Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science"

Nicola Ferro[1] **Norbert Fuhr**[2] **Kalervo Järvelin**[3]
Noriko Kando[4] **Matthias Lippold**[2] **Justin Zobel**[5]

[1] University of Padua, Italy, ferro@dei.unipd.it
[2] University of Duisburg-Essen, Germany, {norbert.fuhr, matthias.lippold}@uni-due.de
[3] University of Tampere, Finland, kalervo.jarvelin@staff.uta.fi
[4] National Institute of Informatics, Japan, kando@nii.ac.jp
[5] University of Melbourne, Australia, jzobel@unimelb.edu.au

Abstract

The Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science", held on 24-29 January 2016, focused on the core issues and approaches to reproducibility of experiments from a multidisciplinary point of view, sharing the experience coming from several fields of computer science.

In this paper, we discuss, summarize, and adapt the main findings of the seminar to the context of IR evaluation – both system-oriented and user-oriented – in order to raise awareness in our community and stimulate the fields towards and increased reproducibility of our experiments.

1 Introduction

Reproducibility has always been a primary concern of science since the experimental method has been established by Galileo Galilei many centuries ago [12], and it is both enabled and emphasised in this era of data-driven science [16].

Information Retrieval (IR) is a discipline strongly rooted in experimentation and has always strived for repeatability of experiments by relying on the common Cranfield paradigm [6], developing shared experimental collections, and running coordinated experimental activities. This has led to successful large-scale evaluation initiatives, such as such as the *Text REtrieval Conference (TREC)*[1] [15] in the United States since 1992, the *NII Testbeds and Community for*

[1] http://trec.nist.gov/

Information access Research (NTCIR)[2] in Japan and Asia since 1999, the *Conference and Labs of the Evaluation Forum (CLEF)*[3] [8] in Europe since 2000, and the *Forum for Information Retrieval Evaluation (FIRE)*[4] in India since 2008.

Nevertheless, reproducibility is a complex concept and the IR community has started only recently to discuss some novel aspects, such as for example infrastructures for managing experimental data [2], off-the-shelf open source IR systems [26], use of private data in evaluation [5], evaluation as a service [14, 17], reproducible baselines [3], as well as considering it as part of the review process of major conferences and in dedicated tracks, such as the new ECIR Reproducibility Track [9, 11].

The Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science"[5] [10] was held in Dagstuhl, Germany, from 24 to 29 January 2016. The overall goal of the seminar was to bring together researchers with a multidisciplinary set of competencies in different fields of computer science – databases, information retrieval, information visualization, semantics, bioinformatics, human-computer interaction, simulation, and more – in order to share experiences, issues with, and approaches to reproducibility across different fields in order to come to a common ground across disciplines, leverage best-of-field approaches, and provide a unifying vision on reproducibility.

In this paper, we distill and apply the main outcomes of the seminar to the specific case of IR evaluation in order to contribute and stimulate the ongoing discussion in the field about this relevant topic. In particular, Section 2 provides an overview of the *Platform, Research goal, Implementation, Method, Actor, and Data (PRIMAD)* model, which has been developed during the seminar to tackle with the different angles of reproducibility; Section 3 discusses reproducibility in the context of system-oriented evaluation, the application of the PRIMAD model to this case, barriers to and actions for improving reproducibility; Section 4 examines reproducibility in the user-oriented case, along the lines of the previous section; finally, we summarise in Section 5.

2 Overview of the PRIMAD Model

As there are many different terms relating to various kinds of reproducibility [7], the *Platform, Research goal, Implementation, Method, Actor, and Data (PRIMAD)* (pronounce "primed") model, developed during the Dagstuhl Seminar [10], acts as a framework to distinguish the major components describing an experiment in computer science (and related fields):

Research Goal characterizes the purpose of a study;

Method is the specific approach proposed or considered by the researcher;

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

Platform describes the underlying hard- and software like the operating system and the computer used;

[2] http://research.nii.ac.jp/ntcir/

[3] http://www.clef-initiative.eu/

[4] http://www.isical.ac.in/~fire/

[5] http://www.dagstuhl.de/no_cache/en/program/calendar/semhp/?semnr=16041

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

Actor refers to the experimenter

As an example, consider a student performing a retrieval experiment. The research goal is to achieve a high retrieval quality, and as method chosen is the BM25 formula. Experiments use the LEMUR system as implementation, under the operating system Ubuntu 15.10 on a Dell xyz server. The GOV2 collection serves as input data, and a specific setting of the BM25 parameters is chosen. The actor is the student performing the runs.

