Report on the Workshop on Barriers to Interactive IR Resources Re-use (BIIRRR 2018)

Toine Bogers[1] **Maria Gäde**[2] **Mark Michael Hall**[3]
Luanne Freund[4] **Marijn Koolen**[5] **Vivien Petras**[2]
Mette Skov[6]

[1] Aalborg University Copenhagen, Denmark
[2] Humboldt-Universität zu Berlin, Germany
[3] Martin Luther University, Halle-Wittenberg, Germany
[4] University of British Columbia, Vancouver, Canada
[5] Huygens ING/KNAW, Netherlands
[6] Aalborg University, Denmark

1 Introduction

The goal of the BIIRRR 2018 workshop [2] was to serve as a starting point for a community-driven effort to design and implement a platform for the collection, organization, maintenance, and sharing of resources for interactive information retrieval (IIR) experimentation.

As in all scientific endeavors, progress in IIR research is contingent on the ability to build on previous ideas, approaches, and resources. Current trends towards open science and funding mandates to preserve and share research data lend support and even urgency to the notion of establishing a shared disciplinary repository of research tools and data for IIR. The need for an IIR (evaluation) framework was further highlighted by Pia Borlund in her 2016 CHIIR keynote [3]. Components from IIR experiments that could be valuable to archive for re-use include: the systems or platforms used for experimentation, the content or resources of the experimental platform (data collections), the search tasks or work situation, the experimental context and other important aspects of the test design, experimental protocols, questionnaire designs, etc., the gathered user and system interaction data, the tools used for analysis as well as the results and measures.

We believe there to be a number of barriers to reproducibility and re-use of resources in IIR research: the fragmentary nature of how the community's resources are organized, the lack of awareness of their existence, insufficient documentation and organization of the resources, the nature of the typical research publication cycle, and the effort required to make such resources available.

The TREC initiative[1] highlights the value of such a repository, as it provides a single access

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

point for the system-based evaluation of many different IR tasks, offering a repository for test collections, topic sets and relevance judgments. The TREC Interactive Track, which ran from 1997 to 2002 [8], contributed to the standardization of protocols for experimental search studies, and involved the use of shared tasks and systems. However, due to the large differences between system-based and user-based evaluation, the fit between IIR tracks and TREC has not been an overly successful one with the TREC repository containing limited IIR data.

As a result of fragmentation, IIR resources are often underutilized due to a lack of awareness of their existence or proper documentation, leading to them falling in disuse. Apart from being publicized in research talks and linked to in publications (with often less-than-persistent links), no dedicated promotion channel or platform exists for these type of resources.

Fundamentally we believe that policy change linked to what level of experiment detail must be disclosed upon publication is needed to fully drive the move towards removing the barriers for re-use. At the same time an online platform dedicated to the collection and organization of IIR resources must be in place so that any change in policy does not lead to several incompatible platforms. This single access point, henceforth referred to as the iRepository, could be used to collect, manage and enable and promote the re-use of a variety of components of IIR experiments. The means by which components such as search tasks, experimental protocols, questionnaire designs, reporting standards, evaluation procedures, data collections, and the search interaction data produced in such experiments could be archived and made accessible for re-use is an unsolved challenge, which needs to consider data modeling and description as well as rights, data security and privacy issues.

While the idea of collecting such resources in a central location is perhaps not a new one, we are aware that the effort required in designing, implementing, and maintaining such a platform can only be borne by the community as a collective effort. The BIIRRR 2018 workshop therefore aimed to serve both as a brainstorming opportunity about the shape this iRepository should take, as well as a way of building support in the community for its implementation.

2 Workshop Topics

The BIIRRR 2018 workshop[2] was an interactive, full-day workshop with interspersed keynotes on the challenges for IIR standardization and previous attempts and experiences with IIR evaluation campaigns including discussions on these topics. To take stock of current efforts, resources, and interest, the workshop also featured a pre-workshop activity. A survey of potential participants and interested IIR researchers gathered their views of, experiences with, and requirements for breaking down the barriers to re-use of IIR resources. This was presented and discussed during the workshop to represent even more voices on IIR re-use.

