the following, we describe the main challenges and barriers in sharing different types of data and resources which are at the core of IR experimentation.

Experimental Collections When it comes to sharing datasets IR has a long tradition, relying on the Cranfield paradigm [Cleverdon, 1962, 1967, Sanderson, 2010] which introduced the notion of an *experimental collection* made up of documents, topics and relevance annotations (more details can be found in Section 2.1). In this context, large-scale evaluation initiatives, namely TREC, NTCIR, CLEF, and FIRE (see Section 2.2), have fostered over the years the development of shared experimental collections, ensuring their quality, fairness, and robustness [Voorhees, 1998, 2018, Zobel, 1998], as well as access to them for both participants in their evaluation activities and other researchers.

Moreover, recent tools such as the `ir_datasets` package [MacAvaney et al., 2021], which provides a Python API to programmatically access many widely used experimental collections, or the *resource papers* tracks in top-conferences, such as SIGIR since 2021 and CIKM since 2020, where new datasets are described and made available, help in further improving the data sharing practices of our field.

However, does this mean that sharing data is a solved issue in IR? Actually no, since IR still has little to no consideration of proper methodologies for describing and making datasets accessible, such as adhering to and implementing the *Findable, Accessible, Interoperable, Reusable (FAIR)* principles [Wilkinson et al., 2016], adopted in many other areas of (open) science. This would require, among other things, to rely on a globally unique and persistent identification mechanism for the data items, such as the *Digital Object Identifier (DOI)* [ISO 26324, 2022, DOI Foundation, 2023], and to use metadata to describe the data items. Metadata need to be based on a formal, accessible, shared, and broadly applicable language for knowledge representation, i.e., relying on shared ontologies and Semantic Web technologies like *Resource Description Framework (RDF)* [W3C, 2024a,b,c]. Already in 2007, Agosti et al. [2007a] pointed out and advocated for the need of proper metadata and identification mechanisms for IR experimental collections but limited progress has been done since then, as noted by Ferro [2017] and reported in many other areas of computer science and engineering [Noy and Goble, 2023].

Runs If all the above is mainly focused on experimental collections, there is another valuable type of data in IR, namely *runs* produced by IR systems executed on an experimental collection (see Section 2.1). Runs are used for multiple purposes, all of them relevant to reproducibility and replicability.

The most immediate one is as a *baseline* for comparing new methods and algorithms. It is often not practical, or even possible, to re-execute an experiment due to several reasons, e.g., the source code is not available or difficult to re-implement, some online and ephemeral data are no more available, the execution of the system is too computationally expensive, etc. In such cases, having the run files available would allow for direct comparison to the baseline that it corresponds to.

Runs also play a fundamental role in developing and assessing our own evaluation methodologies, for example: evaluating robustness and quality of the experimental collections [Voorhees, 1998, Zobel, 1998], developing new pooling methodologies [Losada et al., 2017, Voorhees, 2018], advocating for the use of better statistical methods [Ferro and Sanderson, 2022, Hull, 1993, Savoy, 1997, Tague-Sutcliffe and Blustein, 1995], developing new evaluation measures [Järvelin and Kekäläinen, 2002, Moffat and Zobel, 2008], evaluating the quality of evaluation measures [Sakai, 2006, 2007], and many others.

When it comes to evaluation campaigns, as an example, TREC keeps an archive of the runs submitted to its different tracks and they can be downloaded; CLEF used the DIRECT system [Di Nunzio and Ferro, 2005, 2006, Agosti et al., 2007b, 2010, 2012] to archive and make available runs submitted to several of its tracks.

Other solutions have been developed outside large-scale evaluation campaigns over the years: EvaluatIR [Armstrong et al., 2009a] archived standard baselines on different experimental collections and allowed users to upload their own runs for comparison, computing a set of default evaluation measures; EvALL [Amigó et al.] allowed for uploading runs, computing performance measures, and generating reports in $\LaTeX$; recently, ranxhub [Bassani, 2023] provided similar functionalities allowing users to upload and compare runs but also defining some notion of descriptive metadata about a run and offering a Python API to programmatically access the archived runs.

Besides hosting run data archives, TREC established de facto standard data formats for run data, and related resources like relevance labels and topic files. Even though these data formats are not standardized by any international organization, they are well-adopted by the IR community. For instance, many evaluation toolkits adopt the TREC run file format or at least allow reading in this data type. Unfortunately, we observe a great deal of fragmentation when it comes to sharing runs and other resources, and none of the solutions above has become (yet) a de facto standard approach adopted by most people in the field. This represents a barrier to reproducibility in itself. Moreover, as discussed above, we still lack a principled way for sharing runs based on FAIR recommendations, a global identification mechanism, and well-defined data and metadata schemas. In the best case, a corresponding publication can be cited with a persistent identifier like a DOI but the data itself is currently not registered with

a persistent identifier.

So far, DIRECT has been the only tool adopting a global identification mechanisms for runs based on DOIs [Agosti et al., 2007a]. Silvello et al. [2017] developed an ontology, expressed in RDF, on top of the DIRECT system in order to formally and semantically describe IR data and runs and to share them as *Linked Open Data (LOD)* [Bizer et al., 2009]. Lipani et al. [2014] proposed an ontology for describing and sharing IR experiments and runs in the context of nanopublications.⁸ More recently, Breuer et al. [2022a] released `ir_metadata`, which is defined in a principled way, focused on reproducibility, and is based on the PRIMAD model [Ferro et al., 2016a, Rauber et al., 2016a]. Breuer et al. define a series of metadata fields for describing and archiving runs, provided a Python API to process such metadata and compare runs, and published an example dataset on Zenodo [Breuer et al., 2022b], in an Open Science fashion. Yet, Breuer et al. did not rely on formal languages for defining the metadata and describing runs. Later, Breuer et al. [2024b] built on this work and developed a web application to browse TREC metadata, accessible also via a RESTful interface. All these attempts represent important seeds towards sharing runs according to Open Science and FAIR principles; however, a coherent and widely adopted solution for the field is yet to emerge.

Measures and Statistical Analyses Performance scores, computed according to some measure like AP [Buckley and Voorhees, 2005] or DCG [Järvelin and Kekäläinen, 2002] (see Section 2.4), are usually not considered as data to be shared and/or archived for reproducibility purposes, even if they are the concrete output of the performance measurement process. Solutions like `EvaluatIR` or `EvALL` computed and made available the performance scores, or even prepared standard `LATEX` reports using them, but this was intended more as a convenience for the user rather than a need for reproducibility.

Even less, the statistical analyses and significance tests, such as the Student’s t-test [Student, 1908] or the *ANalysis Of VAriance (ANOVA)* [Fisher, 1925], are not shared as data, e.g., their settings, their detailed output (p-value, test statistics, degrees of freedom, etc.) – and, typically, just their major outcome (significant/not significant) is reported in the papers.

This behaviour is partially due to the availability of standard tools for computing performance scores, such as `trec_eval`,⁹ and the many statistical tools and packages (R,¹⁰ Matlab,¹¹ or Scypi¹²) and the fact that once relevance judg-

⁸A nanopublication is a small knowledge graph snippet with metadata that is treated as an independent (scientific) publication. In the IR context, nanopublications offer the possibility to publish statements about data and experiments, together with references that establish the authorship and provenance of the statements in a machine-readable format. Since nanopublications can be attributed and cited, they provide incentives for researchers to make their data available in standard formats that drive data accessibility and interoperability. For more information about nanopublication, please, visit <https://nanopub.net/>.

⁹https://github.com/usnistgov/trec_eval

¹⁰<https://www.r-project.org/>

¹¹<https://www.mathworks.com/>

¹²<https://scipy.org/>

ments and runs are available, one can easily recompute both performance scores and statistical analyses.

However, performance scores and statistical analyses are an integral part of IR experiments and they should be shared and archived as any other scientific data, as noted also by Dussin and Ferro [2009], who propose to frame the progress from runs to performance scores and statistical analyses in terms of the *Data, Information, Knowledge, Wisdom (DIKW)* hierarchy [Ackoff, 1989, Rowley, 2007].

Online, Private, and User-Data All the above considerations are mostly concerned with various kinds of data that are used in offline and system-oriented evaluation [Sanderson, 2010].

When it comes to user-oriented evaluation [Kelly, 2009] and online evaluation [Hofmann et al., 2016], the situation becomes more and more complicated because they often involve sensitive and personal user data, (often) impossible to share, and there is a lack of standardization and common formats in *Interactive Information Retrieval (IIR)*, as also noted by Petras et al. [2019a]. Examples of various forms of such private data are: personal data, e.g., gender or age, health-related data, financial data, e.g., salary, company and enterprise data, log data, or data created by online experiments, e.g., A/B testing, with real users and customers.

To help addressing the sharing and re-use issues in the IIR domain, Gäde et al. [2021] proposed a set of eight principles that the community is invited to adopt and embody including, among others, how to plan the lifecycle of data, how to document the research resources, how to archive and publish research resources, and more. However, in a recent study, Bogers et al. [2023] surveyed papers published at the CHIIR conferences¹³ from 2016 to 2022 and found that only 14.4% of the papers re-used existing data and just 4.2% shared (part of) their research data. This figures provide an idea of how difficult it still is to share data in the IIR domain.

When it comes to sharing sensitive and private data, the only option is to resort to some data anonymization techniques [Ciriani et al., 2007, Samarati, 2001] such as, for example, suppression or generalization of some data items, *k*-anonymity [Sweeney, 2002], or differential privacy [Dwork and Roth, 2014].

When *America Online (AOL)* released their query logs in 2006,¹⁴ even if users were not explicitly identifiable in the released logs, the New York Times was still able to locate individuals [Barbaro and Zeller, 2006], due to the fact that personally identifiable information was still present in the queries themselves. This highlights how hard it is to properly apply data anonymization techniques: it is not enough to remove some attributes from the data or to hash identifiers and person names or to inject some noise in the data, but one should take into

¹³CHIIR is the ACM SIGIR Conference on Human Information Interaction and Retrieval, provides a forum for the dissemination and discussion of research on the user-centered aspects of information interaction and information retrieval. More information at: <https://chiir.org/>.

¹⁴https://en.wikipedia.org/wiki/AOL_search_log_release

account all the possible means and external resources which could be exploited to identify a person, either directly or indirectly.

Moreover, it is important to understand that all the data anonymization approaches, while ensuring different degrees of privacy, may simultaneously reduce the *utility* of the data. Indeed the resulting anonymized data are “noisier” or less informative than the original data. As a consequence, the analyses carried out over them may lead to similar but less accurate results or, when used to train ML models, they may affect the effectiveness and deliver lower performance.

Finally, when sharing sensitive and personal data, there are also regulatory issues to be taken into consideration. For example, in Europe, the *General Data Protection Regulation (GDPR)* [European Commission, 2016, 2018] requires users to have provided and signed an informed consent for any follow-up use of their personal data.

Code While sharing source code is a seemingly simple step, which requires just to share a code repository, for example using GitHub,¹⁵ or, even better, linking a GitHub repository to an archival platform, such as Zenodo,¹⁶ this is still far from being a solved issue.

Collberg and Proebsting [2016] surveyed 600 papers published in 8 ACM conferences and 5 journals between 2011 and 2013, covering areas such as: Programming Languages, Operating Systems, Computer Systems, Architecture and Code Optimization, Privacy and Security, Data Management, and Databases. They report that the code was available with the papers themselves only in 14% of the cases; in 9% of the cases Collberg and Proebsting had to search (and found) the code on the Web; and, for another 13% of the papers, they obtained the code by mailing to the authors; overall, for only 38% of the papers, the code could be accessed in a way or another. Out of this 38% of the papers, Collberg and Proebsting succeeded in building and running the code in less than half an hour in 58% of the cases, in more than half an hour in 28% of the cases, and they were not able to build and run the code in the remaining 14% of the cases. One of the main barriers, concerning the code itself, pointed out by Collberg and Proebsting is that, even when the source code is available, it is not documented enough to rerun the experiments or it is not adequately prepared for rerunning the experiments.

More recently, in the NLP area, Arvan et al. [2022] found more encouraging results in terms of code sharing, since for roughly 50% of the papers the source code was available. However, when they attempted to reproduce a small sample of papers, they succeeded in just 20% of the cases for the same sort of issues pointed out by Collberg and Proebsting [2016].

Another source code audit in the RS domain found that approximately two thirds out of the 65 analyzed papers shared code artifacts [Shehzad et al., 2025]. However, a more in-depth analysis suggested that only a fraction of these papers provided the entire experimental setup, missing fundamental resources and

¹⁵<https://github.com/>

¹⁶<https://zenodo.org/>

information like baseline implementations, hyperparameters, and others.

Technical Challenges and Barriers

Even when datasets and code are shared and available, there are still lots of challenges for reproducing and replicating the experiments.

As we noted above, source code availability is not enough to guarantee reproducibility. Several studies [Arvan et al., 2022, Collberg and Proebsting, 2016, Ivie and Thain, 2018, Lin and Zhang, 2020, Kusa et al., 2022, Hasibi et al., 2016, Shehzad et al., 2025] have pointed out a set of common issues with the code: lack of sufficiently detailed documentation; missing details about hyperparameters; lack of proper management of or changes in the dependencies and libraries; changes and updates in the operating system, kernel, or system libraries; changes in different versions of the compiler, e.g., in the case of C or C++, or the Java virtual machine; unavailability of or changes in external resources or online *Application Program Interfaces (APIs)*, etc.

Apart from changes in external resources and APIs, all of these issues are typically addressed using some *virtualization*¹⁷ or *containerization*¹⁸ technology to package all the software needed for an experiment together with all its dependencies in a re-executable way (see Section 4.1). TIRA [Potthast et al., 2019] and its extension TIREx [Fröbe et al., 2023] are notable examples of a containerization approach used in the IR field and based on Docker¹⁹ containers, also adopted for running the PAN [Bevendorff et al., 2021, 2022, 2023, Ayele et al., 2024] and Touché [Bondarenko et al., 2022a, 2023, Kiesel et al., 2024] labs within CLEF.

However, it should be noted that even virtualization and containerization might not be the ultimate solution, since even hypervisors or containers such as Docker might evolve causing some changes or they may get superseded by other approaches, as also noted by Lin and Zhang [2020].

The hardware itself may become a barrier to reproducibility. For example, Ferro et al. [2019d] noted different results when reproducing the same neural IR experiments using either *Central Processing Units (CPUs)* or *Graphics Processing Units (GPUs)* on the same machine. While the differences are minor when computed with averaged effectiveness scores (i.e., ± 0.01 in MAP), there are noticeable differences in the ranked lists of documents. The distribution

¹⁷A software called *hypervisor* separates the physical hardware and the host operating system from the *Virtual Machines (VMs)*, allocates hardware resources to and executes one or more VMs, each one running its own guest operating system, possibly different from the host one, and applications within it. VMs are more demanding than containers in terms of computational resources needed but they offer better isolation from each other and from the underlying hardware and host operating system.

¹⁸Containers are lightweight, portable, and self-contained executable images that contain software applications and all their dependencies. Containers share the physical machine’s hardware and the host operating system kernel with other containers; as a consequence, the guest operating system, used by the container, and the host operating system, where the container runs, must be the same. Containers are less demanding than VMs in terms of computational resources and they are more portable.

¹⁹<https://www.docker.com/>

of relevant documents is similar but the actual documents are different, which can possibly lead to different user experiences. Such differences are caused by the non-deterministic behaviour of some basic tensor operation of the *Compute Unified Device Architecture (CUDA)* library,²⁰ when run on a GPU instead of a CPU. As another example, Arvan et al. [2022] encountered problems in reproducing a paper because the Pytorch²¹ implementation was attempting to perform a tensor operation mixing two different devices, a GPU and a CPU.

Sometimes the high computational requirements and the associated costs may represent a barrier to reproducibility [Ivie and Thain, 2018, Raff, 2019]. This also includes the need for specific hardware configurations, e.g., multi-GPU instead of single-GPU [Arvan et al., 2022]. Moreover, computational requirements are becoming more and more demanding due to *Large Language Models (LLMs)*, which are now integrated in IR systems, e.g., in the *Retrieval Augmented Generation (RAG)* paradigm [Fan et al., 2024, Gao et al., 2024, ?], for query expansion and rewriting, etc. [Staudinger et al., 2024, White and Shah, 2025].

Less common but even more severe, the access to extremely specific hardware may be an obstacle to reproducibility in itself. For example, a recent line of research is exploring the use of *Quantum Computing (QC)* for developing and implementing IR and RS algorithms, such as feature selection [Nembrini et al., 2021, Ferrari Dacrema et al., 2022, Niu et al., 2024, Fröbe et al., 2024], clustering [Alvarez-Giron et al., 2024], or instance selection [Pasin et al., 2024a]. All these papers require access to a *Quantum Annealer (QA)*, a specific type of quantum computer, to run the experiments. Access to a QA is not yet common and, typically, requires to resort to a single vendor, D-Wave,²² which is the leading company for this technology. There exist simulators of QA which could be used to run the experiments anyway. However, it goes without saying, that it does not lead to reproduction of results. To mitigate this issue, the QuantumCLEF lab has developed an infrastructure [Pasin et al., 2024b, 2023], based on containerization technologies, which allows for running, measuring, and reproducing QA experiments and which mediates the access to actual QA computers; however, access to the QuantumCLEF infrastructure is still needed to run the experiments.

A last, but not least, item of care concerns differences in the actual implementation of the same algorithm, which may or may not lead to reproducibility issues. For example Kriegel et al. [2017] conducted a study on different implementations of the same hierarchical clustering, k-means, and DBSCAN algorithms and they report substantial efficiency (runtime) differences – unfortunately, Kriegel et al. do not report results concerning the effectiveness. On the other hand, Kamphuis et al. [2020] conducted a study comparing several different implementations of BM25 [Robertson et al., 1995] and, despite minor variations in the effectiveness and efficiency scores, they did not find any

²⁰<https://developer.nvidia.com/cuda-zone> and <https://developer.nvidia.com/cuda-toolkit>

²¹<https://pytorch.org/>

²²<https://www.dwavesys.com/>

statistically significant difference among the various implementations.

Methodological and Experimental Challenges and Barriers

One of the most common experimental barriers is due to lack of enough details in the description of the experimental setup itself: how hyperparameters have been tuned, how the training/validation/test dataset have been split, which basic components (tokenizers, stop lists, stemmers, and so on) have been used in a pipeline and how they have been setup, and more [Ferro and Silvello, 2016a, Silvello et al., 2018, Roy et al., 2018, Berrendorf et al., 2021, Rao et al., 2015, Shehzad et al., 2025].

Lack of enough details about the experimental setup represents an even more severe barrier to reproducibility in the case of IIR experiments. In their manifesto, Gäde et al. [2021] point out the need for: more detailed description, via notes but also screenshots and recordings, of all the (often implicit) aspects concerning study goals, research questions, premises, hypotheses, and decisions that are not part of standardized research designs; definition and discussion of the terminology adopted in the study, ideally referring to existing, published definitions; and, preparation and publication of a data management plan, clarifying the data collection practices and the compliance with data protection rules, e.g., GDPR.

For several years evaluation campaigns have attempted to find a way to abstract IIR experiments and frame them in some sort of Cranfield-like fashion. For example, CLEF ran an Interactive CLEF (iCLEF) lab [Gonzalo and Oard, 2005, Karlgren et al., 2007, Gonzalo et al., 2009, 2010] for several years – 2004, 2006, 2008, 2009 – and then again Interactive Cultural Heritage in CLEF (iCHIC) [Petrás et al., 2013] – 2013 – and Interactive Social Book Search (iSBS) [Hall et al., 2014, Koolen et al., 2015, 2016] – 2014, 2015, 2016. The discontinuity alone in the editions is a clear indicator of how difficult it was to instrument such activities because of, e.g., problems in sharing data, in recruiting user cohorts comparable across participating groups, in synchronizing online experiments across different teams in different time zones, in adopting/sharing/reusing a common user interface to conduct the experiments, and much more. Petrás et al. [2019b] report these elements and others in their analysis of the lessons learned from organizing iCHIC 2013 and iSBS 2014–2016. If these were some of the obstacles encountered by the organizers of these labs, it is not hard to imagine how difficult it would be for a third party to try to reproduce some of these experiments.

Moreover, IIR and user studies face other experimental challenges due to the inherent subjectivity of users’ needs and variability in how they behave and interact with systems. In offline reproducibility studies, researchers try to reproduce the experimental set-up as closely as possible to the original one. However, it is not possible to replicate users, therefore, achieving identical conditions becomes impractical.

Friese and Fuhr [2025] propose a framework for reproducibility of IIR. Similarly to offline evaluation (see Section 3.3), reproducibility is assessed on differ-

ent levels: (1) Key findings, (2) Measurements, and (3) User behavior. In the first level, *Key findings*, a reproducibility experiment is assessed as successful if the overall findings of the reproducibility study are consistent with those of the original study, without directly comparing the user behaviour between studies. The second level, *Measurements*, is stricter. It compares the measured values between the original and reproduced studies pairwise and requires no or minimal difference between the original and reproduced values. The last level, *User Behavior*, is the strictest and requires a comparison of the overall search process. A pre-defined Markov chain is used to model the users’ search behaviour in the original and reproduced studies. Then, a successful reproducibility experiment is required to produce a Markov chain that is equivalent to the one obtained in the original study, i.e., there is no or minimal difference between the transition probabilities and time spent on each state of the Markov chain between the original and reproduced study.

One severe barrier to reproducibility, halfway between being experimental and methodological, is when there is some form of data leakage between training and test datasets or when hyperparameter tuning is performed directly on the test dataset [Berrendorf et al., 2021, Kapoor and Narayanan, 2023, Ferrari Dacrema et al., 2021] - both of which, typically, cause an inflation of the performance and prevent reproducing the results when experiments are re-run without such a flaw.

The problem of data leakage becomes even more severe with LLMs. The datasets used to pre-train LLMs usually include large crawls of the Web, books, scientific papers, code repositories, etc. Unfortunately, most of the private providers of LLMs do not disclose any information related to the exact composition of their pre-training datasets. This means that publicly available datasets and benchmarks, both training and test sets, can be part of such pre-training datasets. This “data contamination” issue, as it is sometimes called, can result in inflated measure scores and possibly unfair comparisons among systems based on LLMs [Zhou et al., 2023, Deng et al., 2024, Kaneko and Baldwin, 2025].

When it comes to experimental collections themselves, a problem that has been studied for a long time concerns the robustness and re-usability of such collections and, in particular, how fair these collections are when it comes to assessing systems which have not been used for creating them, e.g., via pooling [Zobel, 1998], or when it is known that assessors often disagree with each other [Voorhees, 2000]. While past studies have shown that experimental collections are reliable enough to fairly rank systems by their performance, more recently, the emergence of new pooling approaches, e.g., multi-armed bandits, and more sophisticated setups, e.g., living labs, have raised concerns. Specifically, some collections might be biased towards systems which have been used in their construction [Tan et al., 2017, Yilmaz et al., 2020b, Voorhees et al., 2022b, Voorhees, 2018]. Such kind of bias in collections may affect the replicability or generalizability of the experiments, especially for systems which have not taken part in the construction of the collection itself. For example, when trying to replicate results from one collection, where a system participated in its construction, to another one, where the system did not participate.

Another issue is related to comparing new methods against weak baselines [Armstrong et al., 2009b, Yang et al., 2019, Ferrari Dacrema et al., 2019, 2021, Kharazmi et al., 2016]. The consequence is that new methods appear to be more performant than what they actually are. Moreover, the supposed gains over the weak baselines do not hold and cannot be replicated or generalized when comparing to more appropriate state-of-the-art baselines and/or on other datasets.

Sometimes, there is a misalignment between the description of a methodology/algorithm reported in a paper and how it is actually implemented in the corresponding code [Berrendorf et al., 2020]. It also happens that the methodology is explained in an unnecessarily complex way or using an overwhelming mathematical notation, which hampers its actual understanding and reproduction [Lipton and Steinhardt, 2019].

From the methodological point of view, conducting appropriate statistical analysis on the experimental results relates to replicability or generalizability. For example, if the original experiment did not include any statistical analyses or did not use an appropriate statistical test, the results might be misleading and the findings might be hard to replicate or generalize. However, the adoption of proper statistical analysis in IR is still an open issue. Sakai [2016] conducted a survey on papers published in the SIGIR conferences²³ and the TOIS journal²⁴ over ten years (2006–2015). Sakai found that roughly 30% of the papers do not use any statistical significance test to determine which improvements are actually likely to hold, despite the use of significance testing to analyse the experimental results having been advocated for more than 30 years [Hull, 1993].

Moreover, when you compare different methods or algorithms using the same dataset, you run into a *multiple comparison problem*, a well-known issue in statistics [Tukey, 1949, Hochberg and Tamhane, 1987, Hsu, 1996], which causes a substantial inflation in the *Type I error* probability, i.e., the likelihood of detecting a significant difference – a method better than another – when it is not; in other terms, a false positive. The usual Type I error probability adopted in experiments is $\alpha = 0.05$, but, when performing multiple comparisons, the *Family-wise Error Rate (FWER)*, i.e., the probability of committing a Type I error across all the comparisons, grows up to $\text{FWER} = 1 - (1 - \alpha)^c$, where c is the number of comparisons performed [Ferro and Maistro, 2024]. Despite being a known problem, correction for multiple comparisons is not a widely adopted approach yet [Fuhr, 2017, Sakai, 2020, Fuhr, 2020, Ferro and Sanderson, 2024].

Another side of statistical significance testing concerns the *power* of a statistical test [Murphy et al., 2014, Cohen, 1988], i.e., its ability to detect when an effect is truly present. In other terms, it is the probability of avoiding *Type*

²³SIGIR is the Annual International ACM SIGIR Conference on Research & Development in Information Retrieval, the premier international forum for the presentation of new research results and for the demonstration of new systems and techniques in IR. More information at <https://sigir.org/>.

²⁴TOIS is the ACM Transactions on Information Systems, a scholarly journal that publishes previously unpublished high-quality scholarly articles in all areas of IR. More information at <https://dl.acm.org/journal/tois>.

II errors, i.e., failing to detect a significant difference, when there is indeed a significant difference. In a practical example, failing to conclude that a method is better than another when this is indeed the case. The typical power required for an experiment is 0.80. An *overpowered* experiment will detect significant differences, even when the effect size is tiny and, thus, practically meaningless. On the contrary, *underpowered* experiments will fail to detect significant differences, thus missing effects of practical importance.

Among other things, the power of a statistical test is affected by the variance of the sample. In underpowered experiments, variance is usually large, and this can have a subtle side effect. High variance means that estimating the parameters of the statistical test becomes harder, e.g., the sample mean is more widespread. Therefore, when a significant difference is detected, there is a high chance of having a large effect size. For example, there is a higher probability of having a large sample mean in studies with low power (large variance) than high power (small variance) [Gelman and Weakliem, 2009]. This, combined with Type I errors, which still happen with the same probability also in underpowered experiments, makes it possible to have a false positive, i.e., a non-existent significant difference, which looks even more “convincing” than in properly powered or even overpowered experiments, since the effect size is likely to be large. As a consequence, when experiments are underpowered, there is a greater risk of inconsistent and often misleading conclusions or, in other terms, experiments which are not replicable or generalizable. Sakai [2016] found that, among those papers that conducted proper statistical analyses, 64% of the papers have about the right power ($0.50 < \text{power} < 0.99$); 29% are overpowered ($\text{power} \geq 0.99$), and 7% are underpowered ($\text{power} \leq 0.50$). Equations and implementations to compute power for IR experiments are provided in [Sakai, 2016, 2018].

A subtle statistical problem that could arise when re-using the same experimental collection over and over is that of *sequential testing*. Carterette [2015] found that “the intrinsic effectiveness of the best known system could be as much as 20% lower than reported, and systems as much as 20% less effective than the best could have intrinsic effectiveness much higher”. Again, this randomness in the measurement of retrieval performance and its spread could affect replicability or generalizability.

Finally, when we move from reproducibility and replicability to generalizability, the change of context or of some research goals may prevent us from observing similar trends, e.g., when moving to more real-world environments which were simulated or due to temporal changes of the context [Shrestha and Spezzano, 2021, Berrendorf et al., 2021, Fröbe et al., 2022, Breuer et al., 2024a]. For example, Shrestha and Spezzano [2021] study feature importance in the task of fake news classification. They could replicate and generalize the findings of [Horne and Adali, 2017] only partially. For example, they observed substantially different results when changing the classifier (linear SVM vs. Random Forest) or when considering different domains (politics and gossip). Berrendorf et al. [2021] compare two state-of-the-art approaches [Wu et al., 2019, Fey et al., 2020] for Entity Alignment on Knowledge Graphs. Their experiment confirmed the superiority of the state-of-the-art approaches over the baseline

models, but also show that the state-of-the-art approaches struggle when using noisy datasets, which are typical from real world. Finally, Fröbe et al. [2022] investigates the reproducibility of findings on the effect of anchor text for Web search [Craswell et al., 2001, Eiron and McCurley] considering different time windows. Results are mixed; while Fröbe et al. [2022] confirmed that anchor text is still helpful in navigational queries [Craswell et al., 2001], they could not confirm that anchor text has similar characteristics to queries and that retrieval with anchor texts instead of documents yields more homogeneous results yields [Eiron and McCurley].

Cultural Challenges and Barriers

The “Publish or Perish” culture [Moskovkin, 2024], so widespread and overwhelming in academic life, is one of the major cultural barriers to reproducibility. Researchers, especially younger ones, live under the constant pressure of producing more and more publications and of obtaining more and more citations which, in turn, are used to quantify their impact and condition their career progress; the phenomenon is so well-known and pervasive that it has even turned recently into a card game [Kozlov, 2024]. This trend is even more exacerbated in the current era of Generative AI and LLMs, where the competition and the race to be first is pushed to the extreme. The consequence is a large number of not peer-reviewed papers, announcements of the next breaking result, blog posts, and leaderboard chasing [Sculley et al., 2018], which are published almost every day. All of this severely undermines the possibility of conducting reproducibility studies and systematic experiments since they require a lot of time and effort and this, in turn, would lead to fewer publications, a risk of losing one’s place in the competition, as well as the possibility of reproducing even recent research that suddenly becomes outdated.