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 use these letters to characterize specific forms of reproducibility.

When another researcher now tries to reproduce the experiment described above, she will change one or more of the components. In case she tries to rerun the experiment without changing anything else[6], then we have another actor, that is, A is changed to A′, the actor is "primed". If successful, this experiment would demonstrate that the original researcher has supplied enough information to ensure reproducibility. In case the results of the experiment are the same, then the original findings have been successfully reproduced and thus confirmed.

Now let us look at changes of the other components:

$\mathbf{R} \rightarrow \mathbf{R}'$: 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 and implementation usually are also changed, and other components as well.

$\mathbf{M} \rightarrow \mathbf{M}'$: Most of the research in the field of IR deals with the investigation of alternative methods (retrieval models, formulas). This implies also a new implementation I′, which often runs on a different platform. However, for performing comparisons, the (input) data should be the same.

$\mathbf{I} \rightarrow \mathbf{I}'$: Here a researcher uses a different implementation, say Terrier instead of Lemur, or does their own reimplementation.

$\mathbf{P} \rightarrow \mathbf{P}'$: 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.

$\mathbf{D} \rightarrow \mathbf{D}'$: 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.

For ensuring reproducibility, there is the need to be able to share as many PRIMAD components as possible. Research goal and method are what we currently share via publications in conference proceedings or journals (although details of the method are often missing). Sharing

[6]Actually, this would be difficult to achieve

implementations are possible via making it open source and uploading it on Web sites focusing on this task (for example, Github). Platforms can be shared by means of virtual machines or dockers, or by "evaluation as a service". For the input data, there are a number of standard test collections which are generally available. When researchers use their own test collection, however, reproducibility can only be ensured in case this collection is shared with the community, ideally via a trustworthy repository.

Finally, there are two other important aspects that are not part of the core PRIMAD model:

Transparency is the ability to look into all necessary components to verify that the experiment does what it claims; for example, sharing a virtual machine, but not the source code of an experiment, would not satisfy this criterion.

Consistency refers to the success or failure of a reproducibility experiment in terms of consistent outcomes; for example, using a random number generator for breaking ties in a ranking would lead to problems with respect to this criterion.

3 System-oriented Evaluation and Reproducibility

3.1 Methodological Background

System-oriented evaluation is concerned with the ability of a retrieval system to find answers in a test collection.

Evaluation of retrieval systems based on robustly built test collections has been a cornerstone of research in information retrieval since the 1960s, beginning with the highly influential Cranfield collections of 1963–1967 [25]. These early collections established the core principles for system-oriented test data in the field. In these principles, the document collection is static and represented in a uniform way; there is a fixed set of queries; and the relevance of each document to each query is known. In each query-document pair relevance is established by a human assessor.

Such a collection is expected to provide a basis for highly reproducible system-oriented research. A run with a given system on the data should produce identical results, and more generally if the querying methodology (or similarity function) is sufficiently well described then a fresh implementation should likewise be identical. As the collections are often shared, or publicly available, they are an exemplar of a particular kind of reproducibility.

In some respects such reproducibility is not particularly interesting: a new experiment validates the original runs, or demonstrates that a new tool has been correctly implemented, but does not confirm that the same behaviour would be observed on new data, that is, reproduction on the same data provides no information on the generality of the result. However, the format of such test collections is in general straightforward and reasonably standard, with the effect that a system can be validated on one collection and then applied to another without alteration. This portability of outcomes means that systems can often be compared to each other in diverse contexts, allowing, for example, sensitivity analysis of the effect of different system components.

System runs on a test collection provide no information about the behaviour of users when presented with answers, the likelihood that an answer page will trigger reformulation of the query, the need to present results in an appealing form, and so on. The purpose of the experiment is

purely to understand how alterations in the retrieval mechanism influence the quality of what is retrieved.

As test collections grew, and exhaustive relevance assessment became impractical, variants to these principles were adopted for new test collections. The TREC evaluation campaigns [15] were particularly influential. TREC saw the introduction of gigabyte-scale collections with millions of documents; to provide some meaningful volume of relevance assessment, sampling of the collections was required.