2.1 Initial Efforts: INEX, TREC Interactive and RepAST

The workshop started with two invited talks from Nils Pharo (Oslo Metropolitan University) and Luanne Freund (University of British Columbia) who shared their experiences with standardization efforts in IIR.

[2] `http://biirrr2018.aau.dk`

In his keynote on "The importance and challenges for standardization in IIR Evaluation - Basis for an iRepository", Nils Pharo discussed the INEX interactive track [7], which ran from 2004-2010. The interactive track at INEX studied user interactions based on tasks, which were loosely associated with the system-centered INEX tracks in order to compare user and system perspectives. The track aggregated task-based user interaction data on a particular information retrieval system, which were collected by different research groups and then shared the data points across the participating groups.

While corpora (from IEEE research articles to Wikipedia entries to Amazon/LibraryThing book descriptions), information retrieval systems and research groups and participants changed over the years, some overlap was maintained across years. The INEX interactive track demonstrated that large-scale data gathering of interactive data was possible across different research institutions, employing standardized tasks and data collection protocols. Rich background data on users could be collected and compared across institutions and years.

The grand challenge for IIR evaluation, however, is to design realistic experiments: how can the study design and protocol reflect how, in which situations, and on which platforms people search? Task creation for studying user interactions in an IR system is a major aspect of designing realistic IIR experiments. Tasks are not only required to be realistic and fitting to the collection, but also - especially in a collaborative context such as the INEX interactive track - need to be relatable and interesting to a large number of participants on a global, cross-institutional scale.

Equally important is the data collection method, which influences how the data can be analyzed. Transaction logging is easiest to standardize across institutions, however for flexible data analysis, there needs to be agreement which interactions and data points are documented.

The INEX organizers invested a lot of time and effort into designing realistic and effective tasks and standardized data collection protocols, however, after the track was ended, neither the tasks, protocols or the collected data (150-300 individual user sessions per year) were rarely taken up in other studies even though more detailed analyses would have been possible and beneficial. The lack of sustainable preservation of the data and protocols prevented researchers, who were not affiliated with the original experiments, from getting access. Over time, the realism of the tasks and data collection diminishes, reducing the value of the data for more recent studies. However, both task structures and data collection protocols can provide valuable inspiration for other studies.

The second keynote presentation was entitled "Experiences with the Repository for Assigned Search Tasks (RepAST)". In this talk, Luanne Freund began by reviewing some of the early collaborative work in IIR carried out through the TREC Interactive Track[5] (1994 to 2002). She noted that the Interactive Track was influential in helping to establish a standardized protocol for running experimental user studies in IR. The track adopted different collaborative approaches from one year to the next, generally sharing tasks and instruments and sometimes using a common system, with user data collected separately by each of the teams.

One of the major limitations of the track is that it did not succeed in the goal of enabling sharing of research data across the participant teams. Many of the reasons for this, as noted by Freund, continue to act as barriers to IIR research data sharing today, including the lack of shared infrastructure, ethical concerns relating to participant privacy, and proprietary research interests. One of the outcomes of involvement in TREC in 2001 and 2002 was the development of a blueprint

for a web-based IR experimentation system, known as WIiRE[9], which was a precursor to later digital platforms for managing IR studies.

The remainder of the presentation focused on experiences with RepAST[3], an online platform for the analysis and sharing of tasks used in IIR studies[4]. RepAST was created as a means to study the use of assigned search tasks by IIR researchers, to encourage greater conceptual clarity and rigour in task-based studies, and to encourage re-use of study protocols and tasks. RepAST is the result of manual analysis of over 800 research papers, allowing researchers to search by author, task type, and keywords, and to retrieve the full text of search tasks where available. A number of research papers have been written by Freund and colleagues based on the RepAST dataset, which map out the wide range of types and formats of assigned search tasks in use, and which clearly point to opportunities for greater methodological clarity and consistency[4, 10, 11].