Lin [2022] discusses how to counter-balance this constant pressure and how to create a culture of reproducibility, which incentivizes students into putting the effort required to release properly replicable and reproducible experiments. One key point raised by Lin is that it is critical to use self interest rather than altruism to motivate a student towards reproducibility. For example, a reproducible experiment used for a current study can become the stepping stone for an experiment in the next paper for the same student. At the same time, it could also become the starting point for the paper of another student and this is where the self interest becomes a benefit for the research group or the whole community where those students work.

The pressure to publish also leads to methodological and experimental flaws, which constitute unethical behaviors when intentional [Crocker and Cooper, 2011]. For example, to start experimenting without clearly formulated research questions or any hypothesis about the outcomes, leading to post-hoc formulated “assumptions” about favorable or significant results [Fuhr, 2017, Lipton and Steinhardt, 2019]; to resort to data dredging or p-hacking²⁵ for finding patterns in the data which lead to statistically significant results [Wasserstein and

²⁵Cherry-picking results such that the observed findings are statistically significant, for

Lazar, 2016]; to selectively report only significant results, as opposed to all the analyses conducted [Carbine et al., 2019, Kimmel et al., 2023]. Obviously, all such “inflated” results are hardly reproducible.

Overall, all these practices lead to various forms of *publication bias* [Sterling, 1959, Sterling et al., 1995], also known as the *file drawer problem* [Rosenthal, 1979], where results which are not statistically significant are not published. Recall that the Type I error probability, e.g., $\alpha = 0.05$, means that, if an experiment is repeated 100 times, a significant difference can be (erroneously) obtained 5 times, even if the null hypothesis is true, i.e., there is no statistically significant differences. This is due to the randomness intrinsic in the experiment.

Therefore, if we just publish the significant results, we would end-up publish only those 5 false positive, completely misrepresenting the true situation. Thus, publication bias can have very serious consequences: (i) effects that are not real may appear to be supported by evidence; (ii) meta-analysis studies [Kulinskaya et al., 2008, Borenstein et al., 2009], aggregating results from previous research, may get biased by failing to meet the underlying distributional assumptions; and, (iii) researchers may spend time and effort in investigating topics which have been already investigated but not published, because of the lack of significant results. For all these reasons, there is an increase in the number of venues that encourage the publication of negative results such as: *The Missing Pieces: A Collection of Negative, Null and Inconclusive Results*²⁶ by PLOS One, the *The All Results Journals*,²⁷ or the *Journal of Articles in Support of the Null Hypothesis*.²⁸

When it comes to IR, a methodology based on topic splitting has been proposed to estimate the amount of publication bias in system-oriented experiments [Ferro and Sanderson, 2022, 2024]. Topics are split into pairs of topic sets by sampling without replacement. The topic sets in each pair have the same size and multiple pairs are created with varying sizes. Then, given a pair of topic sets, publication bias is observed when there is statistical significance on a topic set but the same experiment shows no statistical significance on the other set. The probability of observing publication bias is estimated as 1 minus the ratio between the number of agreements, i.e., statistical significance is observed on both topic sets in the pair, divided by the total number of pairs of topic sets where statistical analyses are conducted.

Ferro and Sanderson [2024] have studied two separate conditions: the track condition and the participant condition. The track condition is the typical setup used in the IR literature to validate an evaluation methodology or a statistical significance test. In this condition, all the runs submitted to a track of an evaluation campaign are used. Instead, the participant condition uses only the runs of a specific participant. The former condition is more varied since runs from different participants typically adopt very diverse approaches; the

example, by conducting different statistical tests but reporting only those that are significant [Wasserstein and Lazar, 2016].

²⁶<https://collections.plos.org/collection/missing-pieces/>

²⁷<http://www.arjournals.com/>

²⁸<https://www.jasnh.com/>

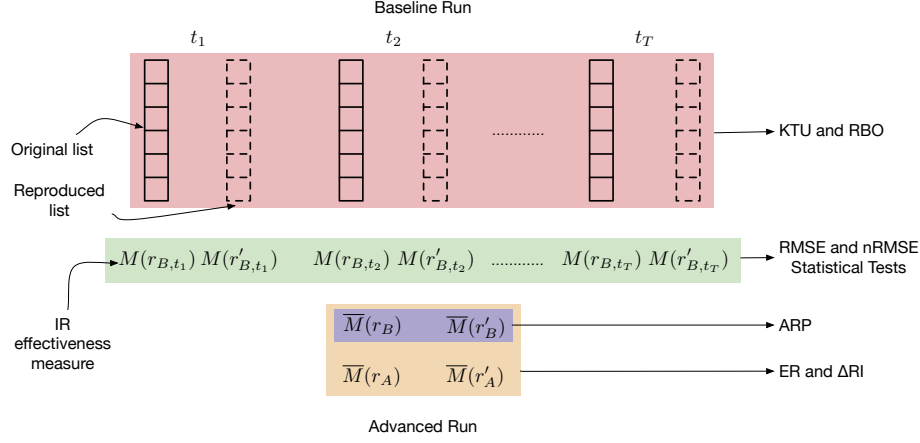


Figure 3.3: Reproducibility measures organized by their inputs: ranked lists of documents for KTU and RBO; per topic scores for RMSE and nRMSE; average scores for ARP; and average scores for a baseline and advanced run for ER and DeltaRI.

latter condition is less varied since runs from a single participant are typically variations of the same system. For this reason, the track condition is usually considered more complex than the participant one and, thus, behaviors and trends found in the track condition are implicitly assumed to hold also in the participant condition.

However, Ferro and Sanderson [2024] found that, when using 50 topics, the amount of publication bias in the track condition can be 21% when using *Average Precision* (AP) and 14% when using nDCG while in the participant condition it can go up to 58% when using AP and to 46% when using nDCG.²⁹ This is a worrisome finding since the participant condition is what researchers experience when they write and publish their own papers, i.e., they compare variations of their own systems, and the amount of publication bias in this condition is much higher. To mitigate this issue, Ferro and Sanderson also proposed a methodology to estimate the probability that each single comparison between two runs leads to publication bias, in order to be able not only to publish those results which perform best and are significantly different, but also those that are less likely to originate publication bias.

3.3 Measures for Reproducibility

Breuer et al. [2020] present a set of measures for reproducibility of IR exper-

²⁹Note that, the publication bias differs when computed with different measures due to the discriminative power of each measure [Sakai, 2006]. The discriminative power of a measure estimates the performance difference required to achieve a given significance level. A smaller performance difference, means a more sensitive measure.

iments, i.e., to quantify the extent to which an attempt to reproduce earlier work is successful. These measures are categorized by their level of granularity: *ordering of documents*, *effectiveness error measures*, *statistical tests*, and *overall effect over a baseline*, as illustrated in Figure 3.3. All these measures can be computed in Python with the `repro_eval` library [Breuer et al., 2021].

Notation Let r be the original run and $r_t = (d_1, \dots, d_n)$ the ranking of documents for topic $t \in \mathcal{T}$, where d_k is the document ranked at rank position k , and the set of topics is $\mathcal{T} = \{1, \dots, T\}$.³⁰ Let r' denote the reproduced run and $r'_t = (d'_1, \dots, d'_n)$ the reproduced list of documents for topic t . Any IR effectiveness measure, like AP or nDCG [Järvelin and Kekäläinen, 2002], is denoted by M . The average score across topics is denoted as: $\overline{M}(r) = 1/T \sum_{t=1}^T M(r_t)$.

Some reproducibility measures require pairs of runs as input: an advanced run and a baseline run, where the advanced run is supposed to perform better than the baseline run with respect to an effectiveness measure M . In these cases, r_A and r_B denote the original advanced and baseline runs, similarly r'_A and r'_B are the reproduced advanced and baseline runs.

Ordering of Documents *Kendall's τ Union (KTU)* [Ferro et al., 2018a, 2019a] and *Rank-Biased Overlap (RBO)* [Webber et al., 2010] can be used to compare the original and reproduced runs at the ranking level, i.e., to quantify how similar the ranking of documents is, for each topic. *Kendall's τ Union (KTU)* is a derivation of Kendall's τ [Kendall, 1945] that can be computed when the two input rankings contain different documents. This is often the case for IR because two systems can retrieve different documents for the same topic.

For each topic, KTU_t is computed as follows: (1) compute the union of the original and replicated rankings: $r_t \cup r'_t$. This is done by concatenating r'_t at the end of r_t and then removing duplicated documents, starting from the bottom of the concatenated ranking; (2) create a list of rank positions l_t for the original run r_t , where the item at position k in l_t is the rank position of document d_k in the union $r_t \cup r'_t$. This is the rank position of the document at rank position k from the original run in the union of the original and reproduced runs. Similarly, l'_t is the list of rank positions for the reproduced run r'_t ; (3) compute Kendall's τ between the list of rank positions as follows:

$$\text{KTU}_t(r_t, r'_t) = \tau(l_t, l'_t) = \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}}$$

where P is the total number of concordant pairs, Q is the total number of discordant pairs, and U and V are the number of ties in l_t and l'_t respectively. KTU is the average across topics:

$$\text{KTU}(r, r') = \frac{1}{T} \sum_{t=1}^T \tau(l_t, l'_t)$$

³⁰To avoid over-complicating the notation, we use the set of topic indices to denote the set of topics itself.

KTU ranges in $[-1, 1]$ with the same interpretation as Kendall’s τ [Kendall, 1945]: -1 denotes perfect anti-correlation, 0 denotes no correlation, and 1 denotes perfect correlation.

Rank-Biased Overlap (RBO) can handle rankings with different documents and is top-heavy, i.e., a greater penalty is imposed when two documents are swapped at the beginning of the ranking than at the end. Intuitively, RBO computes the proportion of overlap between the original and reproduced rankings for each rank position. Then, the overlap is discounted by the rank position because the closer the rank position is to the end of the ranking, the higher the probability to have a large overlap. RBO is computed as follows:

$$\text{RBO}_t(r_t, r'_t) = (1 - \phi) \sum_{k=1}^{\infty} \phi^{k-1} \cdot A_k \quad (3.1)$$

where ϕ is a parameter in $[0, 1]$ which controls the measure top-heaviness: the lower the ϕ , the higher the weight assigned to variations at the top of the ranking; A_k is the overlap between the original and reproduced rankings up to position k , it is defined as the cardinality of the intersection between r_t and r'_t up to rank position k divided by k . RBO is reported as an average across topics:

$$\text{RBO}(r, r') = \frac{1}{T} \sum_{t=1}^T \text{RBO}_t(r_t, r'_t)$$

Note that the sum in Equation (3.1) goes up to ∞ , indeed RBO assumes that rankings of documents have infinite length so that the geometric series converges to $1/(1 - \phi)$, and RBO ranges in $[0, 1]$.

Effectiveness Error Measures *Average Retrieval Performance (ARP)*, *Root Mean Square Error (RMSE)* [Ferro et al., 2018a, 2019a], and *Normalized Root Mean Square Error (nRMSE)* [Maistro et al., 2023] can be used to quantify reproducibility in terms of effectiveness scores. These measures take as input the score of the original and reproduced runs computed with any IR effectiveness measures M , e.g., AP or nDCG, and they disregard the actual ranking of documents.

Average Retrieval Performance (ARP) represents the simplest approach to quantify reproducibility, by comparing the difference between the average score across topics for the original and reproduced runs:

$$\text{ARP}(r, r', M) = \frac{1}{T} \sum_{t=1}^T M(r_t) - \frac{1}{T} \sum_{t=1}^T M(r'_t) = \overline{M}(r_t) - \overline{M}(r'_t) \quad (3.2)$$

Since most of IR effectiveness measures range in $[0, 1]$, ARP ranges in $[-1, 1]$, where 0 represents perfect reproducibility, negative scores mean that the reproduced run performs better than the original run, conversely positive scores mean that the reproduced run performs worse than the original run.

Root Mean Square Error (RMSE) is computed with per-topic scores to account for outliers, i.e., topics where there is a large difference between the performance of the original and reproduced runs. RMSE is computed as follows:

$$\text{RMSE}(r, r', M) = \sqrt{\frac{1}{T} \sum_{t=1}^T (M(r_t) - M(r'_t))^2} \quad (3.3)$$

RMSE returns non negative values, where 0 denotes perfect reproducibility and the higher the score the worse the reproducibility experiment. The upper bound of RMSE depends on the original runs and the IR effectiveness measure (see Equation (3.4) in the following).

Normalized Root Mean Square Error (nRMSE) is a normalized version of RMSE in Equation (3.3), with the normalization factor computed as follows:

$$\text{max-RMSE}(r, M) = \sqrt{\frac{1}{T} \sum_{t=1}^T (\max\{M(r_t), 1 - M(r_t)\})^2} \quad (3.4)$$

therefore

$$\text{nRMSE}(r, r', M) = \frac{\text{RMSE}(r, r', M)}{\text{max-RMSE}(r, M)} \quad (3.5)$$

For any IR effectiveness measures with values in $[0, 1]$, nRMSE ranges in $[0, 1]$, where 0 represents perfect reproducibility and 1 is the worst possible outcome.

Statistical Test Breuer et al. [2020] use statistical testing to compare the original and reproduced runs. A paired two-tailed *t*-test is run between the per-topic scores of the original and reproduced runs computed with any IR effectiveness measures. The *p*-value returned by the statistical test is used to quantify reproducibility. It ranges in $[0, 1]$, where the smaller the value, the higher the confidence in rejecting the null hypothesis, i.e., the scores of the original and reproduced runs come from the same distribution. Therefore, the smaller the *p*-value, the stronger the evidence that the original and reproduced runs are significantly different, meaning that the reproducibility experiment is not successful.

Overall Effect over a Baseline *Effect Ratio (ER)* and *Delta Relative Improvement (DeltaRI)* [Sakai et al., 2019, 2020, 2022] compare the improvement of an advanced run over a baseline run. Both measures consider an original experiment where an advanced run is reported to perform better than a baseline run. Then, the same improvement is expected when comparing the reproduced advanced and baseline runs.

Effect Ratio (ER) computes the overall effect of the advanced run over the baseline run as follows:

$$\text{ER}(r_A, r_B, r'_A, r'_B, M) = \frac{\frac{1}{T} \sum_{t=1}^T M(r'_{A,t}) - M(r'_{B,t})}{\frac{1}{T} \sum_{t=1}^T M(r_{A,t}) - M(r_{B,t})} \quad (3.6)$$

where the numerator in Equation (3.6) is the improvement for the reproduced experiment averaged across topics, and the denominator is the improvement for the original experiment averaged across topics. ER can assume both negative and positive values. The best reproducibility score is $ER = 1$, meaning that the same effect can be observed for the original and reproduced experiments. The farther ER scores are from 1, the worse is the reproducibility experiment. If $ER \leq 0$ the baseline reproduced run outperforms the advanced reproduced run, i.e., the reproducibility experiment is not successful and the opposite effect is observed; conversely if $ER > 1$ the improvement of the reproduced advanced run over the baseline run is larger than in the original experiment; if $0 < ER < 1$ an improvement in the reproduced experiment is observed, but to a smaller extent when compared to the original experiment.

Delta Relative Improvement (DeltaRI) complements ER by accounting for the similarity of relative scores, not only the similarity in terms of observed improvement. This is done by computing the *Relative Improvement (RI)* of the advanced run over the baseline run as follows:

$$RI(r_A, r_B, M) = \frac{\overline{M}(r_A) - \overline{M}(r_B)}{\overline{M}(r_B)} \quad (3.7)$$

and then considering the difference in relative improvements between the original and reproduced pairs of runs:

$$\Delta RI = RI(r_A, r_B, M) - RI(r'_A, r'_B, M) \quad (3.8)$$

DeltaRI ranges in $[-1, 1]$, where 0 means perfect reproducibility because the relative improvement of the advanced run over the baseline run is the same for the original and reproduced runs. The closer ΔRI is to 1 or -1 , the lower the reproducibility. If $\Delta RI < 0$, the relative improvement with the reproduced experiment is larger than the original experiment, conversely, if $\Delta RI > 0$ the relative improvement is lower for the reproduced experiment.

ER and DeltaRI can be visualized in combination by plotting ER on the x -axis against DeltaRI on the y -axis, as shown in Figure 3.4. Perfect reproducibility is obtained at $(1, 0)$, when both the effect over the baseline and the relative improvement are the same for the original and reproduced experiment. The rest of the plane can be divided in 4 regions as follows:

- Region 1, $ER > 0$ and $\Delta RI > 0$: the reproducibility of the effect is partially preserved because the reproduced advanced run outperforms the reproduced baseline run, but the relative improvement is smaller than in the original experiment;
- Region 2, $ER < 0$ and $\Delta RI > 0$: the reproducibility experiment failed both in terms of effect and relative improvement because the reproduced baseline run outperforms the reproduced advanced run and the relative improvement is smaller than in the original experiment;
- Region 3, $ER < 0$ and $\Delta RI < 0$: the reproducibility experiment failed with respect to the effect, even if the relative improvement is larger for the reproduced runs;

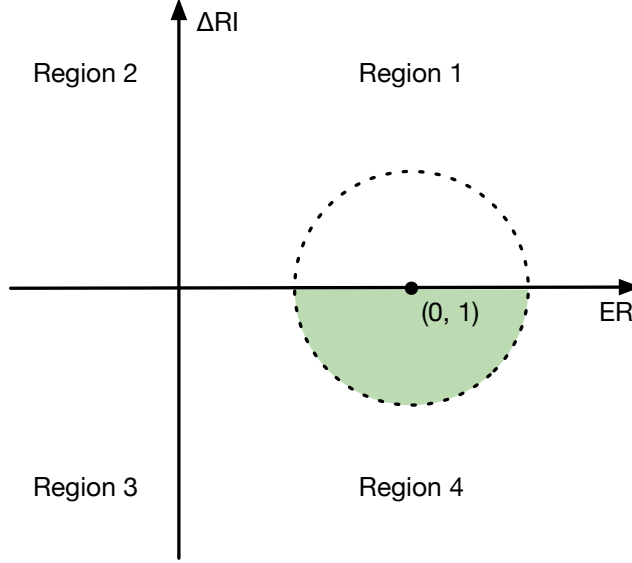


Figure 3.4: Reproducibility regions defined by ER and DeltaRI. Perfect reproducibility is achieved in $(0, 1)$, best if in Region 4 (green area).

- Region 4 $ER > 0$ and $\Delta RI < 0$: both the reproducibility of the effect and relative improvement are demonstrated.

In a practical scenario, the closer a point to $(1, 0)$, the higher the reproducibility of the experiment. However, reproducibility is more faithful if the obtained point is in Region 4, rather than in Region 1, where even if the experiment is not perfectly reproduced, the reproduced advanced run is performing better than the reproduced baseline run both in terms of size of the effect and relative improvement.

3.3.1 Guidelines to Choose a Reproducibility Measure

Breuer et al. [2020], Maistro et al. [2023] present some guidelines to choose the most appropriate reproducibility measures. The first selection criteria regards data availability:

- Case 1: there is access only to the average score over the topics, for example, as reported in the original paper;
- Case 2: there is access to the actual run from a repository, i.e., the original ranking of documents for each topic.

In case 1, the only measure that can be computed is ARP, i.e., a comparison of the average scores. This is a straightforward solution which might be misleading, for example, when the per topic scores are very different but the overall

average scores for the original and reproduced runs are similar. Moreover, ARP is affected by the underlying IR effectiveness measure: in some cases it is easier to achieve better reproducibility scores with recall-oriented measures, e.g., AP, or with large cut-off, e.g., nDCG@100. This is because precision oriented measures tend to have higher variance. Note that, in case 1, the choice of the IR effectiveness measure is restricted to those reported in the original paper. The lack of access to the actual run prevents the computation of any other IR effectiveness measure or to test different cut-offs.

In most of the cases, research papers report the score of a run and those from several baselines. This allows to compute ER and DeltaRI also in case 1. These measures focus on the reproducibility of the effects or relative improvements, rather than the actual scores. This means that the per topic scores can be very different, but leading to a similar overall improvement of the advanced run over the baseline run in the original and reproduced experiments. Moreover, similarly to ARP, ER and DeltaRI are also affected by the choice of the effectiveness measure and the evaluation cut-off.

Case 2 allows to compute any reproducibility measure with any underlying IR effectiveness measures. Among reproducibility measures, those based on the ordering of documents are the strictest. KTU requires that the ordering of documents for the original and reproduced run is exactly the same for each topic. This ensures that the experiment is perfectly reproduced, leading to the same user experience. However, experimental results show that it is very hard to achieve high reproducibility scores with KTU, with correlation scores that tend to be close to 0 [Breuer et al., 2020, Maistro et al., 2023]. RBO is more lenient, since variations that occur close to the bottom of the ranking are weighted less. Therefore, RBO can be a valuable alternative when it is more important to have good reproducibility at the top of the ranking, for example for navigational queries.

RMSE and nRMSE compare the per-topic effectiveness scores and account for outliers. They are more accurate than ARP because large differences even on a single topic will affect the overall reproducibility score. nRMSE is easier to interpret than the standard RMSE because it ranges in $[0, 1]$ independently of the original run, while RMSE score depends on the original run and on the number of relevant documents retrieved by that run. However, both RMSE and nRMSE are affected by the choice of the effectiveness measure as for ARP: it is easier to achieve better reproducibility scores with recall-oriented measures and large cut-offs. Finally, note that perfect reproducibility in terms of RMSE and nRMSE does not mean that the original and reproduced runs are exactly the same but only that there is a one-to-one correspondence between the positions of relevant and non-relevant labels in the original and reproduced runs. In principle, the reproducibility run can rank different documents, still resulting in a very low RMSE score but conveying a completely different user experience.

Statistical tests are more general than RMSE and nRMSE. The p -value informs on the statistical difference between the original and reproduced runs but it is harder to interpret than the other measures. Experimental results show that p -values are affected by the choice of the effectiveness measures and are

very sensitive even to small changes in the runs. However, this measure can be used whenever statistical requirements are needed to support reproducibility.

3.4 Measures for Replicability

Quantifying replicability is more challenging than reproducibility because it is not possible to compare directly a replicated ranking or score against the original version. Indeed, the replicated run is obtained on a different test collection, with different documents and topics. Therefore, among the reproducibility measures presented in Section 3.3, only measures that consider *statistical tests* and the *overall effect over a baseline* can be used for replicability [Breuer et al., 2020].

Notation The notation for reproducibility is extended to account for the test collection used by the replicability experiment. Let C be the original collection, with topics $\mathcal{T} = \{1, \dots, T\}$ and D the collection for the replicability experiment with topics $\mathcal{S} = \{1, \dots, S\}$. Note that $\mathcal{T} \neq \mathcal{S}$ and in general we might have a different number of topics, i.e., $T \neq S$. The original run r is computed on the collection C and the replicated run r'' is computed on D . The advanced replicated run is r''_A and the baseline replicated run is r''_B .

Statistical Test For replicability, an unpaired two-tailed t -test is carried out to compare the original and replicated runs [Breuer et al., 2020]. The interpretation is the same as for reproducibility: the smaller the p -value the stronger the evidence that the original and replicated runs are different, i.e., that the replicability experiment is not successful.

Overall Effect over a Baseline ER and DeltaRI [Sakai et al., 2019, 2020, 2022] quantify replicability in terms of the difference in performance between the advanced and baseline runs in the original and replicated experiment. The intuition is that if an advanced run outperforms a baseline run on a collection C , the replicated advanced run will outperform the replicated baseline run on a different collection D , with the same size of the effect and relative improvement.

For replicability, ER is computed as follows:

$$\text{ER}(r_A, r_B, r''_A, r''_B, M) = \frac{\frac{1}{S} \sum_{s=1}^S M(r''_{A,s}) - M(r''_{B,s})}{\frac{1}{T} \sum_{t=1}^T M(r_{A,t}) - M(r_{B,t})} \quad (3.9)$$

with the same interpretation as in Equation (3.6): perfect replicability for $\text{ER} = 1$, failure of the replicability experiment for $\text{ER} \leq 0$; replicability to some extent if $0 < \text{ER} < 1$ because the advanced run outperforms the baseline run on the new collection but to a smaller extent than in the original collection; successful replicability if $\text{ER} \geq 1$ because the advanced run outperforms the baseline run on the new collection; the farther ER from 1 the larger the extent of the improvement of the advanced run over the baseline when compared to the original experiment and the lower the replicability.

DeltaRI can be computed for replicability as follows:

$$\Delta\text{RI} = \text{RI}(r_A, r_B, M) - \text{RI}(r''_A, r''_B, M) \quad (3.10)$$

and its interpretation is the same as for reproducibility: $\Delta\text{RI} = 0$ means perfect replicability, the closer ΔRI is to -1 the lower the replicability with a larger relative improvement in the replicability experiment than in the original experiment; conversely, the closer ΔRI is to 1 the lower the replicability with a lower relative improvement in the replicability experiment. Finally, ER and DeltaRI can be visually combined as in Figure 3.4, with the same interpretation of the 4 different regions (Section 3.3).

3.4.1 Guidelines to Choose a Replicability Measure

Breuer et al. [2020] present an analysis of replicability measures. As for reproducibility measures (Section 3.3.1), the first criterion to select a replicability measure depends on the availability of data. In case 1, computing ER and DeltaRI requires access to the scores averaged across topics for the original baseline and advanced runs. In case 2, statistical tests require the per topic scores of the original run on 2 different collections.

Note that the shortcomings of ER, DeltaRI and statistical tests, i.e., dependency on the IR effectiveness measure and interpretability of the scores, are further exacerbated for replicability. Experimental results show that these measures are very sensitive to the changes in the collection, i.e., different topics and documents, leading to poor replicability results even when this is not necessarily the case. For example, Breuer et al. [2020] show that conducting an unpaired t -test between the score distributions on two different collections results in very low p -values, thus very low replicability, even when these distributions are generated by exactly the same system.

3.5 Reproducibility Initiatives

The following reviews reproducibility efforts realized in the form of dedicated paper tracks at conferences, community events like seminars, tutorials, workshops, and other collaboration platforms.

Reproducibility Workshops at SIGIR The workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) [Arguello et al., 2015b,a] was held at SIGIR 2015. The aim of the workshop was twofold: it provided a venue where negative results could be published, and an open source challenge on reproducibility for IR. Negative results included challenges in reproducing or replicating previously published results as well as generalizability and inexplicability. Generalizability was defined as obtaining comparable results in different tasks, e.g., text and image retrieval. Inexplicability referred to a well known hypothesis that is expected to be true on different domains or tasks but is disproved by experimental results.

The reproducibility challenge focused on the problem of weak baselines, i.e., IR systems that are below the median of TREC scores on a given collection [Armstrong et al., 2009b]. When such baselines are selected, it is easier to achieve statistical significance but similar results will not hold or add up with a stronger baseline. Therefore, the organizers proposed to index and search the same collections, from TREC and CLEF, on the same device (an Amazon EC2 instance) and compare the results. The participants were asked to provide the source code to index and search the selected collections, then the source code was run by a third person, who was not the code creator. The results from the challenge showed large variations among the participants’ submissions in terms of efficiency, measured using indexing time and size, and average time to search. Variations in terms of effectiveness, with AP@1000, were not as large as for efficiency, however, there were statistical differences between the same model with the same parameters but implemented within different IR platforms. Overall, the results were different from those previously reported in the literature and the reason was attributed to the lack of parameter tuning.

The *Open-Source IR Replicability Challenge (OSIRRC)* workshop [Clancy et al., 2019] built on and extended the RIGOR workshop. Participants submitted reproducible systems and experimental pipelines using Docker-based containerization. Organizers created a lightweight toolkit, the “jig”, which defines standard interfaces and commands for data ingestion, indexing, and ranking, enabling easy integration of ad-hoc retrieval pipelines. The workshop produced a comprehensive library of Docker images that can be rerun together with the jig for reproducible experiments.

Reproducibility of Data-Oriented Experiments in e-Science The Dagstuhl seminar 16041 on Reproducibility of Data-Oriented Experiments in e-Science³¹ [Ferro et al., 2016a, Freire et al., 2016] was held in 2016 in Dagstuhl, Germany. Participants came from Computer Science but with diverse backgrounds, e.g., databases, IR, information visualization, semantics, bioinformatics, human-computer interaction, simulation, etc. The goal of the workshop was to gather experts in reproducibility from different areas so that reproducibility issues and approaches could be tackled from different perspectives. The primary outcome of the workshop was the PRIMAD model (see Section 3.1.3).

Besides PRIMAD, the working groups focused on challenges and solutions related to reproducibility. For example, a researcher can take several actions to ease reproducibility. Some of these actions are proactive: sharing the data and code as well as prototypes and complementing all these resources with appropriate documentation and metadata. Reactive actions are triggered by publications, i.e., independent reproducibility studies, systematic analyses, peer-reviews of artifacts. Finally, supporting actions include formulating shared tasks, challenges, and benchmarks, where experimental results are directly comparable.

In relation to each action, there is a plethora of reproducibility tools available

³¹<https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/16041>

online (see also Section 4) making it hard for a researcher to navigate and select the most appropriate tool. This, combined with the lack of practical guidelines or a standard for sharing data and source code hampers reproducibility. The community does not necessarily need the development of new tools, but existing tools can be used or adapted. ReproMatch³² represents an initial step to organize reproducibility tools within a taxonomy, allowing users to search for the best tool.

Not only are authors responsible of reproducibility but other actors are involved, e.g., reviewers, editors, publishers, institutions, funding agencies, etc. During the seminar, some working groups analysed obstacles and barriers for reproducibility, to name some: the lack of incentives and recognition, the time pressure to publish or review papers, the fear of having mistakes exposed, legal or copyright issues, etc. (see Section 3.2). A set of practical interventions and incentives to overcome these obstacles were proposed for different actors. For example, guidelines for authors and editors may be included in the call for papers along with criteria that can be used to decide when it is worth starting an independent reproducibility study. In terms of incentives to encourage reproducibility, motivation can be driven by the return of value. Examples include publicity and impact because a reproducible paper can possibly get more citations, or reusability because it is easier to reuse artifacts from a paper accompanied by data and code. Incentives to achieve reproducibility can be natural or moral, personal curiosity and willingness to contribute to social good, but also financial, with some sort of reward, and coercive, when reputation is damaged because of failure to reproduce.