A true random sample could easily contain no relevant documents (which in the TREC queries occur at a rate of around one in ten thousand), and thus biased sampling is required. The method used in TREC was to pool the results of a diverse search engines and assess only the document-query pairs in the pool. The use of this methodology was effectively an artifact of the rationale for TREC, namely the comparison of a variety of search tools, but proved highly effective as a mechanism for constructing enduring test collections. Other methodologies for sampling have also been used (such as use of human searchers on a reference retrieval system to quickly find search candidates), but the TREC pooling methodology has remained dominant.

In the TREC approach, the set of contributing systems are run in a blind manner, where the researchers are unaware of which documents may be relevant. In this way a TREC evaluation campaign should provide a reasonably unbiased comparison amongst these systems.

Other evaluation campaigns in information retrieval, notably NTCIR, INEX, and CLEF, have adopted similar methodologies. Like TREC, these have proven to be not only a productive mechanism for creating large shared resources, but have also been an impetus to creation of shared software and to refinement of evaluation methodologies.

3.2 PRIMAD for System-oriented Evaluation

In system-oriented evaluation, many of the experimental settings are determined by the structure of the test collection and in particular of the relevance assessments. For example, if the assessments are binary then investigation of 'best' answers may not be feasible.

- ***R**esearch goal*, the question to be addressed, is in broad terms to produce a high-quality ranking of answers, 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.

- ***M**ethod* 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).

- ***I**mplementation* and ***P**latform* is 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 natural-language processing tool; external resources such as dictionaries or reference collections; and the tools used post-retrieval for computing effectiveness from the document rankings.

- ***A****ctor*, 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.

- ***D****ata*, 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.

3.3 Barriers and Obstacles to Reproducibility

A shared test collection provides a strong basis for reproducible research. Coupled with a public-domain implementation of the method, it allows in a straightforward and well-understood way for work to be reproduced.

However, there are practical barriers to full reproducibility. Retrieval systems tend to be, not just complex pieces of software, but dependent on a range of other resources in the laboratory environment such as dictionaries or machine learning tools. It is not always possible to encapsulate the complete environment of the experiment, and behaviour of the system will change as the environment changes.

Where an external or commercial tool has been used for statistical interpretation, changes in the tool (or lack of later access to it) may mean that numerical summaries of the results are inconsistent with each other, even though the underlying rankings are identical.

Some test collections are private, which is a serious barrier to reproducibility. More commonly, where document sets are (more or less) public, they are associated with query logs that are proprietary; and commercial search engines have sound legal reasons for having to keep such data private. Simply, confidentiality means that some query logs cannot be shared.

The same consideration applies to some specialized collections. For example, medical records contain extensive textual content, and are an obvious target of retrieval experimentation. But they are highly sensitive, and not easily anonymised.

The size of realistic collections is another obstacle, with large web collections requiring significant storage and computational resources. Pragmatically, simply obtaining a copy of a large collection can be impractical because of the need to deliver it via a collection of hard disks, involving significant human and logistical effort.

Perhaps surprisingly, although there are numerous substantial test collections that have been made available or shared, there are few archives of previous experimental runs. That is, the outcomes of experiments are generally kept private, with the notable exception of the runs submitted to TREC (and similar evaluation campaigns) as part of the pooling process.

The test collections can be regarded as reasonably robust, but they rest on assumptions about users, queries, and relevance, and these assumptions are often hidden. In particular, the meaning of relevance is not always clear! The shared collections often contain good numbers of documents where fresh assessment of relevance is inconsistent with the original assessment, and for this reason

methods that are tuned to run on these collections may not always perform as well in new contexts. That is, the results can be repeatable but not reproducible.

Similar problems arise with the queries, which may be not representative of realistic search needs, or which may simply be too few to allow observation of generalizable results.

3.4 Actions to Improve Reproducibility

A shared test collection provides a strong basis for repeatability and reproducibility not available in other contexts. The standardization of the format of these test collections further adds to reproducibility, as it means that public domain retrieval systems are typically able to process a collection from a straightforward uncustomized installation.

However, many collections do not have sufficient queries to provide strong experimental power. That is, there may not be enough queries to observe real effects even when they are present. Having insufficient numbers of queries is also an obstacle to re-use within a series of experiments.