However, the impact of RepAST as a means of sharing and reusing search task descriptions is less clear. While there is anecdotal evidence that researchers use it to find and re-use tasks, there is no clear mechanism or guidelines for attribution. The current version of RepAST has a number of other limitations, including the high level of effort required to manually analyze articles and add to the database and the lack of quality control over search task descriptions, which are not vetted, but are drawn verbatim from the research literature.

2.2 Survey of IIR Researchers

The pre-workshop survey aimed to acquire the views of a wider range of people than were able to attend the workshop. In total, 26 participants were recruited, with almost all participants having conducted at least three IIR studies and most having experience of 5+ studies. The views represented in the study are thus definitely representative of the more experienced segment of IIR researchers. More than half the participants had participated in both shared campaigns and individual studies.

In addition to inquiring about their experience, the survey collected responses on the following aspects: what data participants had collected in IIR studies they had undertaken, the degree to which aspects of the IIR studies were documented and available for re-use, reasons why they had not re-used existing things in previous studies, which aspects of IIR studies they felt had the potential for re-use, open-ended suggestions on how to maximize re-use, and general comments.

Log data and survey data were collected by almost all studies, with observational and interview data collected for some studies. Additionally, some participants reported collecting more complex data such as eye tracking and neurophysiological data.

When asked about the documentation and availability of aspects of the IIR studies they had undertaken, participants indicated high documentation levels and public availability for metrics, protocols, and tasks. Some public documentation of design and evaluation of the studies, and only private availability of log data and interaction data. An open question is how participants interpreted documented and publicly available and what level of documentation is provided. Potentially, the publication is seen as providing these, but whether they are detailed enough for re-use is unclear.

The reportedly biggest hurdle for re-use is a lack of fit between the materials available for re-use and the requirements of the new study. This was followed by lack of awareness of what

[3] `https://ils.unc.edu/searchtasks/search.php`

materials were available for re-use. Potentially the lack of fit is driven by the need to quickly publish, making it easier to just rebuild from scratch, rather than see how the existing materials can be adapted to fit the new study. One major aim of a repository thus clearly needs to be to make re-use easier than rebuilding.

Looking at which aspects participants were open to re-using in the future, participants indicated a high preparedness to re-use all aspects (data, metrics, procedures, protocols, questionnaires, and tasks), with the exception of interaction data. In particular tasks and metrics were seen as primary candidates for re-use. In the open-ended suggestions for how to maximize re-use, suggestions were focused on raising awareness and improving documentation. Ideas for raising awareness included a central, open repository and a monthly mailing list notifying of new materials that could be re-used. Suggestions for documentation included standardizing how things are reported and providing a space where more detail can be provided than is generally possible in a paper.

Finally, the general comments focused on the question of determining where the optimum point was for documentation and re-use versus the cost in providing that and adapting existing IIR study materials. The general consensus mirrored the other responses, in that tasks and metrics were seen as prime candidates for re-use, with less scope for the other aspects.

One of the biggest issues the analysis highlighted is the contentiousness of the term "re-use", where a frequent interpretation seems to be literal 1:1 re-use, rather than the wider "we took X and adapted it to our needs". This led to some very negative comments, as the scope for literal re-use is clearly limited. However, re-use should really be framed in the wider interpretation. Potentially in future work, the word "re-use" should be avoided, instead focusing on improving research quality by building on the best of the past, replication, and comparability of IIR studies. This also provides a clearer goal for the repository, focusing on the efficiency of re-use, quality of research, and the preservation of experience in IIR.

2.3 Viewpoints on Standardization and Re-use

The first discussion session started with a summary of the CHIIR 2017 workshop "Supporting Complex Search Tasks"[1] and the ELIAS-supported Expert Meeting that followed it. These two meetings were direct instigators of this BIIRRR 2018 workshop, as they touched on issues of replication and comparison of IIR studies, and the re-use of existing IIR materials. The participants of the expert meeting felt a need for a more detailed discussion of standardization and re-use.