Reproducibility of IR user studies was also one of the discussion points. Some obstacles and barriers to reproducibility are specific for user studies and in general to any paper that exploits user data, for example system log data. First, datasets may not be publicly available or easily shareable because of privacy and confidentiality issues. In many cases it is not possible to completely anonymise the data while preserving the features needed for reproducibility. User data can also suffer from volatility, e.g., A/B testing or live streams, and in general a user study may not be re-run with the same set of users. External validity is a further issue: user data are so multidimensional that it is hard to ensure some sort of generalizability of the results. Some of the proposed solutions were evaluation as a service [Hopfgartner et al., 2015, 2018], which allows to evaluate a system without direct access to the data, and checklists or well defined standards to design the experiments and data format.

Reproducibility Tracks at ECIR and SIGIR Both the *European Conference on Information Retrieval (ECIR)* and the *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)* host a reproducibility track, since 2015 and 2022 respectively. ECIR and SIGIR call for replicability and reproducibility papers, as well as papers that analyse and generalise prior work. The main focus of reproducibility papers is on the

³²<https://remram44.github.io/repromatch-new/>

process, e.g., reproducibility protocol, evaluation and lessons learnt, rather than on the outcome, whether the reproducibility experiment was successful or not.

Reproducibility tracks have gained popularity over the years. ECIR had 3 accepted reproducibility papers in its first edition, in 2015 and 9 accepted reproducibility papers in 2024. SIGIR started with 7 accepted reproducibility papers in 2022 and had 9 accepted reproducibility papers in 2024. Overall, accepted papers cover a variety of topics besides ad-hoc IR, to mention some, RS, stance detection, classification, summarization, evaluation, entity linking, etc. Based on authors’ self reported assessments, out of 67 ECIR reproducibility papers, 27 papers report successful reproducibility experiments (40%), 20 papers report unsuccessful reproducibility experiments (30%), 18 papers report mixed results (27%), i.e., only some claims could be reproduced or replicated, and the remaining papers provided resources or were comparative studies where no previous results were available to be used as reference.³³ In SIGIR, out of 23 reproducibility papers,³⁴ 7 papers report successful reproducibility (30%), 4 unsuccessful reproducibility (17%), 6 mixed reproducibility (26%), and the remaining papers did not have previous results to compare against or were re-evaluation with different measures, e.g., fairness.

CLEF, NTCIR, TREC REproducibility (CENTRE) With the aim of fostering reproducibility, *CLEF, NTCIR, TREC REproducibility (CENTRE)* was a shared task between CLEF, NTCIR and TREC. The goal was to reproduce and replicate the best and/or most interesting systems from previous editions of CLEF, NTCIR and TREC, using open source code and standard IR tools that could eventually be shared with the community. Moreover, the task focused on exploring possible ways to quantify and measure reproducibility and replicability (see Sections 3.3 and 3.4).

CENTRE ran for two editions at CLEF in 2018 [Ferro et al., 2018a,b] and 2019 [Ferro et al., 2019a,b], one edition at TREC 2018 [Soboroff et al., 2019], and three editions at NTCIR: NTCIR-14 [Sakai et al., 2019] in 2019; NTCIR-15³⁵ [Sakai et al., 2020] in 2020; and NTCIR-16 in 2021 [Sakai et al., 2022]. Unfortunately, in each edition of CENTRE, only a single team managed to submit reproducibility or replicability runs, with the exception of TREC 2018, when there were two submitting teams, one official and one unofficial, and NTCIR-15, when two distinct teams participated.

At CENTRE participants were challenged to reproduce and/or replicate a target system. For example, CENTRE@TREC2018 identified three target systems to reproduce: (task 1) a learning to rank approach enriched with complex features such as word embeddings, pseudo-relevance feedback and query parsing [Gurulingappa et al., 2017] submitted from the Merck team at TREC 2016

³³Note that this analysis does not include reproducibility papers accepted at ECIR 2017 and 2018 because those papers are not explicitly marked as “reproducibility papers” in the conference proceedings

³⁴Resource papers, that were accepted at the reproducibility track, are excluded from this analysis.

³⁵CENTRE was merged with We Want Web (WWW) at NTCIR-15 and NTCIR-16.

Clinical Decision Support (CDS) track [Roberts et al., 2017]; (task 2) an axiomatic retrieval model [Yang and Fang, 2014] submitted from the Delaware team the at TREC 2013 Web track [Collins-Thompson et al., 2014]; and (task 3) a risk-sensitive learning to rank approach [McCreadie et al., 2015] submitted from the Glasgow team at the TREC 2014 Web track [Collins-Thompson et al., 2015].

Overall, all CENTRE editions show that reproducibility is challenging. For example, at CENTRE@TREC2018, the runs submitted to task 1 could reproduce the original ranking of systems, but not the original scores [Di Nunzio and Marchesin, 2019]. For task 2, the submitted runs could not reproduce the original ranking of systems nor the original scores [Yang and Lin, 2019a]. Extra relevance judgements were collected for task 2 because the pools used in the Web track at TREC 2013 were quite shallow and the new reproducibility runs retrieved many unjudged documents. Using the newly judged documents yielded low agreement with respect to the original judgements and a different ranking of systems, highlighting yet another factor, i.e., incompleteness of relevance judgements, that can make reproducibility hard to achieve.

ACM SIGIR Artifact Badging ACM SIGIR inaugurated the ACM “Artifact Review and Badging” policy (v1.1, 20 Aug 2020) for TOIS, SIGIR, CHIIR and ICTIR to recognise authors who provide implementations and/or datasets that enable replication or reproduction, and to acknowledge reviewers and reproducers.³⁶ Artifact evaluation is available only for accepted papers after peer-review and is initiated by the authors submitting the artifact for an additional review. The process is fully asynchronous with respect to the prior peer-review and may be triggered at any time (preferably within three years of publication). The artifact reviews are open so authors and reviewers can interact.

Each artifact is jointly assessed by one junior and one senior member of the Artifact Evaluation Committee. The artifact review judges only whether the artifact supports the claims made in the accepted paper (it is not a re-review of scientific quality). Accepted artifacts receive DOI entries in the ACM Digital Library and become citable. The overall aim is to make artifacts discoverable, citable, and usable so as to accelerate reuse, reduce reimplementations, enable reliable comparisons, and foster reproducibility and cumulative progress in IR research.

Tutorials and Lectures On Reproducibility Lucic et al. [2022a] presented a tutorial on reproducibility at SIGIR 2022. Besides terminology, definitions and examples of reproducibility across different disciplines, the tutorial offered pointers to practical solutions to ease reproducibility. Some examples are: reproducibility checklists [Pineau et al., 2021],³⁷ the papers with code repository,³⁸ li-

³⁶<https://sigir.org/general-information/acm-sigir-artifact-badging/>

³⁷<https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf>

³⁸<https://github.com/paperswithcode>

braries for reproducibility (Weights & Biases,³⁹ CodaLab,⁴⁰ repo2docker [Forde et al., 2018],⁴¹ etc.), and finally the idea of using reproducibility as a teaching aid [Yildiz et al., 2021].⁴²

A lecture on reproducibility was held at the *European Summer School in Information Retrieval (ESSIR)* 2024 edition.⁴³ The lecture provided an overview of Reproducibility in IR as well as practical guidelines and tips. For example, how to publish and polish source code, how to fix random seeds to control for stochasticity, how to choose proper baselines, and how to write reproducibility papers.

Finally, ECIR 2025 [Ferrara et al., 2025] offered another tutorial on reproducibility. This tutorial focused on reproducibility not only in IR, but also in *Recommender Systems (RS)* and offered hands-on sessions with Jupyter Notebooks.

³⁹<https://wandb.ai/site/>

⁴⁰<https://codalab.org/>

⁴¹<https://repo2docker.readthedocs.io/en/latest/>

⁴²<https://reproducedpapers.org/>

⁴³<https://drive.google.com/file/d/1KfLxRxtoygzEgnDbGXfW0yv52LYHGCvZ/view>

Table 3.1: Reproducibility barriers reported in the proceedings of the ECIR Reproducibility Track from 2015 to 2024.

Category	Barrier	Reference(s)
Sharing	Missing source code	<i>Nearly all papers</i>
	Missing data	[Hagen et al., 2015, Pothast et al., 2016]
	Unpublished resources (other than data and code)	[Lin et al., 2016, Pothast et al., 2016, Šihinec et al., 2018]
	External platform unavailable (e.g., API)	[Hasibi et al., 2016]
	Outdated and unavailable resources	[Fernández-Pichel et al., 2021]
	Updated operating system, compiler, runtime	[Lin and Zhang, 2020]
	Updated (web) corpora (without fixed snapshot)	[MacAvaney et al., 2022b]
Technical	Tie breaks unspecified	[Ferro and Silvello, 2015]
	Hyperparameters not reported	[Ferro and Silvello, 2015, Li et al., 2022]
	No dependency management: missing versions of software libraries	[Pothast et al., 2016, Silvello et al., 2018, Kusa et al., 2022]
	Ambiguity in the model architecture description	[Dir et al., 2018, Lajewska and Balog, 2023]
	Missing implementation details / lack of pseudo-code	[Silvello et al., 2018, Mackenzie et al., 2019, Papariello et al., 2020, Lajewska and Balog, 2023]
	Ambiguity in the training process	[Dir et al., 2018]
	No random seed reported	[Yang and Lin, 2019b, Li et al., 2022]
	Missing code documentation	[Rao et al., 2015]
	Ambiguity in (corpus) data preparation	[Tamber et al., 2023]
	Missing cross-validation	[Rao et al., 2015]
	Weak and improperly-tuned baselines	[Rao et al., 2015, Mackie et al., 2016, Althammer et al., 2021, Moreo and Sebastiani, 2021, Kusa et al., 2022, Ma et al., 2022, Pérez Maurera et al., 2022, Shohrad and Jannach, 2024]
	Lack of contextualization, evaluations are not exhaustive, selected baselines	[Pothast et al., 2016, Berrendorf et al., 2021, Kiesel et al., 2021, Boratto et al., 2022, Balloeu et al., 2023, Liu et al., 2024b]
Methodological and experimental	Differences in indexing/pre-processing methods and implementations	[Lin et al., 2016, Wang et al., 2022a]
	Same method, different variants and implementations	[Kamphuis et al., 2020]
	Hyperparameters tuned on test set	[Berrendorf et al., 2021]
	Ambiguity between paper and implementation	[Berrendorf et al., 2021]
	No generalizability with other data or different task	[Manotamrueksa et al., 2019, Yang and Lin, 2019b, Bi et al., 2024, Silvello et al., 2018, Bhatnacharya et al., 2019, Efimov et al., 2023]
	Data bias (e.g., popularity bias or query distribution)	[Boratto et al., 2019, Kowald et al., 2020, Neophytou et al., 2022, Benedetti and Ruggiero, 2023]
	Algorithmic bias	[Boratto et al., 2019, Neophytou et al., 2022]
	Conclusions cannot be confirmed by the reimplementations	[Schliski et al., 2020, Althammer et al., 2021, Shrestha and Spezzano, 2021, Blocker and de Rijke, 2022, Penha and Hauff, 2023]
	Conclusions do not reproduce in practice	[Mukherjee et al., 2021, Cross et al., 2023]
	Lack of robustness (with other data or settings)	[Wang et al., 2021]
Cultural	No reply from the authors	[Papariello et al., 2020]
	Venue does not enforce reproducibility	[Lajewska and Balog, 2023]

Chapter 4

Best Practices for Reproducibility

The previous section outlined the general scope of reproducibility in IR research, gave an overview of the major barriers, and introduced a framework of measures for quantifying the degree of reproducibility and replicability when an experimental setup must be reimplemented. In contrast to this retrospective perspective, which challenges reproducers to recreate the original experiments, **this section reviews existing solutions and toolkits that enable researchers and practitioners to proactively prepare an experimental setup for future reuse.**

Primarily, a proactive approach to reproducibility focuses on sharing the experimental resources. As discussed earlier, resource sharing is one of the fundamental first steps to make experiments reusable and repeatable by others. To this end, this section outlines good practices for sharing the resources, including software, data, and related artifacts. The computational sciences are particularly well positioned in this regard, as most of the experimental components can be made digitally reusable with little or no additional cost.

Referring again to the PRIMAD model, we can identify four technical components that are convenient to share: the *platform*, the *method* and its *implementation*, and the *data*. We will discuss ways and the currently best approaches to share these components in Sections 4.1, 4.2, and 4.3, respectively. In this context, the IR community benefits from a rich ecosystem of reusable open source software. We therefore provide an overview of state-of-the-art open source software toolkits that support IR practitioners in their reproducibility efforts.

However, the mere existence of publicly available resources does not ensure their correct reuse if they are insufficiently documented. Likewise, less technical aspects, i.e., the remaining two PRIMAD components, the *research goal* and the *actor*, play a crucial role in the successful reproduction of experiments. Accordingly, Section 4.4 discusses approaches for documenting experiments, recording data provenance and experimental workflows, and annotating experimental data

with appropriate metadata and protocols.

Importantly, considerations of reproducibility should not be deferred until the final stages of an experiment but should be integrated from the earliest phases of its conceptualization. For this reason, each section concludes with a summary of good practices and actionable recommendations that can serve as guidelines or reference points, supporting IR practitioners in developing reproducible implementations of computational experiments from the outset.

Finally, this section concludes with a discussion of the FAIR principles for data and resource management. We outline directions for future work and consider pathways for advancing IR toward open science. In this context, we review existing solutions and assess the extent to which they support IR research in meeting the requirements of the FAIR principles, while also identifying opportunities to improve experimental practices in the interest of enhanced reproducibility.

4.1 Sharing the Computational Platform

The platform mainly comprises the hardware, operating system, and underlying software components on which a computational experiment is implemented. Platform characteristics can be critical to reproducibility if any of these components are unavailable or have changed when an experiment is rerun. For example, updates to the operating system may prevent the successful execution of previously functional software, while high computational demands or specialized hardware requirements can incur costs that render reproduction attempts infeasible. Moreover, nondeterministic hardware behavior may also lead to divergent experimental outcomes.

Lin and Zhang [2020] revalidated software artifacts four years after their initial release [Lin et al., 2016] and reported substantial limitations in their reusability. Specifically, they were able to reproduce identical results for only one of seven retrieval systems when rerunning the original experiments on newer hardware and an updated operating system. The remaining reproduction attempts failed due to runtime exceptions, segmentation faults, or compilation errors, or depended on external resources that were no longer available.

Likewise, high computational requirements and associated costs can pose a barrier to successful reproduction, especially for researchers who do not have access to adequate hardware or the expertise to operate it [Raff, 2019]. Notably, LLMs rely on tremendous amounts of data and require specialized hardware components like GPUs to make the training and parameterization of the model weights feasible. For certain experiments, these hardware requirements can impose a barrier, effectively leaving any further reproducibility attempts out of consideration.

Beyond availability and cost, the hardware components themselves can directly affect reproducibility. For instance, GPU manufacturer NVIDIA acknowledges that bitwise reproducibility is not guaranteed even when the same GPU

architecture is used, due to nondeterministic floating-point rounding behavior.¹ Ferro et al. [2019d] demonstrated the impact of such inherent GPU-induced randomness on the reproducibility of ranking results when no mitigating measures are applied.

From a general perspective, the most straightforward way to ensure a reusable platform would be to preserve the physical computing environment, including all digital assets and resources, exactly as used in the original experiment. However, this approach is often impractical or infeasible due to logistical constraints and scalability barriers [Ivie and Thain, 2018].

Nevertheless, it is possible to replicate the operating system and software environment with solutions such as *virtualization*, *containerization*, the use of *functional package managers*, or reliance on *shared testbeds* hosted on *external platforms*. In the following, we review these approaches, which are also summarized in Table 4.1, by categorizing different types of platform-sharing solutions and comparing their respective strengths and limitations.

Virtualization

VMs emulate physical computing devices and provide a means of making the operating system and kernel portable across different (physical) hardware platforms and experimental environments [Roegiest and Cormack, 2016]. More specifically, VMs capture snapshots of an operating system’s state, including all installed software components and configuration settings. As a result, code, data, and the execution environment can be shared together as a single image file.

In addition, VMs are highly isolated from the host operating system running the hypervisor or virtual machine monitor, and are therefore generally considered more secure than alternatives such as containers. However, this level of isolation comes at a cost. Virtual machines incur significant technical overhead, as the operating system is packaged in its entirety. Consequently, many files unrelated to the experiment are included, resulting in comparatively large image sizes. Moreover, executing virtual machines is computationally expensive due to the additional layer of hardware emulation.

Containerization

A more lightweight approach to preserving the platform is made possible by containerization or OS-level virtualization [Boettiger, 2015, Nüst et al., 2020]. A popular platform for running containers is Docker, although other implementations, such as Podman² or LXC³, exist and are listed in Table 4.1.

Containers encapsulate the application and its dependencies in an isolated user space, allowing the application to run consistently across different computing environments. In contrast to virtual machines, containers run in isolated

¹<https://github.com/NVIDIA/framework-reproducibility>

²<https://podman.io/>

³<https://linuxcontainers.org/lxc/>

Table 4.1: Platform sharing solutions and example tools for reproducibility.

Type	Examples	Strengths	Limitations
Virtualization	☑ KVM	✓ highest degree of isolation	✗ high computational overhead due to additional runtime
	☑ Hyper-V	✓ share entire state of the experimental setup	✗ unnecessary artifacts may be shared
	☑ VirtualBox		
Containerization	☑ Docker	✓ high portability across different operating systems	✗ fair computational overhead due to additional runtime
	☑ Podman	✓ isolated environment but efficient use of host resources	✗ image builds can be non-deterministic
	☑ LXC/☑ LXD		
Functional Package Managers	☑ Nix(OS)	✓ deterministic package builds	✗ limited portability, scalability, and orchestration
	☑ Guix	✓ lightweight rollbacks	✗ limited isolation from the host
	☑ Spack	✓ efficient use of host resources	
Shared Testbeds and Platforms	☑ CodaLab	✓ low technical overhead for original authors and reproducers	✗ little control of the underlying software and hardware stack
	☑ Binder	✓ notebooks better integrate a paper and the implementation	✗ limited scope of programming languages, hardware, etc.
	☑ Colab		

environments but share some of the underlying host’s resources, which imposes less computational overhead.

Docker-based container images are specified with so-called Dockerfiles, which specify layered software components required to run a containerized application. This approach increases transparency regarding the implementation dependencies and clarifies which components are essential for executing the experiment.

For many applications, pre-built images are readily available. For example, to run a Java application, one can use a base image containing a Java runtime environment and specify only the additional commands needed to execute the application. Containers typically require fewer resources than virtual machines, including reduced memory footprints and lower CPU utilization.

Although it is generally difficult to draw definitive conclusions about the long-term preservability of virtual machines and containers, notable differences exist between them regarding the comprehensiveness of preserved platform components. Archiving an experiment using a Dockerfile, which is a single text-based specification, does not guarantee that all required platform layers will remain accessible in the future [Nüst et al., 2020]. Providing a pre-built Docker image that includes all underlying platform layers via public repositories such as DockerHub is a more sustainable approach. Nonetheless, virtual machines remain more comprehensive, as their corresponding images are typically fully self-contained.

Functional Package Managers

As an alternative to virtual machines and software containers, Linux-based functional package managers like Nix [Dolstra et al., 2004] or Guix [Courtès, 2013] are viable alternatives. Like containers, these systems adopt a declarative configuration approach, in which the entire system configuration, including installed packages, their dependencies, and precise versions, is specified in a single file. These package managers employ a functional build language, conceptually similar to functional programming languages, in which software packages are treated analogously to variables, ensuring reproducible builds and minimizing unintended side effects.

NixOS [Dolstra and Löh, 2008] is an immutable Linux distribution built on the package manager Nix. Once installed, its system state cannot be changed during runtime. Updates and modifications create new instances or snapshots, allowing users to easily roll back to previous configurations if needed. As an alternative to the Nix/NixOS ecosystem, GNU Guix [Courtès, 2013] offers similar functionalities while adhering to the principles of free software (Nix also packages non-free software which could pose a problem to reproducibility). Like Nix, Guix also allows users to manage software and system configurations declaratively but is based on Scheme — a dialect of the Lisp family of programming language — instead of the Nix language.

The primary advantages of functional package managers are deterministic package builds, achieved through explicit declarations of all dependencies. Additionally, package builds occur in isolated environments, ensuring that changes

to the global system do not affect the state of the packages or their underlying dependencies. Compared to containers or virtual machines, functional package managers impose lower computational overhead, as they do not require dedicated runtimes. Instead, built packages are directly integrated with the host system. Storage efficiency is also improved, since containerized builds can sometimes duplicate dependencies despite using layered architectures. Furthermore, rolling back to earlier package versions is typically more efficient than rebuilding an entire container.

On the downside, functional package managers do not provide the same level of isolation as virtual machines or containers. Portability can also be a concern, particularly if the experiment is to be reproduced on a closed or incompatible operating system. In addition, orchestrating multiple containers within a cluster can be more straightforward using dedicated tools such as Kubernetes. Consequently, the choice of platform-preservation solution depends on the specific requirements of the experiment, and no single approach is universally optimal.

Shared Testbeds and External Platforms

When sharing the platform using the previously outlined solutions is not feasible, experiments can be deployed on common platforms that are accessible to potential reproducers. Building on technologies such as virtualization and containerization, *Evaluation-as-a-Service (EaaS)* platforms enable researchers to submit retrieval systems (or other software), execute experiments, and reproduce the resulting outputs [Hopfgartner et al., 2018]. A well-known example of such a platform for scientific and computational competitions is CodaLab [Pavao et al., 2023]. Additionally, we review several IR-specific EaaS platforms in the next section.

These platforms reduce the technical burden of managing the underlying hardware and software environment by providing a centralized, web-accessible service. Researchers can focus entirely on implementing their experimental pipelines, which can then be easily shared and re-executed by others, facilitating reproducibility and broader reuse.

Alternatively, platforms for executable papers [Kluyver et al., 2016, Jimenez et al., 2017a, Bar and Wang, 2021] and cloud-based environments for code notebooks, such as Google Colab⁴ or Binder⁵, allow users to run experiments on web-hosted platforms. Similar to EaaS solutions, these platforms ensure that the same environment is used for both the original and reproduced experiments. Often limited to specific technologies, such as Jupyter Notebooks, they provide a convenient, low-overhead way to demonstrate and rerun computational experiments. However, this convenience comes at the cost of reduced control over the underlying software and hardware stack, which can occasionally introduce reproducibility issues, for example, when customization of the software stack is required or platform operators modify the environment.

⁴<https://colab.research.google.com/>

⁵<https://mybinder.org/>

Solutions and Actionable Recommendations

- **Virtualization or containerization** allow to preserve the partial or entire experimental platform and make it reusable for others [Clancy et al., 2019, Roegiest et al., 2015],
- **Functional package managers** with a declarative and functional build language for deterministic and reproducible package builds [Dolstra et al., 2004, Dolstra and Löh, 2008, Courtès, 2013],
- **Shared testbeds** and external platforms enforce the use of a common and shared platform and democratize access to computational resources [Armstrong et al., 2009a, Yang and Fang, 2016, Potthast et al., 2019], also allow to run executable papers and code notebooks [Kluyver et al., 2016, Jimenez et al., 2017a].

4.2 Sharing the Software and Source Code

Making implementations reusable for others extends beyond merely sharing code and artifacts; it begins at the earliest stages of development, even before the first line of code is written. In general, it is recommended to review existing software libraries and to build upon established open-source retrieval toolkits [Lin et al., 2016]. A typical retrieval pipeline consists of several standardized components, such as stop-word removal, lexical unit generation, and the retrieval method itself, among others. For each processing stage, multiple variants exist, many of which are already implemented in widely used retrieval toolkits and libraries. These libraries are typically mature, have undergone extensive review, are less prone to implementation errors, and facilitate comparability across experimental results.

Open Source IR Software Toolkits

The IR community has fostered a strong culture of software sharing, with many tools released as open source to support the implementation and evaluation of retrieval experiments. Table 4.2 provides an overview of IR research software, categorized according to the different purposes they serve. We do not advocate the exclusive use of any particular tool; rather, our aim is to present a representative overview of available software solutions.

In general, the use of well-curated open-source retrieval toolkits is advantageous, especially when the retrieval pipeline relies on commonly employed indexing and retrieval methods. These tools are typically developed collaboratively within the community, and their integrated features are actively maintained and validated through regression testing, which helps ensure correctness and reliability.

There is a plethora of different retrieval tools with a high degree of compatibility, with some of them building upon each other. Some commonly used retrieval and indexing software frameworks are Anserini [Yang et al., 2017],

Table 4.2: Open source experiment and evaluation information retrieval software.

Category	Software, Resources, and Publication
Lexical indexing and retrieval	➦ Anserini [Yang et al., 2017]
	➦ Pyserini [Lin et al., 2021a]
	➦ Terrier [Ounis et al., 2005]
	➦ Pyterrier [Macdonald et al., 2021]
	➦ Indri [Strohman et al., 2005]
	➦ PISA [Mallia et al., 2019]
	➦ Lucene, ➦ Solr, ➦ Elasticsearch
Neural retrieval	➦ MatchZoo [Guo et al., 2019]
	➦ Capreolus [Yates et al., 2020]
	➦ CHERCHE [Sourty et al., 2022]
	➦ PyGaggle [Pradeep et al., 2023a]
	➦ RankGPT [Sun et al., 2023]
	➦ RankLLM [Pradeep et al., 2023b]
Data collection management	➦ <code>ir_datasets</code> [MacAvaney et al., 2021]
	➦ Datamaestro [Piwowarski, 2020]
Index sharing	➦ CIFF [Lin et al., 2020]
Packaging	➦ OSIRRC jig [Clancy et al., 2019]
Evaluation	➦ <code>trec_eval</code>
	➦ <code>pytrec_eval</code> [Gysel and de Rijke, 2018]
	➦ <code>ranx</code> [Bassani, 2022]
	➦ <code>ir_measures</code> [MacAvaney et al., 2022a]
	➦ <code>trectools</code> [Palotti et al., 2019]
Benchmark suite	➦ BEIR [Thakur et al., 2021]
EaaS platform	➦ TIRA [Potthast et al., 2019]
	➦ TIREx [Fröbe et al., 2023]
	➦ KIMERA [Pasin and Ferro, 2025]
Conversational search	➦ Chatty Goose [Zhang et al., 2021]
	➦ DECAF [Alessio et al., 2023]
Others	➦ TARexp [Yang and Lewis, 2022]
	➦ <code>pybool_ir</code> [Scells and Potthast, 2023]
	➦ <code>ir_axioms</code> [Bondarenko et al., 2022b]

Terrier [Ounis et al., 2005], or Indri [Strohman et al., 2005]. These frameworks are favorably used for implementing lexical retrieval methods and standardized processing steps for generating an inverted index structure.

Python-based interfaces, such as Pyserini [Lin et al., 2021a] and PyTer-

rier [Macdonald et al., 2021], provide more accessible and flexible access to Anserini and Terrier, respectively, facilitating their integration into experimental workflows with Python. In addition to these widely used tools, several other retrieval frameworks exist; accordingly, Table 4.2 provides a non-exhaustive overview. For example, PISA [Mallia et al., 2019] focuses on efficiency-oriented retrieval and is also available through a Python interface [MacAvaney and Macdonald, 2022].

All of these frameworks primarily target academic retrieval experiments. Nevertheless, it is also possible to build upon industrial-grade retrieval software. Most notably, Apache Solr and Elasticsearch should be mentioned, both of which are built on the Java library Lucene. Notably, Anserini itself also builds upon Lucene but provides an interface that is specifically tailored to research use cases. While Anserini introduces additional features that facilitate more detailed analysis, it also serves as a common ground for transferring research outcomes into real-world applications [Grand et al., 2020].

Beyond lexical retrieval methods, several of the aforementioned frameworks also support end-to-end experiments involving deep learning-based retrieval approaches. For example, PyTerrier provides multiple submodules that integrate Transformer-based methods such as ColBERT, T5, and Doc2Query. Similarly, PyGaggle [Pradeep et al., 2023a], which is part of the Anserini/Pyserini ecosystem, supports several deep neural architectures for tasks such as passage ranking and question answering.

In addition, there exist software packages specifically designed for neural retrieval that predate the widespread adoption of Transformer architectures, e.g., MatchZoo [Guo et al., 2019] or Capreolus [Yates et al., 2020]. More recently, instruction-tuned LLMs have been explored for (re-)ranking text passages. RankGPT [Sun et al., 2023] is a framework built on OpenAI’s ChatGPT, whereas RankLLM [Pradeep et al., 2023b,c] adopts a more comprehensive approach by providing a unified interface to multiple (open-weight) LLMs.

Good Practices of Software Development

When implementing an experiment, it is beneficial to adhere to established software development best practices. These include, but are not limited to, the use of version control systems, dependency management, test-driven development with unit and regression tests, configuration files, and comprehensive code documentation. Additional practices such as the use of established programming patterns to improve readability, pair programming, code reviews, and systematic logging of project modifications further contribute to software quality.

Applying these principles from the outset of an IR experiment facilitates the tracing of provenance and increases the transparency of implementation choices and design decisions. More broadly, these best practices can be extended beyond the software itself to encompass the entire research project, including datasets, drafts, protocols, and the final publication [Ram, 2013].

When building upon existing software packages, careful attention must be paid to dependency management [Kusa et al., 2022, Crane, 2018]. Many soft-

ware packages are actively developed and frequently updated, which can have a substantial impact on reproducibility. In addition to reporting which software packages are used, it is therefore essential to specify their exact versions and to document the dependency configuration using appropriate tools, such as Maven for Java or `pip` for Python.

Likewise, when relying on existing implementations of retrieval methods provided by software frameworks, it is necessary to precisely specify which method variants are used in order to avoid ambiguities between different implementations of nominally identical approaches. This issue constitutes a critical aspect of experimental documentation, which we discuss in more detail in Section 4.4. In this context, we refer to the study by Kamphuis et al. [2020], who evaluated multiple variants of the lexical retrieval method BM25 within a standardized evaluation environment. Although these variants yielded comparable retrieval effectiveness, bitwise reproducibility could not be guaranteed.

Test-driven development techniques [Erdogmus et al., 2005] are beneficial for both productivity and reproducibility. Well-defined unit tests help clarify the intended functionality of an implementation and enable systematic assessment of the effects of code changes or updates at a functional level. Similarly, regression tests support the detection and tracking of changes in final ranking outputs resulting from software modifications.