More serious obstacles arise with retrieval systems. Full repeatability is often unrealistic, because of system dependence on resources that cannot be encapsulated with the system, but reproducibility is a more achievable goal. If resources are fully characterized, and standard installations closely approximate the experimental settings, then a researcher undertaking a fresh experiment should be able to attribute changes in behaviour of the system to clear causes.

A straightforward step that should be taken is creation of a full archive of published experimental results. If the rankings that are reported in a paper are publicly available then a meaningful comparison of new results can be undertake. Such archives were developed – EvaluatIR [4] and DIRECT [1, 24] – but saw little or moderate use.

A richer step would be to require experimental materials as part of the paper submission process. For example, EvaluatIR produced a standardized "results" sheet showing outcomes across a range of metrics against previous runs on the same data. By providing a common measurement framework, a reviewer could readily assess performance against earlier benchmarks.

Another approach is for the community to make full use of the cloud. The cost of maintaining key public collections in the cloud would not be great, and could be coupled with installations of reference systems and statistical tools. Institutions and researchers would then have the option of downloading materials or running experiments remotely, probably at lower amortized cost than under current practice. A step in this direction is the TIRA [13] platform, a Web service that supports organizers of shared tasks in computer science to accept the submission of executable software.

4 User-oriented Evaluation and Reproducibility

4.1 Methodological Background: Experiments in Psychology

The knowledge acquired in psychology is based on empirical results from experiments. An experiment is a research method in which one or more independent variables are manipulated to determine the effects on a dependent variable. Other relevant factors besides the independent variable need to be controlled for. For instance, in the case of a user experiment in information retrieval, as the independent variable different user-interfaces could be implemented and as

a dependent variable the time to finish a specific search could be measured. A confounding variable could be the speed of the internet connection, which should be held constant between in all different interface conditions.

Psychological experiments needs to fulfill three criteria: *validity*, *objectivity* and *reliability*.

Validity First of all an experiment needs to be valid. Validity is reached when the measurements used are reflecting the constructs which they are claimed to measure.

Objectivity Secondly, an experiment needs to be objective. It has to be objective in two ways, the result of the experiment should not be influenced by the experimenter and the interpretation of the data should also not depend on the examiner.

Reliability Furthermore, an experiment has to be reliable. When the same experiment is repeated by the same person or another researcher, a similar result should be obtained. To ensure reliability, scientists have to specify their experimental design, they have to describe the conditions, under which the experiment is conducted and share information about the participants. The material and the raw data of the experiments needs to be stored and shared on demand by the corresponding author.

Reproducibility crisis in psychology

In a recent study [21] the results from 100 experiments from four top journals could just be partially replicated. That started a big discussion about the reasons.

Theoretical reason for failed replication can be found in the selection of the theoretical background, particular in the theory specification. An *ill-defined theory*, which fails to specify the outcome of the following experiment can be a threat to reproducibility. When the acquired results of the experiment are interpreted as evidence for the unspecified theory, the result will most likely be not reproducibly, because the original result are just randomly acquired. Therefore, theories need to be specific [22]. Likewise, it is important to state whether the obtained result can be generalized to different populations or are just relevant for a subpopulation in a specific domain. Concerning IR experiments it might be useful to consider that different age groups could use search engines in different ways than students do, who usually used as participants in studies.

Another theoretical threat are post theories and *post hypotheses* or *post predictions*. When the hypotheses and the theoretical background are selected after the result of the experiment is known, the probabilities of the attained p-values have to be interpreted differently. The describe procedure is a form of $p - hacking$, which is unfortunately quite often applied [18].

There are also some methodological problems leading to the failure of replications. Researchers rely almost exclusively on the p-values and do not consider *effect sizes* [23]. But p-values are dependent on sample size and even a small irrelevant effect can have a significant p-value. IR should switch from asking whether there is a difference at all to investigating how big is the difference and whether the user would actually notice such difference. On the other hand, too small sample sizes can also lead to problems, because they lead to low *statistical power*, and these results might also fail to be replicated [20].

4.2 User-oriented IR Evaluation

In IR, we have different kinds of user studies:

- *laboratory experiments*, where users are observed in the lab
- *in situ observation* of users at their workplace
- *living labs*, where the researcher analyses the system logs and possibly also manipulates the system employed by the users for their daily work.

Besides these types of experiments, there are studies that focus mainly on data collection methods, for which the discussion below only partially applies:

- exploratory user studies,
- focus groups, where researchers interview users
- longitudinal studies of users.