This session started with a reiteration that although most IIR studies are unique in overall configuration, there are always common elements that can be compared with others. Without overlap, there would be no lessons for other studies or the broader community. It is important to point out the comparable aspects of IIR studies.

Next, we discussed examples of reusable IIR components, such as the Experiment Support System[4] and the Python Interactive Information Retrieval Evaluation workbench (PyIRE)[5]. The former offers a workflow for setting up experimental designs and running experiments with users. The latter is modeled after the WIiRe[9] system, but in deploying it for multiple IIR experiments, about a third of the code had to be rewritten for each new study, showing that building a system

[4] `https://experiment-support-system.readthedocs.io/en/latest/`

[5] `https://pyiire.readthedocs.io/en/latest/`

generic enough for re-use is difficult. This led to the question of the extent to which software systems for user studies can be standardized. Clearly the challenge is to find a good balance between configuration and comparability.

The main part of the discussion focused on addressing a set of questions regarding the premise of the workshop, namely, to what extent are standardization and re-use possible and desirable.

What is the need for or potential value of a shared repository for IIR study design and evaluation? First and foremost, a repository is valuable in that it helps preservation of research for future consultation. Papers describing IIR studies an evaluation have page limits, so almost by necessity, have to leave out details that are required to properly asses and replicate a study. Another valuable contribution is towards efficiency. Even though rewriting one third of the code is a lot of effort, it still means two thirds of the work can be saved. Third, a shared repository contributes to cohesiveness, because it could increase comparability between studies. Re-use of existing components from earlier studies would ask for and allow a comparison with those earlier studies. Finally, a repository may increase the quality of research as exemplars can be pointed out and researchers new to IIR could more easily learn from and be guided by existing studies.

What is the need for or potential value of standards in IIR study design and evaluation? The most important aspect is comparability of studies. Although overall each study might be unique, comparability of aspects of studies is useful for sanity checking results and assessing how the findings fit in our broader knowledge and understanding of information behaviour and interaction. Second, standards lower the effort needed to replicate and validate previous studies. Finally, in line with the value of a repository, standards allow for quality control. For instance, a badge or voting system can be used to award studies for how well they adhere to standards.

What are the barriers to re-use and standardization of IIR studies and materials? A serious barrier is that there is currently no agreement on design aspects, task design or measures. Even if there are standards for these aspects, there will always be a need to adjust them to a specific scenario or research goal, which would then again challenge the upkeep of these standards. Standardization, proper documentation and preservation of IIR studies requires more effort for researchers. Without an effective motivation (a carrot and/or a stick), it may be difficult to move the community to put this extra effort.

Furthermore, the survey results show that there is a perception that "Every study is unique" and that therefore any efforts to develop standards and a shared repository are not worth it. Uptake of either will require convincing the community of their value.

What are ways of making components reusable and studies comparable? One way is to make materials available via a single repository. Another is to encourage researchers to share their materials via open data sharing platforms such as Dataverse[6] or the Open Science Framework[7]. Where components of a study are related to previous studies (e.g., different editions of the same interactive track), it would be useful to describe the differences between editions/versions, as well

[6] `https://dataverse.org/`

[7] `https://osf.io/`

as reasons for any changes with respect to these previous studies.

Which aspects are candidates for standardization and re-use? Search tasks are already reusable via RepAST. At the moment, it is not clear how often researchers have integrated any of the tasks made available via RepAST, as it is possible that some have used the repository without citing it. Other candidates are questionnaire components, such as questions about demographics, technology adoption, technology expertise/experience, domain knowledge and expertise and user engagement. Both the questions and the scales (answer options) could be re-used. For the latter, it is worth considering to offer standardized scales and documentation or guidelines on how to interpret them. The protocol or general experimental procedure can also be made reusable. Protocols often consist of several standard elements, such as intro, pre-study questionnaire, training task, pre-task questionnaires, task, post-task questionnaire, post-study questionnaire.