The Anserini retrieval toolkit [Yang et al., 2018] provides a representative example of how regression tests over ranking outputs can be integrated into the development workflow. Through dedicated documentation and notebook-based regression tests, experimental outcomes can be independently validated and reviewed, enabling others to reuse baseline retrieval methods as intended.

Distributing the Software and Source Code

Once the research results are ready to be published, the corresponding code and the experimental implementations should be made publicly available — preferably in a persistent code repository with version control — and it should be associated with the publication, for instance, by including a footnote in the paper. There exist several code hosting services based on Git like GitHub⁶, GitLab⁷, or Bitbucket⁸. Alternatively, some publishers and outlets allow hosting additional material and resources.

Regardless of the hosting solution, it is essential to ensure that the referenced resources are associated with persistent identifiers. Even if a code repository continues to evolve after publication, the corresponding URL should remain stable to prevent link rot. If updates are anticipated, it is advisable to reference a specific commit or to create a software release that corresponds exactly to the version of the code used to produce the reported experiments and results.

Although many commercial code hosting platforms are considered “*too big to fail*”, market dynamics can still affect the long-term preservation of open source

⁶<https://github.com/>

⁷<https://gitlab.com/>

⁸<https://bitbucket.org/>

software [Escamilla et al., 2022, 2023]. For example, the Google Code Project Hosting Service⁹ was discontinued in 2016 and is no longer actively maintained. Although an archived version remains accessible online, this case illustrates that long-term code availability is fragile and cannot be assumed without additional precautions. In this context, the *Software Heritage* project [Cosmo and Zachiroli, 2017] represents a dedicated initiative to archive and preserve publicly available source code over the long term.¹⁰

In addition to the code used to run experiments, it is advisable to share the code responsible for generating evaluations, tables, figures, and other outputs derived from intermediate experimental results [Hofman et al., 2020]. Providing these scripts enhances transparency and ensures the integrity of the reported results. Some retrieval toolkits even support the direct export of experimental results as LaTeX tables [Macdonald et al., 2021, Bassani, 2022], facilitating seamless integration between experimental workflows and the final publication.

Since 2015, TREC organizers have encouraged participants to provide public Git repositories alongside their submissions, designating these as *open runs* [Voorhees et al., 2016]. It remains debatable whether making implementations publicly available at the outset of a shared task is advisable, particularly while other groups are still actively participating. Concerns have been raised that early disclosure could reduce the diversity of submissions, as participants might adjust their system design decisions based on knowledge of anticipated approaches from other teams. Nonetheless, making code publicly accessible after the conclusion of the shared task is generally considered beneficial. In practice, however, the inaugural year saw only limited success: a minority of participants made their runs *open* by providing accompanying code repositories.

Nonetheless, there is a noticeable positive trend toward sharing implementations alongside publications [Färber, 2020], and some studies suggest that releasing open source code can increase a paper’s impact, as measured by citations [Färber, 2020, Bhattarai et al., 2022]. However, the presence of a code repository should not be overestimated: it enables the rerunning of the original experiment at best, but does not provide insights beyond the scope of the initial study or guarantee generalizability [Collberg and Proebsting, 2016, Drummond, 2009]. Nevertheless, sharing source code remains a fundamental first step toward making research reusable for others.

Solutions and Actionable Recommendations

- **Open source code releases** facilitate sharing the implementation and is a first fundamental step towards reproducibility, which also has a positive impact on the reception of published research [Färber, 2020, Bhattarai et al., 2022],
- **Following good practices of software development**, including version control, dependency management, test-driven development, code doc-

⁹<https://code.google.com/archive/>

¹⁰<https://www.softwareheritage.org/>

umentation, and others [Ram, 2013, Erdogmus et al., 2005, Hofman et al., 2020],

- **Building upon existing retrieval toolkits** for common retrieval operations to ensure correctness and better comparability with other approaches [Yang et al., 2018, Ounis et al., 2005, Strohman et al., 2005, Mallia et al., 2019].

4.3 Sharing the Experimental Data

To illustrate the importance of sharing data artifacts, consider a conventional ad hoc search experiment following the Cranfield paradigm. The primary data inputs for such a system-oriented experiment include the document corpus and the topic files used to query the system. In addition, the corresponding relevance judgments constitute essential intangible data inputs: although they are not required to generate initial outputs, they are necessary for subsequent evaluation. Finally, the retrieved rankings—typically stored in run files—can be regarded as the experiment’s primary data outputs.

Having both the test collection and the original run file facilitates the validation of a system reimplementations. By comparing the reproduced run file with the original, researchers can assess the fidelity of the reimplemented system. In cases where the source code of the original system is unavailable, this comparison provides the only feasible means to recreate a retrieval system that can serve as a reliable baseline in a different context, such as with a new dataset. In this way, shared data artifacts function as essential reference points for attempts to reproduce a retrieval experiment.

In general, it is important to distinguish between data resources that exist prior to the experiment (*inputs*) and those that are generated as a result of executing the experiment (*outputs*). Ideally, both types of data should be shared to enable the experiment to be rerun with full integrity. Providing identical input and output data allows researchers to validate a reimplemented method and to compare the reproduced rankings, for example using the measures described in Section 3. This approach ensures that the method is correctly reused and can serve as a reliable baseline for subsequent experiments.

Using Open Data

Very often, the separation of data from the experimental setup — that is, from the method and its implementation — constitutes one of the primary barriers to reproducibility [Ivie and Thain, 2018]. If the data is unavailable and not shared alongside other components of the experimental setup, it becomes impossible to rerun and fully validate the experiments. This separation is, to some extent, inevitable, as it is generally not advisable to include large datasets directly within the same repository as the source code. In some cases, sharing data may not even be feasible due to constraints such as privacy concerns, business interests, or other restrictions.

TREC and other shared-task initiatives, such as CLEF and NTCIR, have contributed numerous data resources over the years [Voorhees et al., 2005, Ferro and Peters, 2019b, Tetsuya Sakai, 2021], providing a strong foundation for reproducible experiments. From a practical perspective, the availability of a common test collection and standardized data formats for computed rankings is crucial for conducting large-scale evaluations involving multiple systems and researchers, as well as for validating reproduced systems. In general, open data resources should be preferred over closed ones to enable end-to-end reproducibility. Furthermore, the use of standard or widely adopted datasets facilitates meaningful comparisons with existing research outcomes.

Versioning and Making Data Citable

Whenever a dataset receives significant updates, it is recommended to introduce versioning and make new releases explicit. Sometimes, different versions of a test collection exist. For instance, the TREC Washington Post¹¹ corpus has different versions that emerged from the deduplication of the document set. Another example of a versioned IR test collection is TREC-COVID [Voorhees et al., 2020], which was compiled over multiple rounds when the underlying document set changed and relevance labels were updated. Similarly, some tracks of TREC ran for multiple years, the document collection stayed fixed, and new relevant labels were made for each new iteration. Thus, the data resources should be identified precisely.

Rauber et al. [2016b] introduce data citation principles for the creation of subsets from scholarly data. Their approach emphasizes sharing a re-executable workflow for generating the subset, rather than simply distributing a copy of the subset itself. This method is particularly valuable when multiple subsets are required for an experiment and sharing large volumes of data is impractical. Technically, Rauber et al. [2016b] propose assigning persistent identifiers to the database queries used to generate the subsets. Their methodology incorporates several recommendations designed to ensure complete data provenance and to enhance the reproducibility of subset creation.

Training, validation, and test data splits are critical methodological considerations in how input datasets are used for experimental evaluation. Proper reporting and documentation of these splits are essential for ensuring reproducible outcomes. For example, Rao et al. [2015] and Berrendorf et al. [2021] observed substantial differences between the effectiveness of original and reproduced systems, attributable in part to variations in data splitting procedures. When combined with the subset creation methodology proposed by Rauber et al. [2016b], this approach provides a practical means to make data splits explicit without the need to share duplicate portions of the original raw data.

¹¹<https://trec.nist.gov/data/wapost/>

Sharing Generated and Intermediate Experimental Data

Recent advances in NLP and IR have demonstrated substantial performance improvements when LLMs are incorporated into experimental methodologies. When a language model is used at any stage of an experiment, it is advisable to share the corresponding model weights. Many pre-trained and fine-tuned LLMs are openly available and reusable, eliminating the need for the substantial hardware resources that would otherwise be required to retrain the models and replicate the experimental setup [Lin et al., 2021b, Ma et al., 2022, Wang et al., 2022b]. Notably, the Hugging Face platform¹² serves as a widely used hub for accessing both models and associated datasets for these purposes.

Another important intermediate data artifact in a conventional retrieval pipeline is the inverted index. Using a common index, generated with a consistent pipeline for lexical unit processing, facilitates the reuse of identical index data when rerunning experiments. Although many established indexing methods exist and are not typically the focus of recent research, they can nevertheless have a substantial impact on experimental outcomes, as demonstrated in several studies [Ferro and Silvello, 2016b, Roy et al., 2018]. More recently, Lin et al. [2020] proposed sharing indices using a standardized *Common Index File Format (CIFF)* format. Enabling multiple retrieval toolkits to operate on the same index promotes fairer evaluation of the toolkits themselves and the corresponding implementations of retrieval methods, independent of variations introduced by the indexing process.

When data is unavailable or insufficiently shared, reproducibility can be significantly compromised. Several strategies exist to promote better integration between data and other experimental resources, among which the FAIR principles are particularly notable. According to Wilkinson et al. [2016], scholarly and experimental data artifacts should adhere to the principles of Findability, Accessibility, Interoperability, and Reusability to ensure good data management and to enhance transparency, reproducibility, and reusability. A practical first step toward meeting these principles is the use of dedicated data-sharing platforms, such as OSF¹³, Zenodo¹⁴, or Figshare¹⁵, among others.

User Data and IIR Experiments

In some cases, data cannot be shared due to sensitivity or legal and ethical constraints [Callan and Moffat, 2012]. For example, user interaction logs are often restricted to protect privacy or safeguard business interests. When the original dataset cannot be disclosed, a viable alternative is the use of simulated IR test collections. Hawking et al. [2020] describe various methods for generating synthetic documents, queries, and interaction logs. If experimental results are reported based on private data, researchers can additionally provide results

¹²<https://huggingface.co/>

¹³<https://osf.io/>

¹⁴<https://zenodo.org/>

¹⁵<https://figshare.com/>

derived from the synthetic datasets, enabling others to verify and reproduce findings using shareable data artifacts.

Evaluating a proposed method on multiple datasets is generally recommended and considered best practice, facilitating better assessment of the generalizability of a given methodology. This is particularly important when experiments rely on paywalled or proprietary data. Supplementing such evaluations with open-data alternatives allows others to replicate the experiments without requiring licenses or incurring additional costs. For example, Ferro et al. [2019c] report anticipated workshop submissions that could not proceed due to licensing and cost restrictions on a critical dataset. Complementing these experiments with open data resources helps democratize research and enhances reproducibility.

Beyond system-oriented evaluations, user interaction logs provide another means of assessing the quality of retrieval systems based on actual user behavior. Several datasets share anonymized logs, including query formulations and clicks [Craswell et al., 2020, Rekabsaz et al., 2021]. The widely known AOL search log dataset [Pass et al., 2006], though not anonymized, has been frequently reused and serves as a valuable resource for personalization research [MacAvaney et al., 2022b].

Nevertheless, anonymized datasets are generally preferred, as they can be shared without raising privacy concerns. Protocols from IIR experiments go beyond logged interactions with the search interface, often including additional materials such as questionnaire results or think-aloud protocols. For guidance on sharing resources from IIR studies, we refer the reader to Gäde et al. [2021], who outline a comprehensive manifesto for resource sharing in this context.

Solutions and Actionable Recommendations

- **Open data** like the IR test collections that stem from shared task efforts [Voorhees et al., 2005, Ferro and Peters, 2019b, Kando, 1999] as an alternative to paywalled, private and closed datasets,
- **Data management software and data citation principles** dedicated to IR test collections, e.g. [Piwowarski, 2020, MacAvaney et al., 2021], and for better data provenance of dynamically changing data [Rauber et al., 2016b], can also make train, validation, test data splits more explicit [Rao et al., 2015],
- **Sharing experimental data artifacts** and make them findable and citable with the help dedicated sharing platforms [Wilkinson et al., 2016]; in this context, intermediate data outputs like the inverted index [Lin et al., 2020] or model weights [Ma et al., 2022, Wang et al., 2022b] could also be made available,
- **Simulated test collections and anonymized user logs** are alternatives to the closed variants when sharing the original data is not possible [Hawking et al., 2020, Gäde et al., 2021].

4.4 Documenting the Experimental Setup

Documenting the data and software resources is essential for communicating the experimental setup with others, supporting future (including personal) reuse, and the transparency and provenance of the final outcomes in a publication. When certain experimental components cannot be shared directly, they should at least be thoroughly documented and provided as metadata. Moreover, comprehensive documentation has the potential to reduce the environmental impact and costs of research — one of the most pressing issues in IR [Scells et al., 2022] and NLP [Strubell et al., 2019]. By making experimental resources transparent and findable, documentation directly promotes their reuse and broader accessibility.

Metadata Protocols and Annotations

Metadata and annotations are essential for enhancing the transparency and reproducibility of experiments and their resulting data artifacts [Leipzig et al., 2021]. In recent years, several metadata schemes, protocols, and documentation guidelines have been proposed for ML experiments. A non-exhaustive selection includes data-focused protocols such as *Datasheets for Datasets* [Gebru et al., 2021] and *Data Statements* [Bender and Friedman, 2018], as well as model- and system-focused protocols such as *Model Cards* [Mitchell et al., 2019] and FactSheets [Arnold et al., 2019]. The Hugging Face platform exemplifies the integration of both data and model cards into a unified catalog. Additionally, *Conference on Neural Information Processing Systems (NeurIPS)* has introduced *reproducibility checklists* [Pineau et al., 2021], which are submitted alongside research contributions to enhance transparency and facilitate reproducibility in conference publications.

Many components of a typical IR system pipeline are well established and have been analyzed in previous work. These components are often overlooked in contemporary experiments, yet they are essential for exact reproducibility. For example, Roy et al. [2018] illustrate the effect of (removing) markup artifacts on reproducibility, while Ferro and Silvello [2016b] show that lexical unit generators—such as stemmers, stopword removal, and similar preprocessing steps—can influence final retrieval outcomes. Because these details are largely technical, their inclusion in publications is sometimes debated, particularly when considering page limits or the cognitive load of readers. Although they may not directly contribute to conveying the central research contribution, such details remain critical for reproducibility and can be reported as supplementary metadata for the experiment.

Focusing specifically on IR experiments, Silvello et al. [2017] proposes a resource description framework that uses semantic representations to expose data artifacts as machine-readable linked open data on the web. Similarly, Lipani et al. [2014] introduce an ontology to support machine-readable nanopublications, systematically describing the components of a search system and the overall IR experiment. Reporting such details through supplementary docu-

mentation protocols makes them explicit and findable without requiring their inclusion in the main publication.

Building on these efforts, `ir_metadata` [Breuer et al., 2022a] provides a lightweight metadata scheme for annotating TREC-style run files in IR experiments. Its underlying structure is guided by the PRIMAD framework [Ferro et al., 2016b, Freire et al., 2016], encompassing the six core components, which are further detailed through additional subcomponents in the scheme. To strengthen the linkage between the ranking outputs and their contextual information, the scheme is designed to annotate run files directly with metadata comments.

Workflow Configuration and Management

Although machine learning pipelines generally follow the common paradigm of training, validation, and testing, they are often highly complex, making it infeasible to document all details within a publication. In particular, hyperparameters can have a substantial impact on the reproducibility of experimental outcomes [Crane, 2018]. While these technical details may be of limited importance for the general readership, it is nonetheless advisable to make them explicit to support reproducibility and accurate replication of the experiments.

Configuration frameworks can assist practitioners in documenting and communicating experimental details at a higher level, without embedding them directly in the source code [Yadan, 2019, Anelli et al., 2021]. For instance, Hydra [Yadan, 2019] is a software toolkit specifically designed for managing configurations in complex experimental workflows. Using such a toolkit, modifications to hyperparameters can be performed more conveniently, facilitating robustness tests and potentially improving experimental outcomes. Additionally, several *Machine Learning Operations* (MLOps) platforms support the deployment, tracking, and visualization of standardized pipelines. Examples of these platforms include MLflow¹⁶, Weights and Biases¹⁷, and Neptune.ai¹⁸.

In addition, there are workflow management utilities designed to describe computational experiments in a principled and reproducible manner. For example, Snakemake [Mölder et al., 2021] focuses on data analysis workflows and employs a domain-specific language. It enables the definition of an experiment and its corresponding analyses in a declarative fashion, by decomposing the workflow into individual processing steps with associated input and output data, scripts, command-line calls, and other dependencies. Similarly, *Common Workflow Language (CWL)*¹⁹ [Crusoe et al., 2022] is a standardized and formalized language for describing computational experiment workflows, supporting reproducibility and interoperability across different systems.

¹⁶<https://mlflow.org/>

¹⁷<https://wandb.ai/site>

¹⁸<https://neptune.ai/>

¹⁹<https://www.commonwl.org/>

Logging System and Data Provenance

Another approach to “*automatically document*” a computational experiment and enhance transparency is through logging data provenance. Provenance can be captured at multiple levels [Freire et al., 2008], including the programming language or application level, the workflow level, and the operating system level. Various tools exist to record data provenance and system activity. For example, the Linux utility `strace`²⁰ logs interactions between processes and the kernel. More specialized tools for computational experiments include ReproZip [Chirigati et al., 2016], which records system calls and thereby implicitly documents the experiment in an automated fashion. ReproZip captures all activities associated with the experiment without requiring active use of workflow or configuration management tools. To ensure hardware independence, it can package the experiment within Docker containers or virtual machines, enabling others to rerun the original setup and serving as an alternative to the platform-sharing solutions discussed in Section 4.1.

Executable Papers

Kamphuis et al. [2020] highlight that multiple variants of a given retrieval method can exist, as illustrated by different implementations of BM25. Their study evaluates these variants and implementations, showing that although the differences in average effectiveness are generally small, the exact scores are not identical. Depending on the experimental context, such differences can be critical. For instance, in repeated user studies, reproducing an experiment with a slightly different BM25 variant may yield rankings that are similarly effective overall, but the specific order of documents may differ from the original. This discrepancy matters because users can be sensitive to changes in ranking, even when overall effectiveness metrics appear equivalent.

To minimize ambiguities between the methods described in a publication and their actual implementation *in silico*, executable papers and code notebooks provide a means to tightly integrate the computational method, its implementation, and the resulting outputs [Kluyver et al., 2016, Jimenez et al., 2017a, Bar and Wang, 2021]. In addition to containing the code necessary to run the experiments, these formats allow researchers to document the methodology and present the results within the same executable environment, eliminating the need to reference separate documents or publications.

Focusing specifically on IR experiments, Ferro et al. [2011] describe how the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* framework enables researchers to make their papers “executable”. In particular, DIRECT assigns each experiment a unique DOI, making it directly citable and referenceable from the publication. Ferro et al. [2011] envision reading a paper as an interactive experience, where all experimental assets—such as datasets, statistical analyses, plots, and other relevant materials—are provided alongside the publication. A browser plugin reads the paper and automatically retrieves

²⁰<https://strace.io/>

the corresponding experiments and associated data, thereby tightly linking the narrative to its computational evidence.

Experiment and Evaluation Platforms

Several experimental evaluation and data management platforms have been specifically developed for IR experiments, and a comparison of these platforms is provided in Table 4.3. These platforms support the evaluation of system-oriented IR research and the management of associated data artifacts and resources, with some focusing on specific datasets such as CLEF or TREC, while others target more general use cases. Generally, all of these services are web-based applications that require ongoing maintenance and administration, which is critical for their long-term availability. Notably, some of the older platforms are no longer accessible online.

DIRECT [Nunzio and Ferro, 2005, Agosti et al., 2007a, ?] was a resource management platform developed for shared tasks at CLEF, designed to host experimental data artifacts and enrich them with metadata according to a conceptual model. The framework encompassed eight areas: (1) evaluation activity, (2) experiment, (3) resource management, (4) experimental collection, (5) measurement, (6) metadata, (7) visual analytics, and (8) bibliographical data. Additionally, DIRECT provided a RESTful web service that enabled structured querying of resources, facilitating systematic access to both the data artifacts and their associated metadata.

With a special focus on run files of TREC-style experiment, EvaluatIR [Armstrong et al., 2009a] was an online platform that allowed the uploading and evaluating IR experiments. The resulting outcomes were made referenceable by a unique URL. Hereby, the platform operators aimed to address the need to select stronger baselines [Armstrong et al., 2009b] and make them easier to find. A similar tool for (a sample of) CLEF data was SOIRE [Dittenbach et al., 2009].

RISE [Yang and Fang, 2016] was a Docker-based retrieval and evaluation platform that allowed the execution of standardized experiments with TREC test collections. The authors implemented several retrieval methods, but the software design also allowed others to integrate their experiments into the platform. Following a more holistic approach, EvALL [Amigó et al.] demonstrates the implementation of an evaluation platform for shared tasks of information access systems. By design, it allowed the evaluation of computational results, comparison to other submissions, provision of a reusable data format, and automatic generation of figures.

Other metadata management platforms that do not focus or go beyond the execution of experiments and evaluations are CLAIRE [Angelini et al., 2018] and the IR Anthology [Potthast et al., 2021]. CLAIRE is a visual analytics platform for the principled and systematic evaluation of retrieval systems and their sub-components. The platform visualizes the impact of each system component with the help of grid-based heatmaps. The implementation is demonstrated with a sample of TREC test collections. Inspired by the ACL Anthology [Gildea et al., 2018], the IR Anthology combines a comprehensive corpus of IR publications

Table 4.3: Comparison of experimental evaluation and data management platforms for system-oriented IR experiments.

Platform	Type	Scope/Coverage	Year
DIRECT [Agosti et al., 2007b]	(Meta)data management	CLEF	2005
EvaluatIR [Armstrong et al., 2009a]	Evaluation	TREC (sample)	2009
RISE [Yang and Fang, 2016]	Experiment/evaluation	TREC (sample)	2016
SOIRE [Dittenbach et al., 2009]	Evaluation	CLEF (sample)	2017
EvALL [Amigó et al.]	Evaluation	General purpose	2017
CLAIRE [Angelini et al., 2018]	Visual analytics/evaluation	TREC (sample)	2018
IR Anthology [Potthast et al., 2021]	(Metadata) corpus of IR literature	IR literature	2021
ranx_hub [Bassani, 2023]	(Meta)data management	General purpose	2023
TIREx [Fröbe et al., 2023]	Experiment/evaluation	General purpose	2023
TREC Browser [Breuer et al., 2024b]	(Meta)data management	TREC	2024
KIMERA [Pasin and Ferro, 2025]	Experiment/evaluation	General purpose	2025

with a search interface and the underlying ChatNoir search engine [Bevendorff et al., 2018] that enables full-text search and access to the corresponding meta-data, e.g., BibTeX references.

ranx_hub [Bassani, 2023] is an online repository giving access to pre-computed rankings for standard test collections. The “hub” can be accessed via the interface that is integrated into the retrieval evaluation toolkit **ranx** [Bassani, 2022]. Similarly, *The Information Retrieval Experiment Platform (TIREx)* [Fröbe et al., 2023] bundles multiple of the earlier mentioned toolkits and resources to facilitate reproducible and systematic IR evaluations across a variety of different shared tasks, datasets, and retrieval methods. More specifically, it builds upon the EaaS [Hopfgartner et al., 2018] platform TIRA [Fröbe et al., 2023] and integrates PyTerrier [Macdonald et al., 2021], **ir_datasets** [MacAvaney et al., 2021], and **ir_measures** [MacAvaney et al., 2022a]. TIREx embeds the concepts of containerization and version control, allowing its users to define re-executable and reproducible experiments and evaluations.

The TREC Browser [Breuer et al., 2024b] and the related software suite compose a lightweight metadata management platform. As can be seen from Table 4.3, it is not made for running and evaluating IR experiments. Instead, it facilitates access to metadata of experimental resources, resulting from over 30 years of IR experiments and evaluations at TREC. More specifically, it comprises the TREC metadata, including pointers to the experimental results, i.e., the run submissions, the evaluation protocols, the proceedings, the datasets, and others. By building upon a metadatabase, a web interface allows browsing former TREC submissions²¹ and a corresponding RESTful interface makes more principled and structured queries about the TREC metadata possible.

KIMERA [Pasin and Ferro, 2025] is a cloud-based platform for evaluating information retrieval systems that lets researchers develop and run experiments directly in the browser. Built on Docker and Kubernetes, it offers a scalable, secure, and reproducible environment, monitors resource usage, and supports access to advanced resources like quantum computers. It has been used in QuantumCLEF and in quantum computing tutorials at ECIR and SIGIR.

Crediting Contributors

Considering the *actor* component of the PRIMAD model, there is a less technical but nonetheless crucial aspect of reproducibility: the human element. A reproducibility analysis by Raff [2019] highlights that, despite numerous technologically advanced solutions, direct communication — such as email correspondence — remains a key factor in achieving successful reproduction. Unique researcher identifiers, such as ORCID, help prevent ambiguities between authors with the same name [Haak et al., 2012] and facilitate contact. Furthermore, authors can clarify their individual contributions to both the publication and the experiments by using structured reporting frameworks, such as the CRediT taxonomy [Allen et al., 2019].

²¹<https://pages.nist.gov/trec-browser/>

Solutions and Actionable Recommendations

- **Resource management** with dedicated experiment and evaluation platforms [Agosti et al., 2007b, Armstrong et al., 2009a, Potthast et al., 2019],
- **Documentation by metadata, reproducibility checklists, and other resource-oriented protocols** [Lipani et al., 2014]
- **Workflow management and logging provenance** for better life cycle management and data provenance of experimental resources and results [Mölder et al., 2021, Crusoe et al., 2022, Freire et al., 2008],
- **Executable papers and code notebooks** can reduce disambiguities and bring the publication and experiment together in a more cohesive way [Brammer et al., 2011, Gorp and Mazanek, 2011, Kluyver et al., 2016, Pimentel et al., 2019],
- **Making contributions explicit**, e.g., with the help of an ORCID [Haak et al., 2012] or assigning contributor roles with the help of the CRediT taxonomy [Allen et al., 2019].

4.5 Advancing IR Research Towards an Open Science

As previously noted, the FAIR principles put the four concepts of findability, accessibility, interoperability, and reusability into context with regard to how experimental data and resources are shared [Wilkinson et al., 2016]. In this section, we have reviewed existing approaches for sharing computational platforms, software, and data, as well as best practices for documenting experiments to enhance transparency, reproducibility, and reusability.

Overall, we can conclude that a range of effective solutions exists to support the advancement of IR research toward a more open and transparent scientific practice. In this regard, many research efforts already align, at least implicitly, with FAIR principles. Nevertheless, several open challenges remain that must be addressed to enable IR research to fully and consistently adhere to these principles.

Several test collections and datasets provide a common foundation for standardized IR evaluations. Many of these resources originate from shared task initiatives organized by conferences such as TREC, CLEF, and NTCIR [Voorhees et al., 2005, Ferro and Peters, 2019b, Tetsuya Sakai, 2021]. Tools such as `ir_datasets` [MacAvaney et al., 2021] and `DATAMAESTRO` [Piwowarski, 2020] provide access to these collections via standardized interfaces and internal naming schemes. Nevertheless, many test collections and datasets still lack standardized names and unique persistent identifiers, which would facilitate more reliable discovery, referencing, and reuse.

Similarly, the reusability of a test collection is often not made fully explicit. Although many collections are openly available for research, and their provenance is generally transparent, detailed information about dataset properties is frequently only accessible through overview reports of the corresponding shared tasks. Such details could be more explicitly and systematically documented using structured protocols, for example, following approaches like Datasheets for Datasets [Geburu et al., 2021].

Furthermore, we noted that there are many different open source toolkits that help implement an experiment based on well-maintained software developed in collaboration with the IR community. While many of them have a high degree of compatibility, there are also limitations with regard to their interoperability. For example, it is not trivial to generate an inverted index with one retrieval toolkit and then use the index with another toolkit. Efforts like the CIFF [Lin et al., 2020] demonstrate how a commonly defined index format can facilitate interoperability between a shared index and several retrieval toolkits. In the future, other components could be standardized as well to have more loosely coupled implementations of pipeline components to improve interoperability.

State-of-the-art IR approaches increasingly rely on LLMs and foundation models. Even though many of these models are openly available to different extent [Liesenfeld and Dingemans, 2024], the IR community still lacks a common model that is developed in collaboration. The current landscape is rather fragmented, with many different closed and (semi-)open models being used for similar tasks, limiting the comparability of experimental results. Future efforts could therefore aim to develop principles and guidelines for collaboratively advancing the development of open and fully-reproducible foundation models for IR research.

Ensuring compatibility and interoperability among different tools and components requires standardized formats and common conventions. Currently, certain de facto standards exist for sharing the outcomes of IR experiments, which facilitate interoperability across various operations and processes. For example, the file formats for runs, topics, and relevance judgments (qrels) have been widely adopted in numerous shared tasks and experiments since their introduction in TREC. Although having a common file format may seem mundane, it represents a fundamental requirement for enabling interoperability between retrieval systems, evaluation tools, and other components of a research project.

Grid@CLEF [Ferro and Harman, 2010] was a pioneering initiative that already demonstrated the benefits of having interoperable and loosely coupled components of a retrieval system. The basic idea of these efforts is grounded on a principled and exhaustive evaluation of all possible combinations that would emerge from different experimental constellations. Suppose that we are interested in identifying the most effective system configuration in a multi-lingual evaluation scenario. In that case, we can systematically evaluate all different kinds of combinations, e.g., which would stem from lexical unit generators and retrieval methods, across an array of different languages. Technically, these efforts were made possible with the help of the *Coordinated Information Retrieval Components Orchestration (CIRCO)* framework [Ferro, 2009], which allows dis-

tributed, loosely coupled, and asynchronous experimentation. More specifically, the in- and outputs of each system component are standardized by a commonly-defined XML schema. Besides the clear benefits for reproducibility like transparency or improved data provenance, some participants were also able to debug and improve their retrieval system [Larson, 2010].