4.3 PRIMAD for User-oriented Evaluation

For discussing the reproducibility issues for the specific case of user studies, we follow the PRIMAD model described above:

- ***R****esearch 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.
- ***M****ethod* relates here to the experimental settings, which are used for testing the hypotheses specified before. So, besides the type of study, also the relevant aspects of the settings that refer to the research objectives are part of the model.
- ***I****mplementation* and ***P****latform* 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.
- ***A****ctor* 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.
- ***D****ata* 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 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))

For enabling reproducibility, a researcher should share this context with other users to the maximum extent possible. Research goal and method are usually described in the research paper. In the past, the main research objective was the effectiveness of the methods investigated. Nowadays, also other aspects are considered, which are either more closely related to the actual user task, or to more subjective factors such as user satisfaction or engagement (which, in turn, can be measured via different variables). The more factors are considered, the more important it becomes to state the research hypotheses before actually carrying out the study, in order to achieve statistically valid results.

The environment usually can only be shared partially (mainly the system), while most other aspects (e.g. the users, the hardware, etc.) should be described at a reasonable level of detail in order to ease reproducibility. The same holds for the actor.

For the data, sharing testbeds is widely accepted nowadays, since the state of the art does not allow yet to characterize testbeds to such an extent that an independent researcher would be able to create a comparable testbed that could be expected to give similar results. The observation data, on the other hand, are essential for verifying the claims of a research paper. To a limited extent, it also can be used for simulation studies, depending on the degree of interactivity involved in the study; in classical IR experiments, the only data of this kind are relevance judgments.

4.4 Barriers and Obstacles to Reproducibility

Since the Cranfield project in the early 1960 [6], researchers constructed and shared testbeds (test collections) consisting of the three types of data, namely document collections, a set of search requests, and a static set of the human relevance judgments. However, technology and society have evolved tremendously: interactive online search for various purposes by ordinary persons has become pervasive in everyday life, the various data collections including various web services and the social media were enhanced, search on multiple devices for multi-tasking has become more common. The traditional evaluation paradigm (based on the batch-style one-time judgments) can not cover all the problems in the IR research and we are facing various new challenges and obstacles to make the research reproducible:

For studying users in interactive IR, there are various barriers and obstacles for reproducibility:

- privacy or limitations of anonymization
- confidentiality,
- volatility of data (live streams, when the same situation never happens again, ...),
- validity of the data: the data are so multidimensional that it is difficult to ensure the external validity of the experiment,
  - this complexity is present also in IR test collections,
  - even more if we consider dynamic test collections.

Production search systems are typically based on algorithms which exploit user behaviour data in some way to personalize and fit the system output to user needs and context; the situation is quite similar also in the case of interactive research IR systems. Unfortunately, this data is

intrinsically rich in privacy and often includes confidential information. Although various research efforts have targeted anonymization, there are still limitations which make it difficult to release user behaviour data for external research groups and, in turn, this hampers reproducibility.

Large-scale users logs are generated in commercial search services and substantial studies on modeling and predicting users behaviour have been conducted based on these data but, again, the underlying data are not accessible for other researchers. Not only user modeling studies but also various operational search mechanisms exploit user behavior data in search and ranking, thus making it difficult to reproduce these methods. To tackle information privacy and/or copyright problems, various Evaluation-as-a-Service approaches have been proposed, i.e. the possibility of running the algorithms on the data on some infrastructure without actually accessing them. Even if some of them were implemented successfully, they are still not sufficient for all the data produced by the users in-situ and lab environments.

For volatility, IR experiments can be conducted on live streaming data or commercial search services, in which the data and algorithms are continuously changing and the same data will never be obtained again. Also, user experiments can not be "re-run" with the same users as the users learn from the previous experience.

In IR, interactivity and user behaviour or search experience through whole search sessions (or sometimes even a task involving multiple sessions) are becoming more important in order to consider real-world contexts and various algorithms and softwares to support such interactions have been studied and proposed. The data obtained from the users in such task-based or whole session-based studies are highly complex, comprising e.g. the nature of the tasks conducted as well as the characteristics of each user. More research is needed for developing a framework that is able to describe such complex, multi-dimensional data as well as for devising methods for proper scientific analysis of such data and ensure their reproducibility.