To help researchers adopt these standards and re-use components, the community should develop guidelines for documenting IIR studies and components, adapting study designs, making and (re-)using consent and permissions forms, and collecting and giving access to user data. Several such guidelines already exist, such as Diane Kelly's book on how to do IIR studies[6] - Methods for Evaluating Interactive Information Retrieval Systems with Users, and guidelines and protocols from related fields such as cognitive psychology.

2.4 Requirements for IIR Re-Use and an iRepository

In the second part of the workshop, participants were asked to brainstorm on the requirements for a possible iRepository either from the position of a dreamer, a realist or critic using the Disney brainstorming method[8]. The three break-out groups reported back what their ideas were. From a dreamer point of view, an IIR repository would function as a one-stop solution providing support for researchers at each stage of an IIR study. Through a mostly automated workflow, content could be easily and standardized integrated as well as retrieved. The second breakout group (realists) was mainly concerned with infrastructural and contextual questions such as funding, sustainability, leadership, responsibility and motivational as well as acceptance aspects. Surprisingly, the group of critics did not come up with more critical points compared to realists argumentation. However, all groups shared the maintenance question as a fundamental success factor. Maintenance challenges have been experienced directly by many of the workshop participants who have held leadership roles in collaborative IR projects, which require substantial efforts over many years, and are often unfunded and/or lacking in organizational support.

Following from this discussion, the workshop participants agreed on five main fields of work to take this project forward:

- **Identification of user groups and use cases (Establish value propositions and requirements)**. Participants agreed that especially students could benefit from a repository which points them to relevant researchers, studies and common approaches. Another potential user group would be researchers that change fields or want to plan interdisciplinary studies without or with little knowledge and experience in IIR.

[8] http://www.designorate.com/disneys-creative-strategy/

- **Infrastructure development (use of open source elements, re-use and adaptation of existing repository structures)**. Previous projects have shown that specific infrastructures are often not transparent to the wider community and code or data might disappear with the departure of single persons or groups. The usage of existing, open source and well documented structures saves resources and allows community driven developments.
- **Movement & Infrastructure Maintenance (Institution independent, long-term oriented, senior researcher function)**. The proposed repository needs to be independent of institutional infrastructures or resources. Ideally the project is managed and maintained by a permanent team of researchers that ensures a stable support and development.
- **Scale & Scope (Identify and prioritize study requirements and potential re-using elements)**. The survey results have shown that not all IIR study elements are equally relevant for sustainable usage. As part of this project we aim at identifying which aspects of IIR studies are considered most important, challenging or time-consuming and therefore might be candidates for re-use. Also, it needs to be determined which aspects could be used in future studies and which might be individually designed for a specific problem which is not readily reproducible.
- **Collaboration, Promotion & Attribution (Interdisciplinary collaborations, promotion via conferences, funding options)**. The awareness and acceptance level of projects like RepAST or the proposed iRepository are crucial success factors. For most disciplines specific rules exist that define research standards. Since IIR studies are often designed and conducted by interdisciplinary teams, it needs to be assured that different domains are included in the planning and organization. At the same time an iRepository could contribute to an open research infrastructure providing the necessary incentives for data sharing. This can only be achieved in collaboration with stakeholders encouraging and demanding research data dissemination (in the context of publications).

3 Continuing Activities

Breaking down the barriers to re-use in the IIR field is not an activity that can be completed within a single-day workshop and this shapes the proposed workshop outcomes. The outcomes are thus centered around follow-on activities to move the community forward.