Other domains have adopted some proactive reproducibility practices that — to the best of our knowledge — have not found widespread application in IR research yet. For example, the workflow of a typical IR pipeline could be described in a detailed way with the help of workflow configuration and management toolkits [Mölder et al., 2021, Crusoe et al., 2022]. Thinking further, not only the actual retrieval method could be documented but also the evaluations of the ranking outputs and the code to generate the figures and plots can be documented and made more transparent for the sake of better integrity between the code and paper.

Besides the need for better and IR-specific workflow management tools, there is also a need for solutions to log system calls, the data provenance, and to trace the experimental execution in general. While in other domains, several provenance logging toolkits have been proposed, there is currently no platform-independent solution for IR experiment. The TIREx Tracker [Hagen et al., 2025] is a first step in this direction, which automatically logs system calls and data provenance in accordance with the `ir_metadata` scheme. However, in its current form, it is tightly coupled to the TIREx platform and its underlying architecture. In the future, more general logging solutions could be developed that are independent of a specific platform or experimental environment.

In this regard, executable papers also offer a more immediate way to disseminate the results of a research project and the underlying implementations [Jimenez et al., 2017b,a, Xu et al., 2022]. EaaS platforms like TIRA/TIREx that make it possible to rerun experiments in their entirety on a cloud-based infrastructure. This kind of standardized platform offers an excellent opportunity to enforce the implementation of executable papers, where the experiments can be directly provided alongside the publication in a way that allows validation by re-execution and also fosters the reusability of the experimental resources.

In the future, these endeavors can be enforced by organizing conference and journal submissions that support novel forms of publishing research. For instance, pioneering examples like the ACM Digital Library²² and the arXiv²³ now support rendering HTML-based versions of a publication in a browser. In combination, these technologies pave the way towards wider accessibility. By including pointers to the supporting experiments, there is a more cohesive and interactive solution of bringing the conclusions and the empirical evidence of a computational experiment together.

Our review of the experiment and evaluation platforms also showed that there is a variety of other solutions for documentation and preserving experimental results. Some of them, like the EaaS platforms, make it possible to

²²<https://dl.acm.org/>

²³<https://arxiv.org/>

preserve an IR experiment entirely and also allow rerunning it in a reproducible way. Others align the experimental resources and artifacts to an ontology to improve findability. While these technologies address many of the reproducibility issues that were pointed out earlier, they also come at the cost of administration efforts. We note that many of the older management platforms are not online anymore. Guaranteeing accessibility in the long run is one of the major challenges the community might have to face as a whole. Generally, researchers and practitioners should be encouraged to document their experiments as detailed as possible.

Of course, ensuring reproducibility is not only an entirely technical question but also revolves around sociological and cultural questions. With special regard to the new generations of researchers and practitioners, reproducibility should be considered as a fundamental part of the curriculum. Many scientific fields and also the IR community have already begun to establish tutorials and lectures about best practices for reproducibility [Hofman et al., 2020, Yildiz et al., 2021, Ribotta et al., 2022, Lucic et al., 2022b]. Hereby, students are sensitized for reproducibility issues and learn how to mitigate these. By understanding reproducibility as an ongoing effort that requires continuous revalidation and audits helps to ensure trustworthiness in the long run.

Chapter 5

Reproducibility in the Broader Community

Besides the IR community, other neighboring research communities have also developed a culture of reproducibility. In the following, we provide a brief overview of initiatives and the current state of the art in other communities like *Database (DB)* systems, *Recommender Systems (RS)*, and *Natural Language Processing (NLP)*. We end this section with a listing of aspects that can be learned from other communities and could be enforced in IR in the future.

5.1 Reproducibility in the Database Community

As one of the very first communities in the computational sciences, the DB community started in 2008 to evaluate the reproducibility of submitted works as part of the SIGMOD conference and has since developed several good practices and efforts to make data-driven and computational research more reproducible [Manolescu et al., 2008]. For instance, this community is a prime example of having successfully implemented the badging procedure of accepted papers regarding the availability of artifacts and reproducibility.¹ Similar to the SIGIR Artifact Evaluation Committee, the database conferences have established dedicated reproducibility committees that explicitly review a paper’s reproducibility and the availability of data and code artifacts after acceptance. Even more, the VLDB conference has recently moved towards making the availability of supplementary material and data during the reviewing phase mandatory, unless there are reasons why it is impossible.²

In 2024, SIGMOD inaugurated the Availability & Reproducibility Initiative (ARI) as a continuation of the earlier efforts. The initiative has its own proceedings — the SIGMOD ARI Reports — with citable reproducibility protocols [10.,

¹<https://reproducibility.sigmod.org/>

²<http://vldb.org/pvldb/volumes/17/submission/>

Table 5.1: Overview of open source toolkits for recommender systems.

Toolkit	Use Case(s)	Software and Publication
ClayRS	Various	 [Lops et al., 2023]
Cornac	Multimodal	 [Salah et al., 2020]
DaisyRec	Top-N	 [Sun et al., 2020]
DataRec	Data Management	 [Mancino et al., 2025]
Elliot	Various	 [Malitesta et al., 2023]
FuxiCTR	CTR Prediction	 [Anelli et al., 2021]
iREC	Interactive RS	 [Silva et al., 2022]
Jurity	Fairness Evaluation	 [Kadioglu and Thielbar, 2024]
LensKit	Various	 [Ekstrand, 2020]
Mab2rec	Multi-Armed Bandits	 [Kadioglu and Kleynhans, 2024]
RecBole	Various	 [Zhao et al., 2022]
ReChorus	Various	 [Li et al., 2024]
RecList	Black Box Testing	 [Chia et al., 2022]
Recommenders	Various	 [Argyriou et al., 2020]
RecPack	Top-N	 [Michiels et al., 2022]

2025]. Authors have a reasonable level of interest in having a paper’s artifacts reviewed for availability and reproducibility. The organizers report 72 submitted papers for the artifacts review out of 213 submitted papers to SIGMOD 2024 in total. Still, there is room for improvement. Only a fraction, 24 out of the 72 submissions, were assigned a “Results Reproduced” badge, which is backed by a corresponding reproducibility report in the proceedings. We refer the reader to the round table report by Athanassoulis et al. [2022] who provide a more detailed summary of current reproducibility actions and ambitions in the database community.

5.2 Reproducibility in the Recommender Systems Community

Adjacent to the IR community, the RS community has very similar research challenges, in the light of reproducible research. A fairly recent systematic reproducibility review by Ferrari Dacrema et al. [2021] revealed that reproducibility issues also affect RS research. In this particular study, only a minority of earlier works was reproducible (12 out of 26 papers), and most of these reproduced neural methods were still outperformed by less complex but properly-tuned baseline models (11 out of 12 papers). Similar observations were made by Ludewig et al. [2021], Sun et al. [2020], or Ferrari Dacrema et al. [2026] in the context of additional systematic reproducibility reviews with a focus on published RS works at top-tier conferences and venues.

Comparable to the challenges and barriers we mentioned in Section 3 with a focus on IR, there are very similar reasons why RS research is sometimes

not reproducible. Among others, reasons and common factors that can lead to irreproducible outcomes also include cultural barriers, methodological and experimental flaws or the lack of shared resources (cf. Fig. 3.2).

For instance, several offline RS experiments have erroneous evaluation protocols, as pointed out by Hidasi and Czapp [2023b]. Focusing on sequential recommendations, they discuss four prominent experimental flaws, including 1) dataset-task mismatch, 2) supposedly generalizable conclusions based on intensely preprocessed data, 3) information leakage over time, and 4) negative item sampling. While these factors are a barrier to the experiment’s validity, they also threaten reproducibility in the end.

Another study by Hidasi and Czapp [2023a] emphasizes the importance of correct and rigorous reimplementations. Based on a selection of six reimplementations of a particular neural recommendation approach, they examine the scrutiny of the reproducers when reimplementing the original methodology. They reveal several flaws, leading to severe performance drops in accuracy and, thus, deteriorated baselines. The highlighted flaws include the implementation of a different algorithm, missing important features of the methodology, software bugs and hard-coded parameters, lack of negative sampling during training, and scalability issues. Furthermore, they also point out that some RS frameworks contain unsatisfactory reimplementations. In conclusion, they urge the research community to enforce stronger double-checks of reimplemented methods, especially if an open source version by the original authors is already available.

Focusing on the hyperparameter optimization of baseline methods, Shehzad and Jannach [2023] re-analyze algorithm comparisons with properly tuned recommendation methods. The corresponding experiment suggests that each newly proposed algorithm can be presented as the most effective one in an unsound experimental setup that neglects the tuning of baseline methods. In conclusion, they urge the community to put more effort into documenting the hyperparameter tuning of baseline methods and outline several recommendations for a more rigorous evaluation protocol.

Ferrari Dacrema et al. [2026] identify three major reproducibility issues in their systematic reproducibility review of recent graph-based RS approaches: 1) erroneous data splits, causing information leakage between training and testing data, 2) inconsistencies between the provided code and data artifacts and their descriptions in the paper, causing uncertainty about what is actually being evaluated, and 3) the use of weak but fairly complex baselines, causing an illusionary progress of the state of the art.

We refer the reader to several other works and reproducibility reviews that point out reasons for irreproducible RS experiments aligning with those issues already pointed out for IR, with several similar solutions and countermeasures that were put into practice [Ferrari Dacrema et al., 2019, 2020, Said and Bellogín, 2014, 2015, Beel et al., 2016, Snoek et al., 2012, Petrov and Macdonald, 2022].

Similar to ECIR and SIGIR, the RecSys conference inaugurated a dedicated reproducibility paper track in 2020, which has been continued since then [Santos et al., 2020] to offer a venue for reproducibility issues and more rigorous reval-

idations. Notably, the community can rely on several open source benchmark toolkits that help researchers to follow a more thorough and standardized evaluation environment for RS experiments. A non-exhaustive list of fairly recent and maintained frameworks is given in Table 5.1. In general, there exist several general-purpose RS toolkits that cover various use cases like general, sequential, context-aware, and knowledge-based recommendation approaches such as ClayRS, Elliot, LensKit, Recommenders, RecBole, or ReChorus. While others have a more specialized focus such as Cornac for multimodal recommendations, FuxiCTR for click-through rate prediction, Mab2Rec for multi-armed bandit-based recommendations, iREC for interactive recommendations, or DaisyRec and RecPack for Top-N recommendation tasks. Jurity and RecList have a focus on fairness evaluation and black-box testing, respectively. DataRec is a library that standardizes and simplifies data handling for recommender system research, supporting reproducible dataset preparation, data versioning, and integration with other frameworks. Practitioners may find the compiled list of resources helpful and can use it as a starting point to design and implement an experimental setup in a common evaluation environment.

5.3 Reproducibility in Natural Language Processing

Notably, the REPROLANG [Branco et al., 2020] shared task at LREC in 2020 followed a similar approach as the CENTRE labs for IR that we mentioned in Section 3. The REPROLANG organizers provided a selection of several reproduction candidates in the call for contributions, that were supposed to be reimplemented by the task’s participants based on the information provided by the original authors. In addition to a reproducibility report, the organizers also asked for the submission of the reproduced software in a Docker container, which was later on revalidated by the organizers as an attempt to “*reproduce reproductions*”.

Belz et al. [2021a] provide an overview based on a systematic literature review about current and ongoing efforts towards reproducibility analyses in the NLP domain, including 35 reproducibility papers from 2010 to 2020 contained in the ACL Anthology. As already noted earlier for IR, there is also an inconsistent use of the terminology in NLP. In addition, Belz et al. analyze the distribution of score differences between the original and the reproduced scores. Their analysis reveals a substantial amount of deviating scores even when the same experimental setups are used. Belz et al. also highlight the moderate amount of systematic benchmark evaluations that make use of the same software framework for experimental evaluations.

Another series of workshops concerned with reproducibility in NLP started with the ReproGEN shared task in 2021 [Belz et al., 2021b]. Following a similar task design as REPROLANG, the ReproGEN workshop challenges participants to reproduce a selection of candidate papers, or invite them to reproduce their

work of choice with a focus on earlier work about Natural Language Generation. The shared task had one successive iteration in 2022 [Belz et al., 2022] until it was renamed to ReproNLP in 2023 [Belz and Thomson, 2023, 2024], expanding the scope to the general NLP domain. Over the years, the shared tasks were well-received, and had a good amount of participants, which contributes to insightful task overview reports and future challenges. For instance, Belz et al. emphasize the current lack of common evaluation protocols in NLP.

Dodge et al. [2019] introduced an adapted version for NLP of the The Machine Learning Reproducibility Checklist³ by Pineau for the sake of an improved reporting of the experimental setup. This checklist has not only been included in several call for papers of EMNLP, NAACL, and ACL, but also the feedback by the authors was collected in questionnaires. Authors were asked to confirm the availability of different resources such as data and source code but also important meta-information about the experiment like descriptions of the hardware infrastructure, evaluation protocols, hyperparameter configurations, and others. Magnusson et al. [2023] provide an overview of this data that stresses the positive impact of providing such a checklist to the authors before submission. According to them, there is an observable trend for papers having a higher probability of being accepted, as more items in the checklist can be confirmed. The checklist’s questionnaire is mainly based on closed questions, i.e., providing these additional pieces of metadata does not impose a much higher workload on authors when submitting a paper. Likewise, the checklist provides guidance to the reviewers, making its integration promising for future peer-reviews of IR conferences. We refer the interested reader to the overview article by Magnusson et al. [2023] for the full list of criteria.

5.4 Community-Bridging Efforts

Initially, the NeurIPS conference inaugurated the ML reproducibility challenge in 2018 and also promoted Pineau’s *Machine Learning Reproducibility Checklist*. The checklist was later adopted by AAAI,⁴ CVPR⁵ and also EMNLP.⁶ The basic idea of the checklist is to motivate authors to reflect about the transparency of models, theories, datasets, code, and the experimental results. The checklist is applicable to many computational research fields, and it has been mentioned in several calls for papers of ML, NLP, Computer Vision and other neighboring disciplines of IR.

Starting off as the P-RECS workshop series from 2018 to 2022 at the high-performance computing conference HPDC [Lofstead et al., 2022], the ACM Emerging Interest Group for Reproducibility and Replicability (EIGREP) evolved with the aim of making it a special interest group for reproducibility.⁷ As of

³<https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf>

⁴<https://aaai.org/aaai-conference/reproducibility-checklist/>

⁵<https://cvpr.thecvf.com/Conferences/2023/AuthorSuggestedPractices>

⁶https://2023.emnlp.org/calls/main_conference_papers/

⁷<https://reproducibility.acm.org/>

now, this interest group organizes the ACM REP conference that follows a multi-disciplinary scope around the topics of reproducibility, data provenance, system log analysis, open source software tools, and others.

5.5 What Can Be Learned From Other Communities?

At this point, the reader may agree that the neighboring communities share a substantial amount of the same reproducibility issues already pointed out earlier in this monograph. We end this section by emphasizing aspects that the IR community can adopt from other communities or enforce if similar structures and directions have already been established.

- **Enforce badging efforts and artifact reviewing:** The SIGIR community has already established the infrastructure for reviewing submissions for reproducibility and availability of the artifacts. In the future, these endeavors should be promoted. Besides assigning papers with badges, citable reproducibility protocols co-authored by original experimenters and reproducers could make reproducibility studies more attractive, as exemplified by the SIGMOD Availability & Reproducibility Initiative.
- **Revitalize workshops about reproducibility:** The IR community inaugurated CENTRE, which is currently on hold for CLEF, TREC, and NTCIR (the last edition was a reproducibility sub-task as part of the NTCIR-18 FairWeb task in 2024 [Tao et al., 2025]). In the NLP domain, a thriving series of workshops currently culminates in the ReproNLP workshop. Reviving the CENTRE workshop series is promising, as it could also provide a venue for negative results. Besides, the IR community has the possibility to continue the cross-conference design by including CLEF, NTCIR, and TREC.
- **Checklists alongside paper submissions:** Many recent calls for papers include guidance on how to write a strong paper. However, the IR community currently lacks a commonly acknowledged list of good practices shared across different venues. Organizers should also motivate authors to provide complete checklists alongside a paper submission, as exemplified by the NLP community.
- **Experimental setup with strong and sound baselines:** As Armstrong et al. [2009a] have already emphasized, weak baselines threaten the validity of outcomes and, thus, reproducibility. Several works in the RS domain point out that it is important to properly tune baselines and share the corresponding source code.
- **Validation of reimplementations:** In the RS domain, a recent study revealed that reimplementations may not be good enough to be used as

baselines [Hidasi and Czapp, 2023a]. Thus, a rigorous validation of reimplémentations is required, which could be implemented with the help of the evaluation framework that is introduced in Section 3.

Chapter 6

Conclusion and Future Directions

In this section, we will wrap up the discussion by highlighting the main take away messages from the previous sections. Finally, we give an outlook about future directions and challenges for reproducibility in IR.

6.1 Reproducible Information Retrieval in Summary

In this work, we have outlined the general scope and the state of ongoing efforts regarding reproducibility in IR research. In order to lay the foundations, we provided a review of the terminology around reproducibility and related terms, and gave an overview of common issues and barriers that are often obstacles towards a successful reproduction in Section 3.

We note that reproducibility is a multi-perspective research challenge that can be seen from different points of view. A simple two-folded perspective includes reproducers, i.e., those who want to reuse and repeat an earlier experiment, and the original authors and experimenters, i.e., those who have the possibility, and ideally the obligation and intention, to prepare their work for others.

In this regard, we outlined a framework to measure the quality of a reproduction. Suppose that no experimental artifacts and software resources are available. In that case, the reproducers reimplement earlier work. The most apparent question in this case is concerned with the quality and correctness of the reproduction. Beyond the very strict requirement of having a bit-wise reproduction, we introduced several reproducibility (and replicability) measures to have a better understanding to what extent an experiment is successfully reproduced. The measures range from a fine-grained analysis of the document ranking to more general statistics based on the comparison of score distributions

between the original results and the reimplementations of an IR experiment. In addition to that, we provide the reader with some recommendations when to apply which measure.

Opposed to the reactive way, where reproducers need to reimplement the experiments, the authors can take action and provide reusable artifacts of their experiments. Section 4 reviewed several good practices of how to share technical components like the computational platform, the software and code, and datasets. Even more, less technical components and aspects can be shared by proper documentation and archiving, as we also outlined in this context. Besides, we also discussed how we can advance IR towards open science and how to integrate FAIR principles of sharing artifacts.

In Section 5, we briefly summarized ongoing efforts in other communities. Overall, the issues and barriers towards reproducibility are mainly the same, and it is also possible to apply the conceptual hierarchy of reproducibility challenges shown in Figure 3.2 to the domains of DB, RS, or NLP. Therefore, these communities also established similar countermeasures like re-validations as part of dedicated conference tracks or artifact badging initiatives. Likewise, there are a couple of software resources that provide reusable, stratified, and systematic experimental setups to enforce better comparability with the state of the art. In the future, these communities could work on solutions for reproducible computational experimentation together in a cross-disciplinary attempt to better understand the key factors of reproducibility on a larger scale. To this end, we also reviewed some current community-bridging efforts.

6.2 Future Challenges Ahead and Open Questions

Throughout the earlier sections, we have seen that reproducibility involves different actors, can be approached from different perspectives, and also at different levels, depending on the particular experiment or finding that is reproduced. Last but not least, the success of reproducibility is also largely affected by the technological standards and the state of the art in a specific research field. From this perspective, reproducibility is also time-dependent, leading to broader questions of what is generalizable and can be considered as common knowledge in the long run. Ultimately, reproducibility should be continuously revalidated to put findings on a larger time scale in order to gain deeper insights about what is generalizable and holds over time.

With the advent of Generative Artificial Intelligence, the landscape of traditional information access systems is complemented by novel ways of how information is directed towards the consumers and users. The shift from providing answers as rankings or recommendations, i.e., in a list- or set-based format, towards abstractive and extractive natural language answers in a text-based, summarizing format revitalizes the questions around reproducibility.

For instance, as of now, it is still an open question how we can measure

a successful reproduction of an information access system that is based on a conversational agent. It is not obvious how our current evaluation standards and practices for reproducibility can be transferred to an experimental setup that mainly relies on generated answers that tremendously deviate from conventional system answers in the form of ranked lists.

Considering the vast amount of data and costly hardware that is required to build and train effective Large Language Models, the core components of these generative models, it is simply not feasible to reproduce an experiment in its entirety from the bottom up in many cases. Likewise, many works and research contributions rely on semi-open or closed system architectures that are often not transparent, presenting a barrier to reproducible experiments [Liesenfeld and Dingemans, 2024].

The importance of reproducible IR methods is further emphasized by Retrieval Augmented Generation-based approaches, where an LLM is provided with a list of retrieved answers to better support the generated answer with facts and to avoid hallucinations [Lewis et al., 2020]. As of now, the importance of having a reliable and reproducible ranking for the generated answer is still under-investigated, even though some recent works suggest that the order and relevance of the retrieved items matters and has an impact on the generated text answers [Liu et al., 2024a, Cuconasu et al., 2024, Tian et al., 2025].

Currently, the IR community does not yet have a shared, collaboratively developed and commonly used LLM that can serve as a common backbone model for generating synthetic relevance labels, (re-)ranking, or RAG. Instead, the field is characterized by a fragmented ecosystem in which numerous closed and (partially) open models are applied to similar problems, making it difficult to directly compare experimental findings. To address this limitation, future research should focus on establishing shared principles, best practices, and governance structures that support the joint development of open, transparent, and fully reproducible foundation models for IR.

Beyond having an exact bit- or list-wise reproduction, there is also the question of what kind of rigor is required for generating text answers that convey the same degree and extent of information without the requirement to reproduce the text answer of a RAG system word-by-word. *What kind of other, possibly different rankings can be used to generate comparable answers with the same level of information?*

In a similar direction, there are also open questions about the users of IR systems. Do we really need bit-wise reproductions of rankings, as a foundation for reproducing a user-oriented experiment? Given that users can compensate for changes in the ranking results by adapting their search behavior to have a similar level of satisfaction [Turpin and Hersh, 2001, Turpin and Scholer, 2006], this requirement could be too strict. Maybe a different ranking can also convey a similar degree of information as that used in the original experiment, and a certain kind of expectable system robustness is still helpful to the users without delivering exact reproductions.

The primary purpose of an IR system is to satisfy the user’s information need. User satisfaction is a multi-faceted concept [Al-Maskari and Sanderson,

2010] that involves many variables such as system effectiveness, but also the user’s effectiveness, effort, characteristics, and expectations. In this regard, the system effectiveness is one aspect out of many, and its influence on the other factors once it deviates from an original experiment has yet to be understood. Of course, this problem has to be discussed by considering the underlying research question of the reproduced user experiment. For instance, several user experiments suggest that the influence of the presentation layout, i.e., the way system results are presented to the user, may outweigh the influence of differences in system effectiveness in certain settings [Roy et al., 2022, Pathak et al., 2024].

The question is *When would a user consider a system as reproduced?* and what do the introduced reproducibility measures tell us beyond the evaluation of a system-oriented IR experiment [Al-Maskari et al., 2007]? Is it also possible to learn something about the level of reproducibility from user interactions with the system answers? The users’ expectations are an important aspect in this context that goes beyond the system-oriented factors and measures. For instance, a reliable information system should deliver factual, timely, and relevant information over time. Especially for time-sensitive topics, the exact same ranking cannot be returned as it would simply neglect new information. Vice versa, completely different system rankings upon repeated requests would make the system appear untrustworthy. In this case, the scope of reproducibility shifts from system-oriented measures to more high-level criteria and concepts such as reliability, usefulness, and trustworthiness.

Beyond these methodological and technological questions, there is also the question of how we can establish a culture of reproducibility and make it an integrated community standard. As in other domains, the rigorous review of complementary materials like data and code ensures a certain kind of improved trustworthiness and accountability. Still, it is debatable if there is enough reward for both making an experiment reproducible and reviewing the reproducibility of it.

The badging mechanism is a promising first direction towards recognizing the additional work by the authors. On the other hand, the reviewing process should also receive a certain kind of reward. Compared to the typical review process of conference or journal articles, reviewing complementary materials, resources, and software is different. Depending on the experiment, reviewing an experiment for reproducibility can require more time and imposes the need for specialized expertise. In addition, hardware requirements have to be met, and for computationally costly experiment, tradeoffs have to be found between the environmental impact and what is of real interest for a reproducibility analysis [Scells et al., 2022, Zuccon et al., 2023].

Another interesting direction towards establishing a culture of reproducibility and to convey its importance, is already made possible by integrating reproducibility reviews into the curriculum [Hofman et al., 2020]. The resulting benefits are obvious. On the one hand, it is much more engaging for students to learn about scientific experiments with some hands-on experience, and they are also sensitized with regard to related difficulties around ensuring reproducibility.

On the other hand, the existing body of research gets reviewed by the future generation, for which it should be of great importance.

Therefore, we encourage our readers to keep reproducibility in mind from the early beginnings of conceptualizing the experiments and thinking about the hypotheses. If reproducibility is understood as a community good, we do not only help ourselves, but also support others in saving time, and focusing on important and reliable IR research.

Bibliography

- SIGMOD/PODS '24: Reproducibility Reports of the 2024 International Conference on Management of Data*, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711794.
- R. L. Ackoff. From Data to Wisdom. *Journal of Applied Systems Analysis*, 16: 3–9, 1989.
- M. Agosti, G. M. Di Nunzio, and N. Ferro. The Importance of Scientific Data Curation for Evaluation Campaigns. In C. Thanos and F. Borri, editors, *DELOS Conference 2007 Working Notes*, pages 185–193. ISTI-CNR, Gruppo ALI, Pisa, Italy, February 2007a.
- M. Agosti, G. M. Di Nunzio, and N. Ferro. Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In Peters et al. [2007], pages 11–20.
- M. Agosti, G. M. Di Nunzio, M. Dussin, and N. Ferro. 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go. In T. Sakai, M. Sanderson, and W. Webber, editors, *Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010)*, pages 16–24. National Institute of Informatics, Tokyo, Japan, 2010.
- M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, and G. Silvello. DIRECTIONS: Design and Specification of an IR Evaluation Infrastructure. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, and G. Santucci, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012)*, pages 88–99. Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany, 2012.
- Azzah Al-Maskari and Mark Sanderson. A review of factors influencing user satisfaction in information retrieval. *J. Assoc. Inf. Sci. Technol.*, 61(5):859–868, 2010. doi: 10.1002/ASI.21300. URL <https://doi.org/10.1002/asi.21300>.
- Azzah Al-Maskari, Mark Sanderson, and Paul D. Clough. The relationship between IR effectiveness measures and user satisfaction. In Wessel Kraaij,

- Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 773–774. ACM, 2007. doi: 10.1145/1277741.1277902. URL <https://doi.org/10.1145/1277741.1277902>.
- Marco Alessio, Guglielmo Faggioli, and Nicola Ferro. DECAF: A modular and extensible conversational search framework. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 3075–3085. ACM, 2023. doi: 10.1145/3539618.3591913. URL <https://doi.org/10.1145/3539618.3591913>.
- Liz Allen, Alison O’Connell, and Veronique Kiermer. How can we ensure visibility and diversity in research contributions? how the contributor role taxonomy (credit) is helping the shift from authorship to contributorship. *Learn. Publ.*, 32(1):71–74, 2019. doi: 10.1002/LEAP.1210. URL <https://doi.org/10.1002/leap.1210>.
- O. Alonso and R. Baeza-Yates, editors. *Information Retrieval: Advanced Topics and Techniques*, 2024. Association for Computing Machinery (ACM), New York, USA.
- Sophia Althammer, Sebastian Hofstätter, and Allan Hanbury. Cross-Domain Retrieval in the Legal and Patent Domains: A Reproducibility Study. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, volume 12657, pages 3–17. Springer International Publishing, 2021. ISBN 978-3-030-72239-5 978-3-030-72240-1. doi: 10.1007/978-3-030-72240-1_1. URL http://link.springer.com/10.1007/978-3-030-72240-1_1.
- W. Alvarez-Giron, J. Tellezz-Torres, J. Tovar-Cortes, and H. Gómez-Adorno. Team qIIMAS on Task 2 - Clustering. In Faggioli et al. [2024], pages 3064–3074.
- E. Amigó, J. Carrillo de Albornoz, M. Almagro-Cádiz, J. Gonzalo, J. Rodriguez-Vidal, and M. F. Verdejo.
- E. Amigó, P. Castells, J. Gonzalo, B. A. Carterette, J. S. Culpepper, and G. Kazai, editors. *Proc. 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022)*, 2022. ACM Press, New York, USA.
- V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, and T. Di Noia. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In Diaz et al. [2021], pages 2405–2414.

- Marco Angelini, Vanessa Fazzini, Nicola Ferro, Giuseppe Santucci, and Gianmaria Silvello. CLAIRe: A combinatorial visual analytics system for information retrieval evaluation. *Inf. Process. Manag.*, 54(6):1077–1100, 2018. doi: 10.1016/J.IPM.2018.04.006. URL <https://doi.org/10.1016/j.ipm.2018.04.006>.
- A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, and N. Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, 2023. Lecture Notes in Computer Science (LNCS) 14163, Springer, Heidelberg, Germany.
- J. Arguello, M. Crane, F. Diaz, J. Lin, and A. Trotman. Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum*, 49(2):107–116, December 2015a.
- J. Arguello, F. Diaz, J. Lin, and A. Trotman. SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). In R. Baeza-Yates, M. Lalmas, A. Moffat, and B. Ribeiro-Neto, editors, *Proc. 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, pages 1147–1148. ACM Press, New York, USA, 2015b.
- Andreas Argyriou, Miguel González-Fierro, and Le Zhang. Microsoft recommenders: Best practices for production-ready recommendation systems. In Amal El Fallah Seghrouchni, Gita Sukthankar, Tie-Yan Liu, and Maarten van Steen, editors, *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 50–51. ACM / IW3C2, 2020. doi: 10.1145/3366424.3382692. URL <https://doi.org/10.1145/3366424.3382692>.
- T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. EvaluatIR: an Online Tool for Evaluating and Comparing IR Systems. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, page 833. ACM Press, New York, USA, 2009a.
- T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements That Don’t Add Up: Ad-Hoc Retrieval Results Since 1998. In Cheung et al. [2009], pages 601–610.
- Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, Darrell Reimer, John T. Richards, Jason Tsay, and Kush R. Varshney. FactSheets: Increasing trust in AI services through supplier’s declarations of conformity. *IBM J. Res. Dev.*, 63(4/5):6:1–6:13, 2019.