4.5 Actions to Improve Reproducibility

Actions to improve reproducibility of user-oriented experiments include checklists for authors (and reviewers, editors, chairs, ...), sample exemplary papers, method inventories, extended methodological sections in papers, and critical discussions on the components/tools/other data used. These are considered briefly below.

Checklists should be provided on the methodology applied in the study. Kelly's review [19] on interactive IR research methods is a useful source for a checklist for user-oriented IR studies. Examples of items to check and describe are:

- research questions and variables used in the experiment - care is required since the study designs are not as standard as traditional IR study designs;
- experimental design: factors of factorial design neutralizing variation and learning effects (e.g., latin square design, which systematically rotates the order in which conditions like IR techniques and tasks are encountered);
- participant characteristics and the population they are claimed to represent;

- methods of data collection, including the experimental protocol and control of participant fatigue, environmental conditions, and variables used in the study (guidelines for how to define and how to measure);
- the experimenter;
- retrieval systems, their interfaces, and baseline results;
- methods for data analysis, including assumptions of statistical analyses (and adjustments if assumptions are not met);
- degree of control on the system by the experimenter (black box?).

Exemplary papers representing various types of user experiments could be offered in some community-based repository and annotated for their critical features, see also next section.

Inventories of typical variables in various types of user experiments and standard ways to operationalize and measure them in different (sample) study settings could be provided by the community, as further discussed below.

Methodological sections could be emphasized in document templates, author guidelines and review guidelines. More space might be allocated to these sections and / or authors encouraged to provide methodological appendixes or tech reports.

Finally, authors could be encouraged to *critically discuss* how suitable the set of tools and collections is to answer the research questions, what claims can the tools/collection support, describing the generalizability of the findings on the basis of the tools/collection that have been used.

4.6 Community Support to Reproducibility

In order to embody the vision described above and foster reproducibility in user-oriented studies, the involvement of the research community is crucial and it should consist of two complementary actions:

- support to the creation of *shared resources*;
- taking up and implementation of *shared practices*.

When it comes to shared resources, we can foresee several examples of them:

- *inventories*: in order to streamline the reproducibility process, there is a need for catalogues accounting for the most appropriate experimental designs, the kind of independent and dependent variables one typically encounters in these settings, how to describe and measure such variables, the proper data analysis methodologies and statistical validation methods to apply to these variables in the different experimental designs, and so on;

- *do's and don'ts*: in order to facilitate the understanding and adoption of the above facilitators of reproducibility, real and hands-on examples of appropriate and inappropriate ways to carry out user-oriented experiments are needed to clearly explain why a seemingly appropriate experimental setup is or is not working as expected. This could be partnered with a selection of exemplary and well-known papers, which should be annotated and enriched with links and explanations related to the above inventories, in order to clarify to the researcher how and when to apply a given approach by means of concrete and remarkable case studies;

- *repositories*: the adoption of shared repositories to gather collections of documents, interaction data, tasks and topics, and more is a key step to extend the reach of reproducibility in user-oriented experimentation;

- *data formats*: the development of commonly understood and well-documented data formats, which can be extended to specific needs, as well as the introduction of proper metadata (descriptive, administrative, copyright, ...) to model, describe, and annotate the data and the experimental outcomes is a crucial factor in lowering the barriers to reproducibility in user-oriented experimentation.

The methodological instruments, the checklists, the critical discussions, the different kinds of shared resources previously described are all key "ingredients" for successfully reproducing user-oriented experiments but the actual catalyst is the systematic and wide adoption by the community of shared practices, effectively exploiting all of these "ingredients".

5 Conclusions

The Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science" investigated the problems and approaches to reproducibility in various sub-fields of computer science, comprising databases, information retrieval, information visualization, semantics, bioinformatics, human-computer interaction, simulation, and more, in order to gain a common understanding and shared solutions.

Reproducibility is an open challenge for data-driven science in general and for Information Retrieval in particular. IR is a discipline rooted not only in computer science but also in information science, statistics, computational linguistics, and other fields, making reproducibility in IR evaluation especially compelling. Indeed, the spectrum of IR evaluation ranges from system-oriented to user-oriented evaluation and thus involves all of the issues we typically find in the other different fields alone.