The overall goal of the workshop was to come up with a plan of concrete short and long-term action points. For the short term, regular meetings in conjunction with other conferences are planned in order to keep the discussion and development alive. A continuation of this workshop is planned via contributions to CHIIR 2019 either as a position paper, workshop or panel proposal, to include an analysis of previous IIR study documentation, identifying challenges and opportunities in more detail. At the same time, a first iRepository prototype will be designed to serve as a starting point and feedback module for further development. The iRepository will link to and build upon prior collaborative projects, including RepAST. The main driver for the iRepository design must be to minimise the overhead both for contributing to and also maintaining the iRepository. Keeping the overhead low is crucial, as researchers' time is already stretched and any high-overhead activities will almost guarantee that contributions will be minimal, in which case the repository

fails. We plan to draw on ideas from successful open-source community infrastructure projects such as NPM[9], while at the same time re-using as much existing research data infrastructure, such as the Open Science Framework and its integration with the various Dataverse installations around the world. In this way, researchers can contribute their IIR data via Dataverse to remain in control of their own data, and the iRepository could functions as a lightweight shell that gives access specifically to IIR resources. Source code for the iRepository can be made available via GitHub, so that over time it remains easy for new maintainers to take over without risk of losing access to the data themselves.

For the long term, we envision bringing together interested researchers from various fields and backgrounds in a more extensive activity (e.g. a multi-day workshop or a hackathon) in order to enable conceptual and technological movement forward. At the same time, we will also need to work with institutional and research stakeholders on the associated policy issues encouraging the community to include shared standardization and documentation practices into their research workflow. The bottom-up approach of developing re-use and documentation guidelines as well as the development of an infrastructure will hopefully also prepare the breeding ground for a shift in research and scholarly publication practices.

References

[1] Nicholas Belkin, Toine Bogers, Jaap Kamps, Diane Kelly, Marijn Koolen, and Emine Yilmaz. 2017. Second Workshop on Supporting Complex Search Tasks. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, New York, NY, USA, 433–435. DOI:http://dx.doi.org/10.1145/3020165.3022163

[2] Toine Bogers, Maria Gäde, Luanne Freund, Mark Hall, Marijn Koolen, Vivien Petras, and Mette Skov. 2018. Workshop on Barriers to Interactive IR Resources Re-use. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 379–385. DOI:http://dx.doi.org/10.1145/3176349.3176901

[3] Pia Borlund. 2016. Interactive Information Retrieval: An Evaluation Perspective. In *CHIIR '16: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 151–151.

[4] Luanne Freund and Barbara M. Wildemuth. 2014. Documenting and studying the use of assigned search tasks: RepAST. *Proceedings of the American Society for Information Science and Technology* 51, 1 (2014), 1–4. DOI:http://dx.doi.org/10.1002/meet.2014.14505101122

[5] William R Hersh. 2002. TREC 2002 Interactive Track Report. In *TREC*. https://trec.nist.gov/pubs/trec11/papers/INTERACTIVE.OHSU.pdf

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

[9] https://www.npmjs.com

[7] Ragnar Nordlie and Nils Pharo. 2012. Seven Years of INEX Interactive Retrieval Experiments – Lessons and Challenges. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, Tiziana Catarci, Pamela Forner, Djoerd Hiemstra, Anselmo Peñas, and Giuseppe Santucci (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 13–23.

[8] Paul Over. 2001. The TREC interactive track: an annotated bibliography. *Information Processing & Management* 37, 3 (2001), 369–381. DOI:http://dx.doi.org/https://doi.org/10.1016/S0306-4573(00)00053-4 Interactivity at the Text Retrieval Conference (TREC).

[9] Elaine G. Toms, Luanne Freund, and Cara Li. 2004. WiIRE: the Web interactive information retrieval experimentation system prototype. *Information Processing & Management* 40, 4 (2004), 655–675.

[10] Barbara M. Wildemuth and Luanne Freund. 2012. Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*. ACM, 4.

[11] Barbara M. Wildemuth, Luanne Freund, and Elaine G. Toms. 2013. Designing known-item and fact-finding search tasks for studies of interactive information retrieval. In *Proceedings of the second association for information science and technology ASIS&T (European Workshop).* 131–162. http://www.abo.fi/sitebuilder/media/29327/aew2013proceedings.pdf