- Mohammad Arvan, Luís Pina, and Natalie Parde. Reproducibility in computational linguistics: Is source code enough? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2350–2361. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.150. URL <https://doi.org/10.18653/v1/2022.emnlp-main.150>.
- Manos Athanassoulis, Peter Triantafillou, Raja Appuswamy, Rajesh Bordawekar, Badrish Chandramouli, Xuntao Cheng, Ioana Manolescu, Yannis Papakonstantinou, and Nesime Tatbul. Artifacts availability & reproducibility (VLDB 2021 round table). *SIGMOD Rec.*, 51(2):74–77, 2022. doi: 10.1145/3552490.3552511. URL <https://doi.org/10.1145/3552490.3552511>.
- A. A. Ayele, N. Babakov, J. Bevendorff, X. Bonet Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel Pardo, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, N. Wiegmann, S. Muhie Yimam, and E. Zangerle. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification. In Goeuriot et al. [2024].
- M. Baker. First results from psychology’s largest reproducibility test. *Nature*, April 2015.
- M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, May 2016.
- Giacomo Balloccu, Ludovico Boratto, Christian Cancedda, Gianni Fenu, and Mirko Marras. Knowledge is Power, Understanding is Impact: Utility and Beyond Goals, Explanation Quality, and Fairness in Path Reasoning Recommendation. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 3–19. Springer Nature Switzerland, 2023. ISBN 978-3-031-28241-6. doi: 10.1007/978-3-031-28241-6.1.
- Haim Bar and HaiYing Wang. Reproducible science with LATEX. *Journal of Data Science*, 19(1):111–125, 2021. ISSN 1680-743X. doi: 10.6339/21-JDS998.
- L. A. Barba. Terminologies for Reproducible Research. *arXiv.org, Digital Libraries (cs.DL)*, arXiv:1802.03311, February 2018.
- L. A. Barba. Trustworthy Computational Evidence Through Transparency and Reproducibility. *Computing in Science & Engineering*, 23(1):58–64, January/February 2021.

- M. Barbaro and T. Zeller. A Face Is Exposed For AoL Searcher No. 4417749. *New York Times*, 2006. URL <http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=2&ei=5088&en=996f61c946da4d34&ex=1312776000&partner=rssnyt&emc=rss>.
- A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, and N. Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)*, 2022. Lecture Notes in Computer Science (LNCS) 13390, Springer, Heidelberg, Germany.
- E. Bassani. ranxhub: An Online Repository for Information Retrieval Runs. In Chen et al. [2023], pages 3210–3214.
- Elias Bassani. ranx: A blazing-fast python library for ranking evaluation and comparison. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 259–264. Springer, 2022. doi: 10.1007/978-3-030-99739-7_30. URL https://doi.org/10.1007/978-3-030-99739-7_30.
- J. Beel, C. Breitinger, S. Langer, A. Lommatzsch, and B. Gipp. Towards Reproducibility in Recommender-systems Research. *User Model. User Adapt. Interact.*, 26(1):69–101, 2016.
- Anya Belz and Craig Thomson. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In Anya Belz, Maja Popović, Ehud Reiter, Craig Thomson, and João Sedoc, editors, *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL <https://aclanthology.org/2023.humeval-1.4>.
- Anya Belz and Craig Thomson. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In Simone Balloccu, Anya Belz, Rudali Huidrom, Ehud Reiter, Joao Sedoc, and Craig Thomson, editors, *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.humeval-1.9>.
- Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. A systematic review of reproducibility research in natural language processing. In *EACL*, pages 381–393. Association for Computational Linguistics, 2021a.
- Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. The ReproGen shared task on reproducibility of human evaluations in NLG:

- Overview and results. In *INLG*, pages 249–258. Association for Computational Linguistics, 2021b.
- Anya Belz, Anastasia Shimorina, Maja Popović, and Ehud Reiter. The 2022 ReproGen shared task on reproducibility of evaluations in NLG: Overview and results. In Samira Shaikh, Thiago Ferreira, and Amanda Stent, editors, *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 43–51, Waterville, Maine, USA and virtual meeting, July 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.inlg-genchal.8>.
- Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguistics*, 6:587–604, 2018. doi: 10.1162/TACLA_00041. URL https://doi.org/10.1162/tac1_a_00041.
- Alessandro Benedetti and Anna Ruggero. Stat-Weight: Improving the Estimator of Interleaved Methods Outcomes with Statistical Hypothesis Testing. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 20–34. Springer Nature Switzerland, 2023. ISBN 978-3-031-28241-6. doi: 10.1007/978-3-031-28241-6_2.
- Max Berrendorf, Evgeniy Faerman, Valentyn Melnychuk, Volker Tresp, and Thomas Seidl. Knowledge graph entity alignment with graph convolutional networks: Lessons learned. In *ECIR (2)*, volume 12036 of *Lecture Notes in Computer Science*, pages 3–11. Springer, 2020.
- Max Berrendorf, Ludwig Wacker, and Evgeniy Faerman. A Critical Assessment of State-of-the-Art in Entity Alignment. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, volume 12657, pages 18–32. Springer International Publishing, 2021. ISBN 978-3-030-72239-5 978-3-030-72240-1. doi: 10.1007/978-3-030-72240-1_2. URL http://link.springer.com/10.1007/978-3-030-72240-1_2.
- J. Bevendorff, B. Chulvi, G. L. De La Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, and E. Zangerle. Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection. In K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021)*, pages 419–431. Lecture Notes in Computer Science (LNCS) 12880, Springer, Heidelberg, Germany, 2021.
- J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso,

- E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, and E. Zangerle. Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection. In Barrón-Cedeño et al. [2022], pages 382–394.
- J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, and E. Zangerle. Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection. In Arampatzis et al. [2023], pages 459–481.
- Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Elastic chatnoir: Search engine for the clueweb and the common crawl. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 820–824. Springer, 2018. doi: 10.1007/978-3-319-76941-7_83. URL https://doi.org/10.1007/978-3-319-76941-7_83.
- Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 413–428. Springer International Publishing, 2019. ISBN 978-3-030-15712-8. doi: 10.1007/978-3-030-15712-8_27.
- Prajjwal Bhattarai, Mohammad M. Ghassemi, and Tuka Alhanai. Open-source code repository attributes predict impact of computer science research. In *JCDL*, page 16. ACM, 2022.
- Keping Bi, Xiaojie Sun, Jiafeng Guo, and Xueqi Cheng. Reproducibility Analysis and Enhancements for Multi-aspect Dense Retriever with Aspect Learning. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval*, pages 194–209. Springer Nature Switzerland, 2024. ISBN 978-3-031-56066-8. doi: 10.1007/978-3-031-56066-8_17.
- C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.
- Maurits Bleeker and Maarten de Rijke. Do Lessons from Metric Learning Generalize to Image-Caption Retrieval? In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg,

- and Vinay Setty, editors, *Advances in Information Retrieval*, pages 535–551. Springer International Publishing, 2022. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_36.
- Carl Boettiger. An introduction to Docker for reproducible research. *ACM SIGOPS Oper. Syst. Rev.*, 49(1):71–79, 2015.
- T. Bogers, S. Dodson, M. Gäde, K. Freund, M. M. Hall, M. Koolen, V. Petras, N. Pharo, and M. Skov, editors. *Proc. CHIIR Workshop on Barriers to Interactive IR Resources Re-use (BIIRRR 2019)*, 2019. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <https://ceur-ws.org/Vol-2337/>.
- T. Bogers, M. Gäde, M. M. Hall, M. Koolen, V. Petras, and B. Larsen. How we Work, Share, and Re-use at CHIIR. In J. Gwizdka and S. Young Rieh, editors, *Proc. 8th ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023)*, pages 351–356. ACM Press, New York, USA, 2023.
- K. Bollen, J. T. Cacioppo, R. M. Kaplan, J. A. Knosnick, and J. L. Olds. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf, April 2015.
- A. Bondarenko, M. Fröbe, J. Kiesel, S. Syed, T. Gurcke, M. Beloucif, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, and M. Hagen. Overview of Touché 2022: Argument Retrieval. In Barrón-Cedeño et al. [2022], pages 311–336.
- A. Bondarenko, M. Fröbe, J. Kiesel, F. Schlatt, V. Barriere, B. Ravenet, L. Hemamou, S. Luck, J. H. Reimer, B. Stein, M. Potthast, and M. Hagen. Overview of Touché 2023: Argument and Causal Retrieval. In Arampatzis et al. [2023], pages 507–530.
- Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, Benno Stein, Michael Völske, and Matthias Hagen. Axiomatic retrieval experimentation with *ir-axioms*. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3131–3140. ACM, 2022b. doi: 10.1145/3477495.3531743. URL <https://doi.org/10.1145/3477495.3531743>.
- Ludovico Boratto, Gianni Fenu, and Mirko Marras. The Effect of Algorithmic Bias on Recommender Systems for Massive Open Online Courses. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, Lecture Notes

- in Computer Science, pages 457–472. Springer International Publishing, 2019. ISBN 978-3-030-15712-8. doi: 10.1007/978-3-030-15712-8_30.
- Ludovico Boratto, Gianni Fenu, Mirko Marras, and Giacomo Medda. Consumer Fairness in Recommender Systems: Contextualizing Definitions and Mitigations. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørkvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 552–566. Springer International Publishing, 2022. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_37.
- M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. *Introduction to Meta-Analysis*. John Wiley & Sons, New York, USA, 2009.
- Grant R. Brammer, Ralph W. Crosby, Suzanne J. Matthews, and Tiffani L. Williams. Paper mâché: Creating dynamic reproducible science. In *ICCS*, volume 4 of *Procedia Computer Science*, pages 658–667. Elsevier, 2011.
- António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan van Noord, Dieter Van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5539–5545. European Language Resources Association, 2020. URL <https://aclanthology.org/2020.lrec-1.680/>.
- T. Breuer, P. Schaer, N. Tavakolpoursaleh, J. Schaible, B. Wolff, and B. Müller. STELLA: Towards a Framework for the Reproducibility of Online Search Experiments. In R. Clancy, N. Ferro, C. Hauff, T. Sakai, and Z. Z. Wu, editors, *Proc. of the Open-Source IR Replicability Challenge (OSIRRC 2019)*, pages 8–11. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-2409/>, 2019.
- T. Breuer, N. Ferro, N. Fuhr, M. Maistro, T. Sakai, P. Schaer, and I. Soboroff. How to Measure the Reproducibility of System-oriented IR Experiments. In Chang et al. [2020], pages 349–358.
- T. Breuer, N. Ferro, M. Maistro, and P. Schaer. **repro_eval**: A Python Interface to Reproducibility Measures of System-oriented IR Experiments. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, editors, *Advances in Information Retrieval. Proc. 43rd European Conference on IR Research (ECIR 2021) – Part II*, pages 481–486. Lecture Notes in Computer Science (LNCS) 12657, Springer, Heidelberg, Germany, 2021.

- T. Breuer, J. Keller, and P. Schaer. *ir_metadata: An Extensible Metadata Schema for IR Experiments*. In Amigó et al. [2022], pages 3078–3089.
- T. Breuer, J. Keller, and P. Schaer. *ir_metadata: An Extensible Metadata Schema for Information Retrieval Experiments*. Zenodo, <https://doi.org/10.5281/zenodo.5997491>, February 2022b.
- T. Breuer, N. Fuhr, and P. Schaer. Validating Synthetic Usage Data in Living Lab Environments. *ACM Journal of Data and Information Quality (JDIQ)*, 16(1):5:1–5:33, March 2024a.
- Timo Breuer, Ellen M. Voorhees, and Ian Soboroff. Browsing and Searching Metadata of TREC. In *SIGIR '24: The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, July 14–18, 2024, Washington, DC, USA. ACM, 2024b. doi: 10.1145/3626772.3657873. URL <https://doi.org/10.1145/3626772.3657873>.
- C. Buckley and E. M. Voorhees. Retrieval System Evaluation. In D. K. Harman and E. M. Voorhees, editors, *TREC. Experiment and Evaluation in Information Retrieval*, pages 53–78. MIT Press, Cambridge (MA), USA, 2005.
- Jamie Callan and Alistair Moffat. Panel on use of proprietary data. *SIGIR Forum*, 46(2):10–18, 2012. doi: 10.1145/2422256.2422258. URL <https://doi.org/10.1145/2422256.2422258>.
- K. A. Carbine, Lindsey. H. M., R. E. Rodeback, and M. J. Larson. Quantifying evidential value and selective reporting in recent and 10-year past psychophysiological literature: A pre-registered P-curve analysis. *International Journal of Psychophysiology*, 142:33–49, 2019.
- Ben Carterette. The best published result is random: Sequential testing and its effect on reported effectiveness. In *SIGIR*, pages 747–750. ACM, 2015.
- Y. Chang, X. Cheng, J. Huang, Y. Lu, J. Kamps, V. Murdock, J.-R. Wen, A. Diriye, J. Guo, and O. Kurland, editors. *Proc. 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, 2020. ACM Press, New York, USA.
- O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In Cheung et al. [2009], pages 621–630.
- H.-H. Chen, W.-J. Duh, H.-H. Huang, M. P. Kato, J. Mothe, and B. Poblete, editors. *Proc. 46th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, 2023. ACM Press, New York, USA.
- D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors. *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009)*, 2009. ACM Press, New York, USA.

- Patrick John Chia, Jacopo Tagliabue, Federico Bianchi, Chloe He, and Brian Ko. Beyond ndcg: Behavioral testing of recommender systems with reclist. WWW '22 Companion, page 99–104, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391306. doi: 10.1145/3487553.3524215. URL <https://doi.org/10.1145/3487553.3524215>.
- Fernando Chirigati, Rémi Rampin, Dennis E. Shasha, and Juliana Freire. ReproZip: Computational reproducibility with ease. In *SIGMOD Conference*, pages 2085–2088. ACM, 2016.
- D. E. Chubin. Open Science and Closed Science: Tradeoffs in a Democracy. *Science, Technology, & Human Values*, 10(2):73–81, April 1985.
- Valentina Ciriani, Sabrina De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati. Microdata protection. In Ting Yu and Sushil Jajodia, editors, *Secure Data Management in Decentralized Systems*, volume 33 of *Advances in Information Security*, pages 291–321. Springer, 2007. doi: 10.1007/978-0-387-27696-0_9. URL https://doi.org/10.1007/978-0-387-27696-0_9.
- R. Clancy, N. Ferro, C. Hauff, T. Sakai, and Z. Z. Wu. The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). In B. Piwowarski, M. Chevalier, E. Gaussier, Y. Maarek, J.-Y. Nie, and F. Scholer, editors, *Proc. 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1432–1434. ACM Press, New York, USA, 2019.
- C. W. Cleverdon. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK, 1962.
- C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. *Aslib Proceedings*, 19(6):173–194, 1967.
- J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Mahwah (NJ), USA, 2nd edition, 1988.
- Christian S. Collberg and Todd A. Proebsting. Repeatability in computer systems research. *Commun. ACM*, 59(3):62–69, 2016.
- K. Collins-Thompson, F. Diaz, C. L. A. Clarke, and E. M. Voorhees. TREC 2013 Web Track Overview. In Voorhees [2014].
- K. Collins-Thompson, C. Macdonald, P. N. Bennett, and E. M. Voorhees. TREC 2014 Web Track Overview. In Voorhees and Ellis [2015].
- Roberto Di Cosmo and Stefano Zacchiroli. Software heritage: Why and how to preserve software source code. In Shoichiro Hara, Shigeo Sugimoto, and Makoto Goto, editors, *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017, Kyoto, Japan, September 25-29, 2017*, 2017. URL <https://hdl.handle.net/11353/10.931064>.

- Ludovic Courtès. Functional package management with guix. In Christian Queinnec and Manuel Serrano, editors, *Proceedings of ELS 2013 - 6th European Lisp Symposium, Madrid, Spain, June 3-4, 2013*, pages 4–14. ELSAA, 2013. URL <https://european-lisp-symposium.org/static/proceedings/2013.pdf#page=10>.
- Matt Crane. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics*, 6:241–252, 2018.
- Nick Craswell, David Hawking, and Stephen E. Robertson. Effective site finding using link anchor information. pages 250–257. ACM, 2001. doi: 10.1145/383952.383999. URL <https://doi.org/10.1145/383952.383999>.
- Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. ORCAS: 20 million clicked query-document pairs for analyzing search. In *CIKM*, pages 2983–2989. ACM, 2020.
- Jennifer Crocker and M. Lynne Cooper. Addressing scientific fraud. *Science*, 334(6060):1182–1182, December 2011. ISSN 1095-9203. doi: 10.1126/science.1216775. URL <http://dx.doi.org/10.1126/science.1216775>.
- W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors. *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, 1998. ACM Press, New York, USA.
- Sebastian Cross, Guido Zuccon, and Ahmed Mourad. A Reproducibility Study of Question Retrieval for Clarifying Questions. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 35–50. Springer Nature Switzerland, 2023. ISBN 978-3-031-28241-6. doi: 10.1007/978-3-031-28241-6_3.
- Michael R. Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, Stian Soiland-Reyes, Bogdan Gavrilović, Carole Goble, and The CWL Community. Methods included: standardizing computational reuse and portability with the common workflow language. *Commun. ACM*, 65(6):54–63, may 2022. ISSN 0001-0782. doi: 10.1145/3486897. URL <https://doi.org/10.1145/3486897>.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 719–729. ACM, 2024. doi: 10.1145/3626772.3657834. URL <https://doi.org/10.1145/3626772.3657834>.

- M. de Rijke. Retrieval as Interaction. Tony Kent Strix Annual Memorial Lecture. London, UK. Video of the lecture available at <https://www.youtube.com/watch?v=Zb6YGoiPt8M>, 23 November 2018.
- C. Deng, Y. Zhao, X. Tang, M. Gerstein, and A. Cohan. Investigating Data Contamination in Modern Benchmarks for Large Language Models. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024*, pages 8706–8719. Association for Computational Linguistics, 2024.
- G. M. Di Nunzio and N. Ferro. DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In A. Rauber, S. Christodoulakis, and A. Min Tjoa, editors, *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, pages 483–484. Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 2005.
- G. M. Di Nunzio and N. Ferro. Scientific Evaluation of a DLMS: a service for evaluating information access components. In J. Gonzalo, C. Thanos, M. F. Verdejo, and R. C. Carrasco, editors, *Proc. 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, pages 536–539. Lecture Notes in Computer Science (LNCS) 4172, Springer, Heidelberg, Germany, 2006.
- G. M. Di Nunzio and S. Marchesin. The University of Padua IMS research group at CENTRE@TREC 2018. In Voorhees and Ellis [2019].
- F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai, A. Bellogín, and M. Yoshioka, editors. *Proc. 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, 2021. ACM Press, New York, USA.
- Michael Dittenbach, Bernhard Pflügfelder, Andreas Pesenhofer, Giovanna Roda, and Helmut Berger. SOIRE: a service-oriented IR evaluation architecture. In David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy Lin, editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 2101–2102. ACM, 2009. doi: 10.1145/1645953.1646324. URL <https://doi.org/10.1145/1645953.1646324>.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL <https://aclanthology.org/D19-1224>.

- DOI Foundation. *DOI Handbook*. DOI Foundation, <https://doi.org/10.1000/182>, April 2023.
- Eelco Dolstra and Andres Löb. Nixos: a purely functional linux distribution. In James Hook and Peter Thiemann, editors, *Proceeding of the 13th ACM SIGPLAN international conference on Functional programming, ICFP 2008, Victoria, BC, Canada, September 20-28, 2008*, pages 367–378. ACM, 2008. doi: 10.1145/1411204.1411255. URL <https://doi.org/10.1145/1411204.1411255>.
- Eelco Dolstra, Merijn de Jonge, and Eelco Visser. Nix: A safe and policy-free system for software deployment. In Lee Damon, editor, *Proceedings of the 18th Conference on Systems Administration (LISA 2004), Atlanta, USA, November 14-19, 2004*, pages 79–92. USENIX, 2004. URL <http://www.usenix.org/publications/library/proceedings/lisa04/tech/dolstra.html>.
- Chris Drummond. Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, Montreal, Canada, 2009*, Evaluation Methods for Machine Learning Workshop, the 26th ICML, June 14-18, 2009, Montreal, Canada, Collection / Collection : NRC Publications Archive / Archives des publications du CNRC, 2009.
- M. Dussin and N. Ferro. Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany, 2009.
- Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. doi: 10.1561/04000000042. URL <https://doi.org/10.1561/04000000042>.
- Alexander Dür, Andreas Rauber, and Peter Filzmoser. Reproducing a Neural Question Answering Architecture Applied to the SQuAD Benchmark Dataset: Challenges and Lessons Learned. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 102–113. Springer International Publishing, 2018. ISBN 978-3-319-76941-7. doi: 10.1007/978-3-319-76941-7_8.
- Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. The Impact of Cross-Lingual Adjustment of Contextual Word Representations on Zero-Shot Transfer. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 51–67. Springer Nature Switzerland, 2023. ISBN 978-3-031-28241-6. doi: 10.1007/978-3-031-28241-6_4.

- Nadav Eiron and Kevin S. McCurley. Analysis of anchor text for web search. pages 459–460. doi: 10.1145/860435.860550. URL <https://doi.org/10.1145/860435.860550>.
- M. D. Ekstrand. Lenskit for python: Next-generation software for recommender systems experiments. In M. d’Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudre Mauroux, editors, *Proc. 29th International Conference on Information and Knowledge Management (CIKM 2020)*, pages 2999–3006. ACM Press, New York, USA, 2020.
- Hakan Erdogmus, Maurizio Morisio, and Marco Torchiano. On the effectiveness of the test-first approach to programming. *IEEE Trans. Software Eng.*, 31(3):226–237, 2005.
- Emily Escamilla, Martin Klein, Talya Cooper, Vicky Rampin, Michele C. Weigle, and Michael L. Nelson. The rise of github in scholarly publications. In Gianmaria Silvello, Óscar Corcho, Paolo Manghi, Giorgio Maria Di Nunzio, Koraljka Golub, Nicola Ferro, and Antonella Poggi, editors, *Linking Theory and Practice of Digital Libraries - 26th International Conference on Theory and Practice of Digital Libraries, TPD L 2022, Padua, Italy, September 20-23, 2022, Proceedings*, volume 13541 of *Lecture Notes in Computer Science*, pages 187–200. Springer, 2022. doi: 10.1007/978-3-031-16802-4_15. URL https://doi.org/10.1007/978-3-031-16802-4_15.
- Emily Escamilla, Lamia Salsabil, Martin Klein, Jian Wu, Michele C. Weigle, and Michael L. Nelson. It’s not just github: Identifying data and software sources included in publications. In Omar Alonso, Helena Cousijn, Gianmaria Silvello, Mónica Marrero, Carla Teixeira Lopes, and Stefano Marchesin, editors, *Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPD L 2023, Zadar, Croatia, September 26-29, 2023, Proceedings*, volume 14241 of *Lecture Notes in Computer Science*, pages 195–206. Springer, 2023. doi: 10.1007/978-3-031-43849-3_17. URL https://doi.org/10.1007/978-3-031-43849-3_17.
- European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union, OJ L 119, 4.5.2016*, 59: 1–88, May 2016.
- European Commission. Corrigendum to Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union, OJ L 127, 23.5.2018*, 61:2–5, May 2018.

- G. Faggioli, N. Ferro, P. Galuščáková, and A. Garcia Seco de Herrera, editors. *CLEF 2024 Working Notes*, 2024. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <https://ceur-ws.org/Vol-3740/>.
- W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In R. Baeza-Yates and F. Bonchi, editors, *Proc. 30th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2024)*, pages 6491–6501. ACM Press, New York, USA, 2024.
- Michael Färber. Analyzing the GitHub repositories of research papers. In *JCDL*, pages 491–492. ACM, 2020.
- Marcos Fernández-Pichel, David E. Losada, Juan C. Pichel, and David Elswailer. Reliability Prediction for Health-Related Content: A Replicability Study. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, volume 12657, pages 47–61. Springer International Publishing, 2021. ISBN 978-3-030-72239-5 978-3-030-72240-1. doi: 10.1007/978-3-030-72240-1_4. URL https://link.springer.com/10.1007/978-3-030-72240-1_4.
- M. Ferrante, N. Ferro, and M. Maistro. Injecting User Models and Time into Precision via Markov Chains. In S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, and K. Järvelin, editors, *Proc. 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*, pages 597–606. ACM Press, New York, USA, 2014.
- A. Ferrara, C. Pomo, and N. Tonellotto. Enhancing Reproducibility and Replicability in Information Retrieval: A Path Towards Scientific Integrity and Effective Research. In C. Hauff, G. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, and N. Tonellotto, editors, *Advances in Information Retrieval. Proc. 47th European Conference on IR Research (ECIR 2025) – Part V*, pages 266–272. Lecture Notes in Computer Science (LNCS) 15576, Springer, Heidelberg, Germany, 2025.
- M. Ferrari Dacrema, P. Cremonesi, and D. Jannach. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In T. Bogers, A. Said, P. Brusilovsky, and D. Tikk, editors, *Proc. 13th ACM Conference on Recommender Systems (RecSys 2019)*, pages 101–109. ACM Press, New York, USA, 2019.
- M. Ferrari Dacrema, P. Cremonesi, and D. Jannach. Methodological Issues in Recommender Systems Research (Extended Abstract). In C. Bessiere, editor, *Proc. 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pages 4706–4710. ijcai.org, 2020.
- M. Ferrari Dacrema, F. Moroni, R. Nembrini, N. Ferro, G. Faggioli, and P. Cremonesi. Towards Feature Selection for Ranking and Classification Exploiting Quantum Annealers. In Amigó et al. [2022], pages 2814–2824.

- M. Ferrari Dacrema, M. Benigni, and N. Ferro. Reproducibility and Artifact Consistency of the SIGIR 2022 Recommender Systems Papers Based on Message Passing. *ACM Transactions on Information Systems (TOIS)*, 2026.
- Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Trans. Inf. Syst.*, 39(2):20:1–20:49, 2021.
- N. Ferro. Specification of the CIRCO Framework, Version 0.10. Technical Report IMS.2009.CIRCO.0.10, Department of Information Engineering, University of Padua, Italy, 2009.
- N. Ferro. Reproducibility Challenges in Information Retrieval Evaluation. *ACM Journal of Data and Information Quality (JDIQ)*, 8(2):8:1–8:4, February 2017.
- N. Ferro and D. Harman. CLEF 2009: Grid@CLEF Pilot Track Overview. In Peters et al. [2010], pages 552–565.
- N. Ferro and D. Kelly. SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10, June 2018.
- N. Ferro and M. Maistro. Evaluation of IR Systems. In Alonso and Baeza-Yates [2024].
- N. Ferro and C. Peters. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer International Publishing, Germany, 2019a.
- N. Ferro and M. Sanderson. How do you Test a Test? A Multifaceted Examination of Significance Tests. In K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, and J. Tang, editors, *Proc. 15th ACM International Conference on Web Searching and Data Mining (WSDM 2022)*, pages 280–288. ACM Press, New York, USA, 2022.
- N. Ferro and M. Sanderson. Uncontextualized significance considered dangerous. In G. Hui Yang, Wang. H., S. Han, C. Hauff, G. Zuccon, and Y. Zhang, editors, *Proc. 47th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*, pages 261–270. ACM Press, New York, USA, 2024.
- N. Ferro and G. Silvello. A General Linear Mixed Models Approach to Study System Component Effects. In Perego et al. [2016], pages 25–34.
- N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on Reproducibility of Data-Oriented Experiments in e-Science. *SIGIR Forum*, 50(1):68–82, June 2016a.

- N. Ferro, M. Maistro, T. Sakai, and I. Soboroff. Overview of CENTRE@CLEF 2018: a First Tale in the Systematic Reproducibility Realm. In P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J.-Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, pages 239–246. Lecture Notes in Computer Science (LNCS) 11018, Springer, Heidelberg, Germany, 2018a.
- N. Ferro, M. Maistro, T. Sakai, and I. Soboroff. CENTRE@CLEF2018: Overview of the Replicability Task. In L. Cappellato, N. Ferro, J.-Y. Nie, and L. Soulier, editors, *CLEF 2018 Working Notes*. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-2125/>, 2018b.
- N. Ferro, N. Fuhr, M. Maistro, T. Sakai, and I. Soboroff. Overview of CENTRE@CLEF 2019: Sequel in the Systematic Reproducibility Realm. In F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)*, pages 287–300. Lecture Notes in Computer Science (LNCS) 11696, Springer, Heidelberg, Germany, 2019a.
- N. Ferro, N. Fuhr, M. Maistro, T. Sakai, and I. Soboroff. CENTRE@CLEF2019: Overview of the Replicability and Reproducibility Tasks. In L. Cappellato, N. Ferro, D. E. Losada, and H. Müller, editors, *CLEF 2019 Working Notes*. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-2380/>, 2019b.
- Nicola Ferro and Carol Peters, editors. *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer, 2019b.
- Nicola Ferro and Gianmaria Silvello. Rank-Biased Precision Reloaded: Reproducibility and Generalization. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 768–780. Springer International Publishing, 2015. ISBN 978-3-319-16354-3. doi: 10.1007/978-3-319-16354-3_83.
- Nicola Ferro and Gianmaria Silvello. A general linear mixed models approach to study system component effects. In *SIGIR*, pages 25–34. ACM, 2016b.
- Nicola Ferro, Allan Hanbury, Henning Müller, and Giuseppe Santucci. Harnessing the scientific data produced by the experimental evaluation search engines and information access systems. In Mitsuhsa Sato, Satoshi Matsuoka, Peter M. A. Sloot, G. Dick van Albada, and Jack J. Dongarra, editors, *Proceedings of the International Conference on Computational Science, ICCS 2011, Nanyang Technological University, Singapore, 1-3 June, 2011*, volume 4 of *Procedia Computer Science*, pages 740–749. Elsevier, 2011. doi:

- 10.1016/J.PROCS.2011.04.078. URL <https://doi.org/10.1016/j.procs.2011.04.078>.
- Nicola Ferro, Norbert Fuhr, Kalervo Järvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. Increasing reproducibility in IR: Findings from the dagstuhl seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum*, 50(1):68–82, 2016b.
- Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. Overview of CENTRE@CLEF 2019: Sequel in the systematic reproducibility realm. In *CLEF*, volume 11696 of *Lecture Notes in Computer Science*, pages 287–300. Springer, 2019c.
- Nicola Ferro, Stefano Marchesin, Alberto Purpura, and Gianmaria Silvello. A docker-based replicability study of a neural information retrieval model. In *OSIRRC@SIGIR*, volume 2409 of *CEUR Workshop Proceedings*, pages 37–43. CEUR-WS.org, 2019d.
- Matthias Fey, Jan Eric Lenssen, Christopher Morris, Jonathan Masci, and Nils M. Kriege. Deep graph matching consensus. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=HyeJf1HKvS>.
- R. A. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, UK, 1925.
- Jessica Forde, Tim Head, Chris Holdgraf, Yuvi Panda, Gladys Nalvarete, Benjamin Ragan-Kelley, and Erik Sundell. Reproducible research environments with repo2docker. In *Workshop on Reproducibility in Machine Learning RML@ICML 2018*, 2018.
- J. Freire, N. Fuhr, and A. Rauber, editors. *Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science*, volume 6 of *Dagstuhl Reports*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany, 2016.
- Juliana Freire, David Koop, Emanuele Santos, and Cláudio T. Silva. Provenance for computational tasks: A survey. *Comput. Sci. Eng.*, 10(3):11–21, 2008. doi: 10.1109/MCSE.2008.79. URL <https://doi.org/10.1109/MCSE.2008.79>.
- J. I. Friesse and N. Fuhr. Towards Reproducibility of Interactive Retrieval Experiments: Framework and Case Study. In C. Hauff, G. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, and N. Tonellotto, editors, *Advances in Information Retrieval. Proc. 47th European Conference on IR Research (ECIR 2025) – Part IV*, pages 146–160. Lecture Notes in Computer Science (LNCS) 15575, Springer, Heidelberg, Germany, 2025.
- M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, and M. Potthast. The Information Retrieval Experiment Platform. In Chen et al. [2023], pages 2826–2836.