We discussed the outcomes of the Dagstuhl Seminar putting them in the specific context of IR evaluation. In particular, we introduced the *Platform, Research goal, Implementation, Method, Actor, and Data (PRIMAD)* model which allows us to systematically described the major components involved in an experiment and, consequently, to clearly understand how they affect reproducibility. Then, we discussed how to fit PRIMAD in both system-oriented and user-oriented evaluation, highlighting obstacles to reproducibility as well as actions to improve it.

This paper represents a step in the direction of increasing reproducibility in IR evaluation and it contributes to the ongoing discussions in the field on the many different angles of reproducibility. However, much is still ahead of us in order to guarantee full reproducibility of our experiments:

we need to develop a shared culture for reproducibility, adopt common experimental protocols, introduce practices to properly manage and curate our experimental data, make all of this part of our review and publication modalities, and much more.

Acknowledgement

This work was supported by Deutsche Forschungsgemeinschaft (DFG) under grant no 2167, Research Training Group "User-Centred Social Media".

References

[1] M. Agosti and N. Ferro. Towards an Evaluation Infrastructure for DL Performance Evaluation. In G. Tsakonas and C. Papatheodorou, editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, pages 93–120. Chandos Publishing, Oxford, UK, 2009.

[2] M. Agosti, N. Ferro, and C. Thanos. DESIRE 2011 Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation. *SIGIR Forum*, 46(1):51–55, June 2012.

[3] 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 2015.

[4] 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, 2009.

[5] J. Callan and A. Moffat. Panel on Use of Proprietary Data. *SIGIR Forum*, 46(2):10–18, December 2012.

[6] C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In Spärck Jones and Willett [25], pages 47–60.

[7] D. De Roure. The future of scholarly communications. *Insights*, 27(3):233–238, November 2014.

[8] N. Ferro. CLEF 15th Birthday: Past, Present, and Future. *SIGIR Forum*, 48(2):31–55, December 2014.

[9] N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello, editors. *Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016)*, 2016. Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany.

[10] J. Freire, N. Fuhr, and A. Rauber. Reproducibility of data-oriented experiments in e-science. *Dagstuhl Reports*, 6(1), 2016.

[11] N. Fuhr, A. Rauber, G. Kazai, and A. Hanbury, editors. *Advances in Information Retrieval. Proc. 37th European Conference on IR Research (ECIR 2015)*, 2015. Lecture Notes in Computer Science (LNCS) 9022, Springer, Heidelberg, Germany.

[12] G. Galilei. *Dialogues Concerning Two New Sciences*. Dover Publications, USA, 1954.

[13] T. Gollub, B. Stein, and S. Burrows. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In W. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 1125–1126. ACM Press, New York, USA, 2012.

[14] A. Hanbury, H. Müller, K. Balog, T. Brodt, G. V. Cormack, I. Eggel, T. Gollub, F. Hopfgartner, J. Kalpathy-Cramer, N. Kando, A. Krithara, J. Lin, S. Mercer, and M. Potthast. Evaluation-as-a-Service: Overview and Outlook. *CoRR*, abs/1512.07454, 2015.

[15] D. K. Harman and E. M. Voorhees, editors. *TREC. Experiment and Evaluation in Information Retrieval*, 2005. MIT Press, Cambridge (MA), USA.

[16] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, USA, 2009.

[17] 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.

[18] L. K. John, G. Loewenstein, and D. Prelec. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5):524–532, May 2012.

[19] D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval (FnTIR)*, 3(1–2):1–224, 2009.

[20] S. Maxwell and H. D. Delaney. *Designing Experiments and Analyzing Data. A Model Comparison Perspective*. Lawrence Erlbaum Associates, Mahwah (NJ), USA, 2nd edition, 2004.

[21] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):943–952, August 2015.

[22] S. Roberts and H. Pashler. How Persuasive Is a Good Fit? A Comment on Theory Testing. *Psychological Review*, 107(2):358–367, April 2000.

[23] T. Sakai. Statistical Reform in Information Retrieval? *SIGIR Forum*, 48(1):3–12, June 2014.

[24] 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)*, 2016.

[25] K. Spärck Jones and P. Willett, editors. *Readings in Information Retrieval*, 1997. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

[26] A. Trotman, C. L. A. Clarke, I. Ounis, J. S. Culpepper, M.-A. Cartright, and S. Geva. Open Source Information Retrieval: a Report on the SIGIR 2012 Workshop. *ACM SIGIR Forum*, 46(2):95–101, December 2012.