- M. Fröbe, D. Alexander, G. Hendriksen, F. Schlatt, M. Hagen, and M. Potthast. Team OpenWebSearch at CLEF 2024: QuantumCLEF. In Faggioli et al. [2024], pages 3075–3081.
- Maik Fröbe, Sebastian Günther, Maximilian Probst, Martin Potthast, and Matthias Hagen. The power of anchor text in the neural retrieval era. In *ECIR (1)*, volume 13185 of *Lecture Notes in Computer Science*, pages 567–583. Springer, 2022.
- Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. Continuous integration for reproducible shared tasks with tira.io. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, volume 13982 of *Lecture Notes in Computer Science*, pages 236–241. Springer, 2023. doi: 10.1007/978-3-031-28241-6_20. URL https://doi.org/10.1007/978-3-031-28241-6_20.
- N. Fuhr. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum*, 51(3):32–41, December 2017.
- Norbert Fuhr. Proof by experimentation? Towards better IR research. In *SIGIR*, page 2. ACM, 2020.
- Maria Gäde, Marijn Koolen, Mark M. Hall, Toine Bogers, and Vivien Petras. A manifesto on resource re-use in interactive information retrieval. In *CHIIR*, pages 141–149. ACM, 2021.
- G. Galilei. *Dialogues Concerning Two New Sciences*. Dover Publications, USA, 1954.
- Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv.org, Computation and Language (cs.CL)*, arXiv:2312.10997, March 2024.
- T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, and H. Daumé III. Datasheets for datasets. *Communications of the ACM (CACM)*, 64(12):86–92, December 2021.
- A. Gelman and D. Weakliem. Of Beauty, Sex and Power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, 97(4):310–316, July-August 2009.
- E. Gibney. This AI researcher is trying to ward off a reproducibility crisis. *Nature*, 577:14, January 2020.

- Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. The ACL Anthology: Current state and future directions. In Eun-jeong L. Park, Masato Hagiwara, Dmitrijs Milajevs, and Liling Tan, editors, *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2504. URL <https://aclanthology.org/W18-2504>.
- L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, and N. Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024) – Part 2*, 2024. Lecture Notes in Computer Science (LNCS) 14959, Springer, Heidelberg, Germany.
- J. Gonzalo and D. W. Oard. iCLEF 2004 Track Overview: Pilot Experiments in Interactive Cross-Language Question Answering. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004) Revised Selected Papers*, pages 310–322. Lecture Notes in Computer Science (LNCS) 3491, Springer, Heidelberg, Germany, 2005.
- J. Gonzalo, P. Clough, and J. Karlgren. Overview of iCLEF 2008: Search Log Analysis for Multilingual Image Retrieval. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, and A. Peñas, editors, *Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross-Language Evaluation Forum (CLEF 2008). Revised Selected Papers*, pages 227–235. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany, 2009.
- J. Gonzalo, V. Peinado, P. Clough, and J. Karlgren. Overview of iCLEF 2009: Exploring Search Behaviour in a Multilingual Folksonomy Environment. In C. Peters, T. Tsikrika, H. Müller, J. Kalpathy-Cramer, G. J. F. Jones, J. Gonzalo, and B. Caputo, editors, *Multilingual Information Access Evaluation Vol. II Multimedia Experiments – Tenth Workshop of the Cross-Language Evaluation Forum (CLEF 2009). Revised Selected Papers*, pages 13–20. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2010.
- Pieter Van Gorp and Steffen Mazanek. SHARE: A web portal for creating and sharing executable research papers. In *ICCS*, volume 4 of *Procedia Computer Science*, pages 589–597. Elsevier, 2011.
- Adrien Grand, Robert Muir, Jim Ferenczi, and Jimmy Lin. From MAXSCORE to block-max wand: The story of how lucene significantly improved query evaluation performance. In *ECIR (2)*, volume 12036 of *Lecture Notes in Computer Science*, pages 20–27. Springer, 2020.

- Jiafeng Guo, Yixing Fan, Xiang Ji, and Xueqi Cheng. Matchzoo: A learning, practicing, and developing system for neural text matching. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1297–1300. ACM, 2019. doi: 10.1145/3331184.3331403. URL <https://doi.org/10.1145/3331184.3331403>.
- H. Gurulingappa, A. Bauer, L. Toldo, C. Schepers, and G. Megaro. Semi-supervised Information Retrieval System for Clinical Decision Support. In Voorhees and Ellis [2017].
- Christophe Van Gysel and Maarten de Rijke. Pytrec.eval: An Extremely Fast Python Interface to trec.eval. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *SIGIR*, pages 873–876. ACM, 2018. doi: 10.1145/3209978.3210065. <https://doi.org/10.1145/3209978.3210065>.
- Laurel L. Haak, Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. Orcid: a system to uniquely identify researchers. *Learned Publishing*, 25(4):259–264, October 2012. ISSN 1741-4857. doi: 10.1087/20120404. URL <http://dx.doi.org/10.1087/20120404>.
- Matthias Hagen, Martin Potthast, Michel Büchner, and Benno Stein. Twitter Sentiment Detection via Ensemble Classification Using Averaged Confidence Scores. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval, Lecture Notes in Computer Science*, pages 741–754. Springer International Publishing, 2015. ISBN 978-3-319-16354-3. doi: 10.1007/978-3-319-16354-3_81.
- Tim Hagen, Maik Fröbe, Jan Heinrich Merker, Harris Scells, Matthias Hagen, and Martin Potthast. Tirex tracker: The information retrieval experiment tracker. In Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne, editors, *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 3764–3771. ACM, 2025. doi: 10.1145/3726302.3730297. URL <https://doi.org/10.1145/3726302.3730297>.
- M. M. Hall, H. Huurdeman, M. Koolen, M. Skov, and D. Walsh. Overview of the INEX 2014 Interactive Social Book Search Track. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Labs and Workshops, Notebook Papers*, pages 480–493. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-1180/>, 2014.
- D. K. Harman, editor. *The Third Text REtrieval Conference (TREC-3)*, 1995. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA.

- D. K. Harman. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA, 2011.
- D. K. Harman and E. M. Voorhees. *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge (MA), USA, 2005.
- Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. On the Reproducibility of the TAGME Entity Linking System. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 436–449. Springer International Publishing, 2016. ISBN 978-3-319-30671-1. doi: 10.1007/978-3-319-30671-1.32.
- David Hawking, Bodo von Billerbeck, Paul Thomas, and Nick Craswell. *Simulating Information Retrieval Test Collections*. Morgan & Claypool, 2020. ISBN 978-1-68173-957-1.
- M. A. Heroux, L. A. Barba, M. Parashar, V. Stodden, and M. Taufer. Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences. SANDIA REPORT, SAND2018-11186, October 2018.
- T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, USA, 2009.
- Balázs Hidasi and Ádám Tibor Czapp. The effect of third party implementations on reproducibility. In Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song, editors, *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, pages 272–282. ACM, 2023a. doi: 10.1145/3604915.3609487. URL <https://doi.org/10.1145/3604915.3609487>.
- Balázs Hidasi and Ádám Tibor Czapp. Widespread flaws in offline evaluation of recommender systems. In Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song, editors, *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, pages 848–855. ACM, 2023b. doi: 10.1145/3604915.3608839. URL <https://doi.org/10.1145/3604915.3608839>.
- Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, USA, 1987.
- Jake M. Hofman, Daniel G. Goldstein, Siddhartha Sen, and Forough Poursabzi-Sangdeh. Expanding the scope of reproducibility research through data analysis replications. In *WWW (Companion Volume)*, pages 567–571. ACM / IW3C2, 2020.

- K. Hofmann, L. Li, and F. Radlinski. Online Evaluation for Information Retrieval. *Foundations and Trends in Information Retrieval (FnTIR)*, 10(1): 1–117, June 2016.
- F. Hopfgartner, A. Hanbury, H. Müller, N. Kando, S. Mercer, J. Kalpathy-Cramer, M. Potthast, T. Gollub, A. Krithara, J. Lin, K. Balog, and I. Eggel. Report on the Evaluation-as-a-Service (EaaS) Expert Workshop. *SIGIR Forum*, 49(1):57–65, June 2015.
- F. Hopfgartner, A. Hanbury, H. Müller, I. Eggel, K. Balog, T. Brodt, G. V. Cormack, J. Lin, J. Kalpathy-Cramer, N. Kando, M. P. Kato, A. Krithara, T. Gollub, M. Potthast, E. Viegas, and S. Mercer. Evaluation-as-a-Service for the Computational Sciences: Overview and Outlook. *ACM Journal of Data and Information Quality (JDIQ)*, 10(4):15:1–15:32, November 2018.
- Benjamin D. Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 759–766, 2017. doi: 10.1609/icwsm.v11i1.14976.
- J. C. Hsu. *Multiple Comparisons. Theory and methods*. Chapman and Hall/CRC, USA, 1996.
- D. A. Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, pages 329–338. ACM Press, New York, USA, 1993.
- P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, Heidelberg, Germany, 2005.
- ISO 26324. Information and documentation – Digital object identifier system. Recommendation ISO 26324:2022, 2022.
- Peter Ivie and Douglas Thain. Reproducibility in scientific computing. *ACM Comput. Surv.*, 51(3):63:1–63:36, 2018.
- K. Järvelin and J. Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, October 2002.
- Ivo Jimenez, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, Jay F. Lofstead, Carlos Maltzahn, Kathryn Mohror, and Robert Ricci. PopperCI: Automated reproducibility validation. In *INFOCOM Workshops*, pages 450–455. IEEE, 2017a.
- Ivo Jimenez, Michael Sevilla, Noah Watkins, Carlos Maltzahn, Jay F. Lofstead, Kathryn M. Mohror, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. The popper convention: Making reproducible systems evaluation

- practical. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops, IPDPS Workshops 2017, Orlando / Buena Vista, FL, USA, May 29 - June 2, 2017*, pages 1561–1570. IEEE Computer Society, 2017b. doi: 10.1109/IPDPSW.2017.157. URL <https://doi.org/10.1109/IPDPSW.2017.157>.
- Serdar Kadioglu and Bernard Kleynhans. Building higher-order abstractions from the components of recommender systems. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22998–23004. AAAI Press, 2024. doi: 10.1609/AAAI.V38I21.30341. URL <https://doi.org/10.1609/aaai.v38i21.30341>.
- Serdar Kadioglu and Melinda Thielbar. Surrogate modeling to address the absence of protected membership attributes in fairness evaluation. *ACM Trans. Evol. Learn. Optim.*, 2024. doi: 10.1145/3700145. URL <https://doi.org/10.1145/3700145>.
- Chris Kamphuis, Arjen P. de Vries, Leonid Boytsov, and Jimmy Lin. Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 28–34. Springer International Publishing, 2020. ISBN 978-3-030-45442-5. doi: 10.1007/978-3-030-45442-5_4.
- Noriko Kando, editor. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, NTCIR-1, Tokyo, Japan, August 30 - September 1, 1999*. National Center for Science Information Systems (NACSIS), 1999.
- M. Kaneko and T. Baldwin. Investigating How Pre-training Data Leakage Affects Models’ Reproduction and Detection Capabilities. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*, pages 23557–23567. Association for Computational Linguistics, 2025. ISBN 979-8-89176-164-3.
- S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):1–12, September 2023.
- J. Karlgren, J. Gonzalo, and P. Clough. iCLEF 2006 Overview: Searching the Flickr WWW Photo-Sharing Repository. In Peters et al. [2007], pages 186–194.
- D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval (FnTIR)*, 3(1–2):1–224, 2009.

- M. G. Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33 (3):239–251, November 1945.
- S. Kharazmi, F. Scholer, D. Vallet, and M. Sanderson. Examining Additivity and Weak Baselines. *ACM Transactions on Information Systems (TOIS)*, 34 (4):23:1–23:18, June 2016.
- J. Kiesel, C. Çöltekin, H. Heinrich, M. Fröbe, M. Alshomary, B. De Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevicius, T. Reitis-Münstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, and B. Stein. Overview of Touché 2024: Argumentation Systems. In Goeuriot et al. [2024].
- Johannes Kiesel, Lars Meyer, Florian Kneist, Benno Stein, and Martin Potthast. An Empirical Comparison of Web Page Segmentation Algorithms. In Djord Hienstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, volume 12657, pages 62–74. Springer International Publishing, 2021. ISBN 978-3-030-72239-5 978-3-030-72240-1. doi: 10.1007/978-3-030-72240-1_5. URL http://link.springer.com/10.1007/978-3-030-72240-1_5.
- K. Kimmel, M. L. Avolio, and P. J. Ferraro. Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nature Ecology & Evolution*, 7:1525–1536, September 2023.
- Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E. Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B. Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter Development Team. Jupyter Notebooks - a publishing format for reproducible computational workflows. In *ELPUB*, pages 87–90. IOS Press, 2016.
- M. Koolen, T. Bogers, M. Gäde, M. A. Hall, H. C. Huurdeman, J. Kamps, M. Skov, E. Toms, and D. Walsh. Overview of the CLEF 2015 Social Book Search Lab. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. J. F. Jones, E. SanJuan, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixth International Conference of the CLEF Association (CLEF 2015)*, pages 545–564. Lecture Notes in Computer Science (LNCS) 9283, Springer, Heidelberg, Germany, 2015.
- M. Koolen, T. Bogers, M. Gäde, M. Hall, I. Hendrickx, H. C. Huurdeman, J. Kamps, M. Skov, S. Verberne, and D. Walsh. Overview of the CLEF 2016 Social Book Search Lab. In N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016)*, pages 351–370. Lecture Notes in Computer Science (LNCS) 9822, Springer, Heidelberg, Germany, 2016.

- Dominik Kowald, Markus Schedl, and Elisabeth Lex. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 35–42. Springer International Publishing, 2020. ISBN 978-3-030-45442-5. doi: 10.1007/978-3-030-45442-5_5.
- M. Kozlov. ‘Publish or Perish’ is now a card game — not just an academic’s life. *Nature*, 632:483, August 2024.
- Hans-Peter Kriegel, Erich Schubert, and Arthur Zimek. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*, 52(2):341–378, 2017.
- T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, USA, 3rd edition, 1996.
- E. Kulinskaya, S. Morgenthaler, and R. G. Staudte. *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*. John Wiley & Sons, New York, USA, 2008.
- Wojciech Kusa, Allan Hanbury, and Petr Knuth. Automation of Citation Screening for Systematic Literature Reviews Using Neural Networks: A Replicability Study. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 584–598. Springer International Publishing, 2022. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_39.
- Weronika Lajewska and Krisztian Balog. From Baseline to Top Performer: A Reproducibility Study of Approaches at the TREC 2021 Conversational Assistance Track. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 177–191. Springer Nature Switzerland, 2023. ISBN 978-3-031-28241-6. doi: 10.1007/978-3-031-28241-6_12.
- R. R. Larson. Decomposing Text Processing for Retrieval: Cheshire tries GRID@CLEF. In Peters et al. [2010].
- Jeremy Leipzig, Daniel Nüst, Charles Tapley Hoyt, Karthik Ram, and Jane Greenberg. The role of metadata in reproducible computational research. *Patterns*, 2(9):100322, 2021.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference*

- on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- Hang Li, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback: A Reproducibility Study. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørsvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 599–612. Springer International Publishing, 2022. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_40.
- Jiayu Li, Hanyu Li, Zhiyu He, Weizhi Ma, Peijie Sun, Min Zhang, and Shaoping Ma. Rechorus2. 0: A modular and task-flexible recommendation library. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 454–464, 2024.
- Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative AI: open washing and the EU AI act. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*, pages 1774–1787. ACM, 2024. doi: 10.1145/3630106.3659005. URL <https://doi.org/10.1145/3630106.3659005>.
- J. Lin. Building a Culture of Reproducibility in Academic Research. *arXiv.org, Information Retrieval (cs.IR)*, arXiv:2212.13534, December 2022.
- Jimmy Lin and Qian Zhang. Reproducibility is a Process, Not an Achievement: The Replicability of IR Reproducibility Experiments. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval, Lecture Notes in Computer Science*, pages 43–49. Springer International Publishing, 2020. ISBN 978-3-030-45442-5. doi: 10.1007/978-3-030-45442-5_6.
- Jimmy Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig Macdonald, and Sebastiano Vigna. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval, Lecture Notes in Computer Science*, pages 408–420. Springer International Publishing, 2016. ISBN 978-3-319-30671-1. doi: 10.1007/978-3-319-30671-1_30.
- Jimmy Lin, Joel M. Mackenzie, Chris Kamphuis, Craig Macdonald, Antonio Mallia, Michal Siedlaczek, Andrew Trotman, and Arjen P. de Vries. Supporting interoperability between open-source search engines with the common index file format. In *SIGIR*, pages 2149–2152. ACM, 2020.

- Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *SIGIR*, pages 2356–2362. ACM, 2021a.
- Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2021b.
- A. Lipani, F. Piroi, L. Andersson, and A. Hanbury. An Information Retrieval Ontology for Information Retrieval Nanopublications. In E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms, editors, *Information Access Evaluation – Multilinguality, Multimodality, and Interaction. Proceedings of the Fifth International Conference of the CLEF Initiative (CLEF 2014)*, pages 44–49. Lecture Notes in Computer Science (LNCS) 8685, Springer, Heidelberg, Germany, 2014.
- Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *ACM Queue*, 17(1):80, 2019.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024a. doi: 10.1162/TACL_A_00638. URL https://doi.org/10.1162/tacl_a_00638.
- Yuanna Liu, Ming Li, Mozhdeh Ariannezhad, Masoud Mansoury, Mohammad Aliannejadi, and Maarten De Rijke. Measuring Item Fairness in Next Basket Recommendation: A Reproducibility Study. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval*, volume 14611, pages 210–225. Springer Nature Switzerland, 2024b. ISBN 978-3-031-56065-1 978-3-031-56066-8. doi: 10.1007/978-3-031-56066-8_18. URL https://link.springer.com/10.1007/978-3-031-56066-8_18.
- Jay F. Lofstead, Carlos Maltzahn, and Ivo Jimenez. P-recs’22: 5th international workshop on practical reproducible evaluation of systems. In Jon B. Weissman, Abhishek Chandra, Ada Gavrilovska, and Devesh Tiwari, editors, *HPDC ’22: The 31st International Symposium on High-Performance Parallel and Distributed Computing, Minneapolis, MN, USA, 27 June 2022 - 1 July 2022*, pages 298–299. ACM, 2022. doi: 10.1145/3502181.3535106. URL <https://doi.org/10.1145/3502181.3535106>.
- Pasquale Lops, Marco Polignano, Cataldo Musto, Antonio Silletti, and Giovanni Semeraro. Clayrs: An end-to-end framework for reproducible knowledge-aware recommender systems. *Inf. Syst.*, 119:102273, 2023.
- D. E. Losada, J. Parapar, and A. Barreiro. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management*, 53(5):1005–1025, September 2017.

- A. Lucic, M. Bleeker, M. de Rijke, K. Sinha, S. Jullien, and R. Stojnic. Towards Reproducible Machine Learning Research in Information Retrieval. In Amigó et al. [2022], pages 3459–3461.
- Ana Lucic, Maurits J. R. Bleeker, Sami Jullien, Samarth Bhargav, and Maarten de Rijke. Reproducibility as a mechanism for teaching fairness, accountability, confidentiality, and transparency in artificial intelligence. In *AAAI*, pages 12792–12800. AAAI Press, 2022b.
- Malte Ludewig, Noemi Mauro, Sara Latifi, and Dietmar Jannach. Empirical analysis of session-based recommendation algorithms. *User Model. User Adapt. Interact.*, 31(1):149–181, 2021. doi: 10.1007/s11257-020-09277-1. URL <https://doi.org/10.1007/s11257-020-09277-1>.
- Xueguang Ma, Kai Sun, Ronak Pradeep, Minghan Li, and Jimmy Lin. Another Look at DPR: Reproduction of Training and Replication of Retrieval. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørkvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 613–626. Springer International Publishing, 2022. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_41.
- S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian. Simplified Data Wrangling with `ir_datasets`. In Diaz et al. [2021], pages 2429–2436.
- Sean MacAvaney and Craig Macdonald. A python interface to pisa! In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3339–3344. ACM, 2022. doi: 10.1145/3477495.3531656. URL <https://doi.org/10.1145/3477495.3531656>.
- Sean MacAvaney, Craig Macdonald, and Iadh Ounis. Streamlining evaluation with `ir-measures`. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørkvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 305–310. Springer, 2022a. doi: 10.1007/978-3-030-99739-7_38. URL https://doi.org/10.1007/978-3-030-99739-7_38.
- Sean MacAvaney, Craig Macdonald, and Iadh Ounis. Reproducing Personalised Session Search Over the AOL Query Log. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørkvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 627–640. Springer International Publishing, 2022b. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_42.

- Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. Pyterrier: Declarative experimentation in python from BM25 to dense retrieval. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 4526–4533. ACM, 2021. doi: 10.1145/3459637.3482013. URL <https://doi.org/10.1145/3459637.3482013>.
- Joel Mackenzie, Antonio Mallia, Matthias Petri, J. Shane Culpepper, and Torsten Suel. Compressing Inverted Indexes with Recursive Graph Bisection: A Reproducibility Study. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 339–352. Springer International Publishing, 2019. ISBN 978-3-030-15712-8. doi: 10.1007/978-3-030-15712-8_22.
- Stuart Mackie, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Experiments in Newswire Summarisation. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 421–435. Springer International Publishing, 2016. ISBN 978-3-319-30671-1. doi: 10.1007/978-3-319-30671-1_31.
- Ian Magnusson, Noah A. Smith, and Jesse Dodge. Reproducibility in NLP: what have we learned from the checklist? In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12789–12811. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.809. URL <https://doi.org/10.18653/v1/2023.findings-acl.809>.
- M. Maistro, T. Breuer, P. Schaer, and N. Ferro. An in-depth Investigation on the Behavior of Measures to Quantify Reproducibility. *Information Processing & Management*, 60(3), 2023.
- D. Malitesta, C. Pomo, V. W. Anelli, T. Di Noia, and A. Ferrara. An Out-of-the-Box Application for Reproducible Graph Collaborative Filtering extending the Elliot Framework. In *Adjunct Proc. 31st ACM Conference on User Modeling, Adaptation and Personalization, (UMAP 2023)*, pages 12–15. ACM Press, New York, USA, 2023.
- Antonio Mallia, Michal Siedlaczek, Joel M. Mackenzie, and Torsten Suel. PISA: Performant indexes and search for academia. In *OSIRRC@SIGIR*, volume 2409 of *CEUR Workshop Proceedings*, pages 50–56. CEUR-WS.org, 2019.
- Alberto Carlo Maria Mancino, Salvatore Bui, Angela Di Fazio, Antonio Ferrara, Daniele Malitesta, Claudio Pomo, and Tommaso Di Noia. Datarec: A python

- library for standardized and reproducible data management in recommender systems. In *SIGIR*, pages 3478–3487. ACM, 2025.
- Ioana Manolescu, Loredana Afanasiev, Andrei Arion, Jens Dittrich, Stefan Manegold, Neoklis Polyzotis, Karl Schnaitter, Pierre Senellart, Spyros Zoupanos, and Dennis E. Shasha. The repeatability experiment of SIGMOD 2008. *SIGMOD Rec.*, 37(1):39–45, 2008. doi: 10.1145/1374780.1374791. URL <https://doi.org/10.1145/1374780.1374791>.
- Jarana Manotumruksa, Dimitrios Rafailidis, Craig Macdonald, and Iadh Ounis. On Cross-Domain Transfer in Venue Recommendation. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 443–456. Springer International Publishing, 2019. ISBN 978-3-030-15712-8. doi: 10.1007/978-3-030-15712-8_29.
- R. McCreddie, R. Deveaud, D. Albakour, S. Mackie, N. Limsopatham, C. Macdonald, I. Ounis, and T. Thonet. University of Glasgow at TREC 2014: Experiments with Terrier in Contextual Suggestion, Temporal Summarisation and Web Tracks. In Voorhees and Ellis [2015].
- Lien Michiels, Robin Verachtert, and Bart Goethals. Recpack: An(other) experimentation toolkit for top-n recommendation using implicit feedback data. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys ’22, page 648–651, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. doi: 10.1145/3523227.3551472. URL <https://doi.org/10.1145/3523227.3551472>.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *FAT*, pages 220–229. ACM, 2019.
- A. Moffat and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1): 2:1–2:27, December 2008.
- Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher Tomkins-Tinch, Vanessa V. Sochat, Jan Forster, Soohyun Lee, Sven Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with snakemake. *F1000Research*, 10:33, 2021. doi: 10.12688/F1000RESEARCH.29032.1. URL <https://doi.org/10.12688/f1000research.29032.1>.
- Alejandro Moreo and Fabrizio Sebastiani. Re-assessing the “Classify and Count” Quantification Method. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, volume 12657, pages 75–91. Springer International Publishing, 2021. ISBN 978-3-030-72239-5 978-3-030-72240-1. doi:

- 10.1007/978-3-030-72240-1_6. URL http://link.springer.com/10.1007/978-3-030-72240-1_6.
- V. M. Moskovkin. Origin and evolution of the “publish or perish” phenomenon. Leiden Madtrics, Centre for Science and Technology Studies (CWTS) at Leiden University, <https://dx.doi.org/10.59350/tbhbb-spm02>, June 2024.
- Rajdeep Mukherjee, Shreyas Shetty, Subrata Chattopadhyay, Subhadeep Maji, Samik Datta, and Pawan Goyal. Reproducibility, Replicability and Beyond: Assessing Production Readiness of Aspect Based Sentiment Analysis in the Wild. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, volume 12657, pages 92–106. Springer International Publishing, 2021. ISBN 978-3-030-72239-5 978-3-030-72240-1. doi: 10.1007/978-3-030-72240-1_7. URL http://link.springer.com/10.1007/978-3-030-72240-1_7.
- K. R. Murphy, B. Myers, and A. Wolach. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. Routledge, Taylor & Francis Group, UK, 4th edition, 2014.
- NASEM. *Reproducibility and Replicability in Science*. National Academies of Sciences, Engineering, and Medicine. The National Academies Press, Washington, USA, 2019.
- R. Nembrini, M. Ferrari Dacrema, and P. Cremonesi. Feature Selection for Recommender Systems with Quantum Computing. *Entropy*, 23(8):970:1–970:22, July 2021.
- Nicola Neophytou, Bhaskar Mitra, and Catherine Stinson. Revisiting Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørkvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 641–654. Springer International Publishing, 2022. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_43.
- NISO. *NISO RP-31-2021 – Reproducibility Badging and Definitions*. National Information Standards Organization (NISO), January 2021.
- J. Niu, J. Li, K. Deng, and Y. Ren. CRUISE on Quantum Computing for Feature Selection in Recommender Systems. In Faggioli et al. [2024], pages 3096–3104.
- N. F. Noy and C. A. Goble. Are We Cobblers without Shoes?: Making Computer Science Data FAIR. *Communications of the ACM (CACM)*, 66(1):36–38, January 2023.
- Giorgio Maria Di Nunzio and Nicola Ferro. DIRECT: A system for evaluating information access components of digital libraries. In *ECDL*, volume 3652 of *Lecture Notes in Computer Science*, pages 483–484. Springer, 2005.

- Daniel Nüst, Vanessa Sochat, Ben Marwick, Stephen J. Eglen, Tim Head, Tony Hirst, and Benjamin D. Evans. Ten simple rules for writing Dockerfiles for reproducible data science. *Plos Computational Biology*, 16(11), 2020.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):943–952, August 2015.
- Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In *ECIR*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005.
- João R. M. Palotti, Harrisen Scells, and Guido Zuccon. Trectools: an open-source python library for information retrieval practitioners involved in trec-like campaigns. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1325–1328. ACM, 2019. doi: 10.1145/3331184.3331399. URL <https://doi.org/10.1145/3331184.3331399>.
- Luca Papariello, Alexandros Bampoulidis, and Mihai Lupu. On the Replicability of Combining Word Embeddings and Retrieval Models. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 50–57. Springer International Publishing, 2020. ISBN 978-3-030-45442-5. doi: 10.1007/978-3-030-45442-5_7.
- H. Pashler and E.-J. Wagenmakers. Editors’ Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7(6):528–530, November 2012.
- A. Pasin and N. Ferro. KIMERA: From Evaluation-as-a-Service to Evaluation-in-the-Cloud. In N. Ferro, M. Maistro, G. Pasi, O. Alonso, A. Trotman, and S. Verberne, editors, *Proc. 48th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025)*, pages 3584–3594. ACM Press, New York, USA, 2025.
- A. Pasin, M. Ferrari Dacrema, P. Cremonesi, and N. Ferro. qCLEF: a Proposal to Evaluate Quantum Annealing for Information Retrieval and Recommender Systems. In Arampatzis et al. [2023], pages 97–108.
- A. Pasin, W. Cunha, M. A. Gonçalves, and N. Ferro. A Quantum Annealing Instance Selection Approach for Efficient and Effective Transformer Fine-Tuning. In H. Oosterhuis, H. Bast, and C. Xiong, editors, *Proc. 10th ACM SIGIR International Conference on the Theory of Information Retrieval (IC-TIR 2024)*, pages 205–214. ACM Press, New York, USA, 2024a.
- A. Pasin, M. Ferrari Dacrema, P. Cremonesi, and N. Ferro. Overview of QuantumCLEF 2024: The Quantum Computing Challenge for Information Retrieval and Recommender Systems at CLEF. In Goeuriot et al. [2024].

- Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In Xiaohua Jia, editor, *Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006*, volume 152 of *ACM International Conference Proceeding Series*, page 1. ACM, 2006. doi: 10.1145/1146847.1146848. URL <https://doi.org/10.1145/1146847.1146848>.
- Kanaad Pathak, Leif Azzopardi, and Martin Halvey. The influence of presentation and performance on user satisfaction. In Paul D. Clough, Morgan Harvey, and Frank Hopfgartner, editors, *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2024, Sheffield, United Kingdom, March 10-14, 2024*, pages 77–86. ACM, 2024. doi: 10.1145/3627508.3638335. URL <https://doi.org/10.1145/3627508.3638335>.
- Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6, 2023. URL <http://jmlr.org/papers/v24/21-1436.html>.
- Gustavo Penha and Claudia Hauff. Do the Findings of Document and Passage Retrieval Generalize to the Retrieval of Responses for Dialogues? In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 132–147. Springer Nature Switzerland, 2023. ISBN 978-3-031-28241-6. doi: 10.1007/978-3-031-28241-6_9.
- R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, and J. Zobel, editors. *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*, 2016. ACM Press, New York, USA.
- C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors. *Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006). Revised Selected Papers*, 2007. Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany.
- C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda, editors. *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross-Language Evaluation Forum (CLEF 2009). Revised Selected Papers*, 2010. Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany.
- V. Petras, T. Bogers, M. Hall, J. Savoy, P. Malak, A. Pawlowski, N. Ferro, and I. Masiero. Cultural Heritage in CLEF (CHiC) 2013. In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the*

- Fourth International Conference of the CLEF Initiative (CLEF 2013)*, pages 192–211. Lecture Notes in Computer Science (LNCS) 8138, Springer, Heidelberg, Germany, 2013.
- V. Petras, T. Bogers, and M. Gäde. Elements of IIR Studies: A Review of the 2006-2018 IiX and CHIIR Conferences. In Bogers et al. [2019], pages 37–41.
- V. Petras, M. Koolen, M. Gäde, and T. Bogers. Experiences with the 2013-2016 CLEF Interactive Information Retrieval Tracks. In Bogers et al. [2019], pages 29–36.
- A. V. Petrov and C. Macdonald. Effective and Efficient Training for Sequential Recommendation using Recency Sampling. In J. Golbeck, F. M. Harper, V. Murdock, M. D. Ekstrand, B. Shapira, B. Basilico, K. T. Lundgaard, and E. Oldridge, editors, *Proc. 16th ACM Conference on Recommender Systems (RecSys 2022)*, pages 81–91. ACM Press, New York, USA, 2022.
- João Felipe Pimentel, Leonardo Murta, Vanessa Braganholo, and Juliana Freire. A large-scale study about quality and reproducibility of jupyter notebooks. In *MSR*, pages 507–517. IEEE / ACM, 2019.
- J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Lariviere, A. Beygelzimer, F. d’Alche Buc, E. Fox, and H. Larochelle. Improving Reproducibility in Machine Learning Research. A Report from the NeurIPS 2019 Reproducibility Program. *Journal of Machine Learning Research (JMLR)*, 22(164):1–20, May 2021.
- Benjamin Piwowarski. Experimaestro and datamaestro: Experiment and dataset managers (for IR). In *SIGIR*, pages 2173–2176. ACM, 2020.
- H. E. Plesser. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, 11:76:1–76:4, January 2018.
- K. Popper. *The Logic of Scientific Discovery*. Routledge, Taylor & Francis Group, UK, 2nd edition, 2002.
- M. Potthast, T. Gollub, M. Wiegmann, and B. Stein. TIRA Integrated Research Architecture. In N. Ferro and C. Peters, editors, *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 123–160. Springer International Publishing, Germany, 2019.
- Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Löttsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efsthios Stamatatos, and Matthias Hagen. Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens,

- Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 393–407. Springer International Publishing, 2016. ISBN 978-3-319-30671-1. doi: 10.1007/978-3-319-30671-1_29.
- Martin Potthast, Sebastian Günther, Janek Bevendorff, Jan Philipp Bittner, Alexander Bondarenko, Maik Fröbe, Christian Kahmann, Andreas Niekler, Michael Völske, Benno Stein, and Matthias Hagen. The information retrieval anthology. In *SIGIR*, pages 2550–2555. ACM, 2021.
- Ronak Pradeep, Haonan Chen, Lingwei Gu, Manveer Singh Tamber, and Jimmy Lin. PyGaggle: A gaggle of resources for open-domain question answering. In *ECIR (3)*, volume 13982 of *Lecture Notes in Computer Science*, pages 148–162. Springer, 2023a.
- Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *CoRR*, abs/2309.15088, 2023b. doi: 10.48550/ARXIV.2309.15088. URL <https://doi.org/10.48550/arXiv.2309.15088>.
- Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *CoRR*, abs/2312.02724, 2023c. doi: 10.48550/ARXIV.2312.02724. URL <https://doi.org/10.48550/arXiv.2312.02724>.
- Fernando Benjamín Pérez Maurera, Maurizio Ferrari Dacrema, and Paolo Cremonesi. An Evaluation Study of Generative Adversarial Networks for Collaborative Filtering. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 671–685. Springer International Publishing, 2022. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_45.
- W. V. Quine. *From Stimulus to Science*. Harvard University Press, Cambridge (MA), USA, 1998.
- Edward Raff. A step toward quantifying independently reproducible machine learning research. In *NeurIPS*, pages 5486–5496, 2019.
- Karthik Ram. Git can facilitate greater reproducibility and increased transparency in science. *Source Code Biol. Medicine*, 8:7, 2013.
- Jinfeng Rao, Jimmy Lin, and Miles Efron. Reproducible Experiments on Lexical and Temporal Feedback for Tweet Search. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 755–767. Springer International Publishing, 2015. ISBN 978-3-319-16354-3. doi: 10.1007/978-3-319-16354-3_82.

- A. Rauber, V. Braganholo, J. Dittrich, N. Ferro, J. Freire, N. Fuhr, D. Garijo, C. Goble, K. Järvelin, B. Ludäscher, B. Stein, and R. Stotzka. PRIMAD – Information gained by different types of reproducibility. In J. Freire, N. Fuhr, and A. Rauber, editors, *Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science*, Dagstuhl Reports, Volume 6, Number 1, pages 128–132. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany, 2016a.
- Andreas Rauber, Ari Asmi, Dieter Van Uytvanck, and Stefan Pröll. Identification of reproducible subsets for data citation, sharing and re-use. *Bull. IEEE Tech. Comm. Digit. Libr.*, 12(1), 2016b.
- Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. TripClick: The log files of a large health web search engine. In *SIGIR*, pages 2507–2513. ACM, 2021.
- Blandine Ribotta, Ronan Bellemin, Jon E Grahe, Hans IJzerman, Jordan R Wagge, and Mélody Mailliez. Bringing replication into classroom: Benefits for education, science, and society. *PsyArXiv*, February 2022. doi: 10.31234/osf.io/8dhbg. <https://doi.org/10.31234/osf.io/8dhbg>.
- K. Roberts, D. Demner-Fushman, E. M. Voorhees, and W. R. Hersh. Overview of the TREC 2016 Clinical Decision Support Track. In Voorhees and Ellis [2017].
- S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC–3. In Harman [1995], pages 109–126.
- Adam Roegiest and Gordon V. Cormack. An architecture for privacy-preserving and replicable high-recall retrieval experiments. In Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel, editors, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 1085–1088. ACM, 2016. doi: 10.1145/2911451.2911456. URL <https://doi.org/10.1145/2911451.2911456>.
- Adam Roegiest, Gordon V. Cormack, Charles L. A. Clarke, and Maura R. Grossman. TREC 2015 total recall track overview. In *TREC*, volume 500–319 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2015.
- R. Rosenthal. The “File Drawer Problem” and Tolerance for Null Results. *Psychological Bulletin*, 86(3):638–641, 1979.
- J. Rowley. The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science*, 33(2):163–180, 2007.
- Dwaipayan Roy, Mandar Mitra, and Debasis Ganguly. To clean or not to clean: Document preprocessing and reproducibility. *ACM J. Data Inf. Qual.*, 10(4): 18:1–18:25, 2018.

- Nirmal Roy, David Maxwell, and Claudia Hauff. Users and contemporary SERPs: A (re-)investigation. In *SIGIR*, pages 2765–2775. ACM, 2022.
- I. Ruthven and D. Kelly. *Interactive Information Seeking, Behaviour and Retrieval*. Facet Publishing, UK, 2011.
- A. Said and A. Bellogín. Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks. In A. Kobsa, M. Z. Zhou, M. Ester, and Y. Koren, editors, *Proc. 8th ACM Conference on Recommender Systems (RecSys 2014)*, pages 129–136. ACM Press, New York, USA, 2014.
- A. Said and A. Bellogín. Replicable Evaluation of Recommender Systems. In H. Werthner, M. Zanker, J. Golbeck, and G. Semeraro, editors, *Proc. 9th ACM Conference on Recommender Systems (RecSys 2015)*, pages 363–364. ACM Press, New York, USA, 2015.
- T. Sakai. Evaluating Evaluation Metrics based on the Bootstrap. In E. N. Efthimiadis, S. Dumais, D. Hawking, and K. Järvelin, editors, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 525–532. ACM Press, New York, USA, 2006.
- T. Sakai. On the reliability of information retrieval metrics based on graded relevance. *Information Processing & Management*, 42(2):531–548, March 2007.
- T. Sakai. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In Perego et al. [2016], pages 5–14.
- T. Sakai. *Laboratory Experiments in Information Retrieval*, volume 40 of *The Information Retrieval Series*. Springer Singapore, 2018.
- T. Sakai. On Fuhr’s Guideline for IR Evaluation. *SIGIR Forum*, 54(1):p14:1–p14:8, June 2020.
- T. Sakai, N. Ferro, I. Soboroff, Z. Zeng, P. Xiao, and M. Maistro. Overview of the NTCIR-14 CENTRE Task. In E. Ishita, N. Kando, M. P. Kato, and Y. Liu, editors, *Proc. 14th NTCIR Conference on Evaluation of Information Access Technologies*, pages 494–509. National Institute of Informatics, Tokyo, Japan, 2019.
- T. Sakai, S. Tao, Z. Zeng, Y. Zheng, J. Mao, Z. Chu, Y. Liu, M. Maistro, Z. Dou, N. Ferro, and I. Soboroff. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. In N. Kando, M. P. Kato, and Y. Liu, editors, *Proc. 15th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-15)*. National Institute of Informatics, Tokyo, Japan, 2020.
- T. Sakai, D. W. Oard, and N. Kando. *Evaluating Information Retrieval and Access Tasks – NTCIR’s Legacy of Research Impact*, volume 43 of *The Information Retrieval Series*. Springer International Publishing, Germany, 2021.

- T. Sakai, S. Tao, Z. Chu, M. Maistro, Y. Li, N. Chen, N. Ferro, J. Wang, I. Soboroff, and Y. Liu. Overview of the NTCIR-16 We Want Web with CENTRE (WWW-4) Task. In N. Kando, C. L. A. Clarke, M. P. Kato, and Y. Liu, editors, *Proc. 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16)*. National Institute of Informatics, Tokyo, Japan, 2022.
- Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. Cornac: A comparative framework for multimodal recommender systems. *J. Mach. Learn. Res.*, 21: 95:1–95:5, 2020.
- Pierangela Samarati. Protecting respondents’ identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, 13(6):1010–1027, 2001. doi: 10.1109/69.971193. URL <https://doi.org/10.1109/69.971193>.
- M. Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR)*, 4(4):247–375, 2010.
- Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura, editors. *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*. ACM, 2020.
- J. Savoy. Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management*, 33(44):495–512, 1997.
- Harrison Scells and Martin Potthast. pyboolir: A toolkit for domain-specific search experiments. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 3190–3194. ACM, 2023. doi: 10.1145/3539618.3591819. URL <https://doi.org/10.1145/3539618.3591819>.
- Harrison Scells, Shengyao Zhuang, and Guido Zuccon. Reduce, reuse, recycle: Green information retrieval research. In *SIGIR*, pages 2825–2837. ACM, 2022.
- Fabian Schliski, Jörg Schlötterer, and Michael Granitzer. Influence of Random Walk Parametrization on Graph Embeddings. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 58–65. Springer International Publishing, 2020. ISBN 978-3-030-45442-5. doi: 10.1007/978-3-030-45442-5.8.
- D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and Ali Rahimi. Winner’s curse? On pace, progress, and empirical rigor. In *ICLR (Workshop)*. Open-Review.net, 2018.

- Faisal Shehzad and Dietmar Jannach. Everyone’s a winner! on hyperparameter tuning of recommendation models. In Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song, editors, *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, pages 652–657. ACM, 2023. doi: 10.1145/3604915.3609488. URL <https://doi.org/10.1145/3604915.3609488>.
- Faisal Shehzad and Dietmar Jannach. Performance Comparison of Session-Based Recommendation Algorithms Based on GNNs. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval*, pages 115–131. Springer Nature Switzerland, 2024. ISBN 978-3-031-56066-8. doi: 10.1007/978-3-031-56066-8_12.
- Faisal Shehzad, Timo Breuer, Maria Maistro, and Dietmar Jannach. ”we share our code online”: Why this is not enough to ensure reproducibility and progress in recommender systems research. In Mária Bielíková, Pavel Kordík, Markus Schedl, Marco de Gemmis, Sole Pera, Rodrigo Alves, Olivier Jeunen, and Vito Ostuni, editors, *Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys 2025, Prague, Czech Republic, September 22-26, 2025*, pages 884–893. ACM, 2025. doi: 10.1145/3705328.3748157. URL <https://doi.org/10.1145/3705328.3748157>.
- Anu Shrestha and Francesca Spezzano. Textual Characteristics of News Title and Body to Detect Fake News: A Reproducibility Study. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, volume 12657, pages 120–133. Springer International Publishing, 2021. ISBN 978-3-030-72239-5 978-3-030-72240-1. doi: 10.1007/978-3-030-72240-1_9. URL http://link.springer.com/10.1007/978-3-030-72240-1_9.
- Thiago Silva, Nicollas Silva, Heitor Werneck, Carlos Mito, Adriano C. M. Pereira, and Leonardo Rocha. irect: An interactive recommendation framework. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR ’22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3165–3175. ACM, 2022. doi: 10.1145/3477495.3531754. URL <https://doi.org/10.1145/3477495.3531754>.
- G. Silvello, G. Bordea, N. Ferro, P. Buitelaar, and T. Bogers. Semantic Representation and Enrichment of Information Retrieval Experimental Data. *International Journal on Digital Libraries (IJDL)*, 18(2):145–172, June 2017.
- Gianmaria Silvello, Riccardo Bucco, Giulio Busato, Giacomo Fornari, Andrea Langeli, Alberto Purpura, Giacomo Rocco, Alessandro Tezza, and Maristella Agosti. Statistical Stemmers: A Reproducibility Study. In Gabriella Pasi,

- Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 385–397. Springer International Publishing, 2018. ISBN 978-3-319-76941-7. doi: 10.1007/978-3-319-76941-7_29.
- A. Singhal, J. Choi, D. Hindle, and F. C. N. Pereira. AT&T at TREC-6: SDR Track. In E. M. Voorhees and D. K. Harman, editors, *The Sixth Text REtrieval Conference (TREC-6)*, pages 227–232. National Institute of Standards and Technology (NIST), Special Publication 500-240, Washington, USA, 1997.
- J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Proc. 26th International Conference on Neural Information Processing Systems (NIPS 2012)*, pages 2960–2968. MIT Press, Cambridge (MA), USA, 2012.
- I. Soboroff, N. Ferro, M. Maistro, and T. Sakai. Overview of the TREC 2018 CENTRE Track. In Voorhees and Ellis [2019].
- Raphaël Sourty, José G. Moreno, Lynda Tamine, and François-Paul Servant. CHERCHE: A new tool to rapidly implement pipelines in information retrieval. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3283–3288. ACM, 2022. doi: 10.1145/3477495.3531695. URL <https://doi.org/10.1145/3477495.3531695>.
- K. Spärck Jones. *Information Retrieval Experiment*. Butterworths, London, United Kingdom, 1981.
- M Staudinger, W. Kusa, F. Piroi, A. Lipani, and A. Hanbury. A Reproducibility and Generalizability Study of Large Language Models for Query Generation. In T. Sakai, E. Ishita, H. Ohshima, F. Hasibi, J. Mao, and J. M. Jose, editors, *Proc. of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2024)*, pages 186–196. ACM Press, New York, USA, 2024.
- T. D. Sterling. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, 54(285):30–34, March 1959.
- T. D. Sterling, W. L. Rosenbaum, and J. J. Weinkam. Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician*, 49(1):108–112, February 1995.
- Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. Indri: A language model-based search engine for complex queries. In *Proceedings*

- of the *International Conference on Intelligent Analysis*, volume 2, pages 2–6. Citeseer, 2005.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1355. URL <https://doi.org/10.18653/v1/p19-1355>.
- Student. The Probable Error of a Mean. *Biometrika*, 6(1):1–25, March 1908.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14918–14937. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.923. URL <https://doi.org/10.18653/v1/2023.emnlp-main.923>.
- Z. Sun, D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang, and C. Geng. Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison. In R. L. T. Santos, L. B. Marinho, E. M. Daly, L. Chen, K. Falk, N. Koenigstein, and E. S. de Moura, editors, *Proc. 14th ACM Conference on Recommender Systems (RecSys 2020)*, pages 23–32. ACM Press, New York, USA, 2020.
- Latanya Sweeney. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10(5):557–570, 2002. doi: 10.1142/S0218488502001648. URL <https://doi.org/10.1142/S0218488502001648>.
- J. M. Tague-Sutcliffe and J. Blustein. A Statistical Analysis of the TREC-3 Data. In Harman [1995], pages 385–398.
- Manveer Singh Tamber, Ronak Pradeep, and Jimmy Lin. Pre-processing Matters! Improved Wikipedia Corpora for Open-Domain Question Answering. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 163–176. Springer Nature Switzerland, 2023. ISBN 978-3-031-28241-6. doi: 10.1007/978-3-031-28241-6_11.
- Luchen Tan, Gaurav Baruah, and Jimmy Lin. On the reusability of "Living Labs" test collections: : A case study of real-time summarization. In *SIGIR*, pages 793–796. ACM, 2017.
- S. Tao, T. Sakai, J. Wang, H. Fang, Y. Zhang, H. Li, Y. Tu, Chen N., and M. Maistro. Overview of the NTCIR-18 FairWeb-2 Task. In M. P. Kato,

- N. Kando, C. L. A. Clarke, and Y. Liu, editors, *Proc. 18th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-18)*. National Institute of Informatics, Tokyo, Japan, 2025.
- Noriko Kando Tetsuya Sakai, Douglas W. Oard, editor. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Springer Singapore, 2021. ISBN 9789811555541. doi: 10.1007/978-981-15-5554-1. URL <http://dx.doi.org/10.1007/978-981-15-5554-1>.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html>.
- F. Tian, D. Ganguly, and C. Macdonald. Is Relevance Propagated from Retriever to Generator in RAG? In C. Hauff, G. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, and N. Tonellotto, editors, *Advances in Information Retrieval. Proc. 47th European Conference on IR Research (ECIR 2025) – Part I*, pages 32–48. Lecture Notes in Computer Science (LNCS) 15572, Springer, Heidelberg, Germany, 2025.
- N. Tonellotto. Neural Information Retrieval. In Alonso and Baeza-Yates [2024].
- J. W. Tukey. Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2):99–114, June 1949.
- Andrew Turpin and William R. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR*, pages 225–231. ACM, 2001.
- Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *SIGIR*, pages 11–18. ACM, 2006.
- UNESCO. UNESCO Recommendation on Open Science. UNESCO, Paris, France, SC-PCB-SPP/2021/OS/UROS, 2021.
- E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In Croft et al. [1998], pages 315–323.
- E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, September 2000.
- E. M. Voorhees, editor. *The Twenty-Second Text REtrieval Conference Proceedings (TREC 2013)*, 2014. National Institute of Standards and Technology (NIST), Special Publication 500-302, Washington, USA.

- E. M. Voorhees. On Building Fair and Reusable Test Collections using Bandit Techniques. In A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Broder, M. J. Zaki, S. Candan, A. Labrinidis, A. Schuster, and H. Wang, editors, *Proc. 27th International Conference on Information and Knowledge Management (CIKM 2018)*, pages 407–416. ACM Press, New York, USA, 2018.
- E. M. Voorhees and A. Ellis, editors. *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*, 2015. National Institute of Standards and Technology (NIST), Special Publication 500-308, Washington, USA.
- E. M. Voorhees and A. Ellis, editors. *The Twenty-Fifth Text REtrieval Conference Proceedings (TREC 2016)*, 2017. National Institute of Standards and Technology (NIST), Special Publication 500-321, Washington, USA.
- E. M. Voorhees and A. Ellis, editors. *The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018)*, 2019. National Institute of Standards and Technology (NIST), Special Publication 500-331, Washington, USA.
- E. M. Voorhees, I. Soboroff, and J. Lin. Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? *arXiv.org, Information Retrieval (cs.IR)*, arXiv:2201.11086, January 2022a.
- Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and Evaluation in Information Retrieval*, volume 63. Citeseer, 2005.
- Ellen M. Voorhees, Shahzad Rajput, and Ian Soboroff. Promoting repeatability through open runs. In *EVIA@NTCIR*. National Institute of Informatics (NII), 2016.
- Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1):1:1–1:12, 2020.
- Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. Too many relevants: Whither cranfield test collections? In *SIGIR*, pages 2970–2980. ACM, 2022b.
- W3C. RDF 1.2 Concepts and Abstract Syntax – W3C Working Draft 01 August 2024. <https://www.w3.org/TR/rdf12-concepts/>, August 2024a.
- W3C. RDF 1.2 Semantics – W3C Working Draft 27 June 2024. <https://www.w3.org/TR/rdf12-semantics/>, June 2024b.
- W3C. RDF Schema 1.2 – W3C Working Draft 18 April 2024. <https://www.w3.org/TR/rdf12-schema/>, April 2024c.
- Shuai Wang, Harris Scells, Ahmed Mourad, and Guido Zuccon. Seed-Driven Document Ranking for Systematic Reviews: A Reproducibility Study. In

- Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 686–700. Springer International Publishing, 2022a. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_46.
- Shuyi Wang, Shengyao Zhuang, and Guido Zuccon. Federated Online Learning to Rank with Evolution Strategies: A Reproducibility Study. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, volume 12657, pages 134–149. Springer International Publishing, 2021. ISBN 978-3-030-72239-5 978-3-030-72240-1. doi: 10.1007/978-3-030-72240-1_10. URL http://link.springer.com/10.1007/978-3-030-72240-1_10.
- Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. An inspection of the reproducibility and replicability of TCT-ColBERT. In *SIGIR*, pages 2790–2800. ACM, 2022b.
- R. L. Wasserstein and N. A. Lazar. The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, June 2016.
- W. Webber, A. Moffat, and J. Zobel. A Similarity Measure for Indefinite Rankings. *ACM Transactions on Information Systems (TOIS)*, 4(28):20:1–20:38, November 2010.
- R. W. White. *Interactions with Search Systems*. Cambridge University Press, Cambridge, UK, March 2016.
- R. W. White and C. Shah. *Information Access in the Era of Generative AI*. Springer Cham, Switzerland, 2025.
- M. D. Wilkinson, M. Dumontier, and et al. The FAIR Guiding Principles for scientific data management and stewardship. *Nature Scientific Data*, 3 (160018), 2016.
- Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. Relation-aware entity alignment for heterogeneous knowledge graphs. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5278–5284. ijcai.org, 2019. doi: 10.24963/IJCAI.2019/733.
- Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7): 100543, 2022. ISSN 2666-3899. doi: <https://doi.org/10.1016/j.patter.2022.100543>. URL <https://www.sciencedirect.com/science/article/pii/S2666389922001465>.
- Omry Yadan. Hydra - A framework for elegantly configuring complex applications. Github, <https://github.com/facebookresearch/hydra>, 2019.

- Eugene Yang and David D. Lewis. Tarexp: A python framework for technology-assisted review experiments. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3256–3261. ACM, 2022. doi: 10.1145/3477495.3531663. URL <https://doi.org/10.1145/3477495.3531663>.
- P. Yang and H. Fang. Evaluating the Effectiveness of Axiomatic Approaches in Web Track. In Voorhees [2014].
- P. Yang and J. Lin. Anserini at TREC 2018: CENTRE, common core, and news tracks. In Voorhees and Ellis [2019].
- Peilin Yang and Hui Fang. A reproducibility study of information retrieval models. In *ICTIR*, pages 77–86. ACM, 2016.
- Peilin Yang and Jimmy Lin. Reproducing and Generalizing Semantic Term Matching in Axiomatic Information Retrieval. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, volume 11437, pages 369–381. Springer International Publishing, 2019b. ISBN 978-3-030-15711-1 978-3-030-15712-8. doi: 10.1007/978-3-030-15712-8_24. URL http://link.springer.com/10.1007/978-3-030-15712-8_24.
- Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In *SIGIR*, pages 1253–1256. ACM, 2017.
- Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Reproducible ranking baselines using lucene. *ACM J. Data Inf. Qual.*, 10(4):16:1–16:20, 2018.
- Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. Critically examining the "Neural Hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *SIGIR*, pages 1129–1132. ACM, 2019.
- Andrew Yates, Siddhant Arora, Xinyu Zhang, Wei Yang, Kevin Martin Jose, and Jimmy Lin. Capreolus: A toolkit for end-to-end neural ad hoc retrieval. In James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang, editors, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 861–864. ACM, 2020. doi: 10.1145/3336191.3371868. URL <https://doi.org/10.1145/3336191.3371868>.
- Burak Yildiz, Hayley Hung, Jesse H. Krijthe, Cynthia C. S. Liem, Marco Loog, Gosia Migut, Frans A. Oliehoek, Annibale Panichella, Przemyslaw Pawelczak, Stjepan Picek, Mathijs de Weerdt, and Jan van Gemert. ReproducedPapers.org: Openly teaching and structuring machine learning reproducibility. In *RRPR*, volume 12636 of *Lecture Notes in Computer Science*, pages 3–11. Springer, 2021.

- E. Yilmaz and J. A. Aslam. Estimating Average Precision With Incomplete and Imperfect Judgments. In P. S. Yu, V. Tsotras, E. A. Fox, and C.-B. Liu, editors, *Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006)*, pages 102–111. ACM Press, New York, USA, 2006.
- E. Yilmaz, N. Craswell, B. Mitra, and D. Campos. On the Reliability of Test Collections for Evaluating Systems of Different Types. In Chang et al. [2020], pages 2101–2104.
- Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Daniel Campos. On the reliability of test collections for evaluating systems of different types. In *SIGIR*, pages 2101–2104. ACM, 2020b.
- Edwin Zhang, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. Chatty goose: A python framework for conversational search. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2521–2525. ACM, 2021. doi: 10.1145/3404835.3462782. URL <https://doi.org/10.1145/3404835.3462782>.
- W. X. Zhao, Y. Hou, X. Pan, C. Yang, Z. Zhang, Z. Lin, J. Zhang, S. Bian, J. Tang, W. Sun, Y. Chen, L. Xu, G. Zhang, Z. Tian, C. Tian, S. Mu, X. Fan, X. Chen, and J.-R. Wen. Recbole 2.0: Towards a more up-to-date recommendation library. In M. Al Hasan and L. Xiong, editors, *Proc. 31st International Conference on Information and Knowledge Management (CIKM 2022)*, pages 4722–4726. ACM Press, New York, USA, 2022.
- K. Zhou, Y. Zhu, Z. Chen, W. Chen, W. X. Zhao, X. Chen, Y. Lin, J.-R. Wen, and J. Han. Don’t Make Your LLM an Evaluation Benchmark Cheater. *arXiv.org, Computation and Language (cs.CL)*, arXiv:2311.01964, November 2023.
- J. Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments. In Croft et al. [1998], pages 307–314.
- J. Zobel, W. Webber, M. Sanderson, and A. Moffat. Principles for Robust Evaluation Infrastructure. In M. Agosti, N. Ferro, and C. Thanos, editors, *Proc. Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation (DESIRE 2011)*, pages 3–6. ACM Press, New York, USA, 2011.
- Guido Zuccon, Harris Scells, and Shengyao Zhuang. Beyond CO2 emissions: The overlooked impact of water consumption of information retrieval models. In Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi, editors, *Proceedings of the 2023 ACM SIGIR International Conference on Theory of*

Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023, pages 283–289. ACM, 2023. doi: 10.1145/3578337.3605121. URL <https://doi.org/10.1145/3578337.3605121>.

Jan Štihec, Martin Žnidaršič, and Senja Pollak. Simplified Hybrid Approach for Detection of Semantic Orientations in Economic Texts. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 692–698. Springer International Publishing, 2018. ISBN 978-3-319-76941-7. doi: 10.1007/978-3-319-76941-7_64.