An Asynchronous Scheme for the Distributed Evaluation of Interactive Multimedia Retrieval

Sauter, Loris ; Gasser, Ralph ; Bernstein, Abraham ; Schuldt, Heiko ; Rossetto, Luca

Abstract: Evaluation campaigns for interactive multimedia retrieval, such as the Video Browser Shodown (VBS) or the Lifelog Search Challenge (LSC), so far imposed constraints on both simultaneity and locality of all participants, requiring them to solve the same tasks in the same place, at the same time and under the same conditions. These constraints are in contrast to other evaluation campaigns that do not focus on interactivity, where participants can process the tasks in any place at any time. The recent travel restrictions necessitated the relaxation of the locality constraint of interactive campaigns, enabling participants to take place from an arbitrary location. Born out of necessity, this relaxation turned out to be a boon since it greatly simplified the evaluation process and enabled organisation of ad-hoc evaluations outside of the large campaigns. However, it also introduced an additional complication in cases where participants were spread over several time zones. In this paper, we introduce an evaluation scheme for interactive retrieval evaluation that relaxes both the simultaneity and locality constraints, enabling participation from any place at any time within a predefined time frame. This scheme, as implemented in the Distributed Retrieval Evaluation Server (DRES), enables novel ways of conducting interactive retrieval evaluation and bridged the gap between interactive campaigns and non-interactive ones.

DOI: https://doi.org/10.1145/3552467.3554797

Originally published at:
Sauter, Loris; Gasser, Ralph; Bernstein, Abraham; Schuldt, Heiko; Rossetto, Luca (2022). An Asynchronous Scheme for the Distributed Evaluation of Interactive Multimedia Retrieval. In: MM '22: The 30th ACM International Conference on Multimedia, Lisboa Portugal, 10 November 2022 - 14 November 2022. ACM, 33-39.
DOI: https://doi.org/10.1145/3552467.3554797

An Asynchronous Scheme for the Distributed Evaluation of Interactive Multimedia Retrieval

Loris Sauter
loris.sauter@unibas.ch
University of Basel
Basel, Switzerland

Ralph Gasser
ralph.gasser@unibas.ch
University of Basel
Basel, Switzerland

Abraham Bernstein
bernstein@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Heiko Schuldt
heiko.schuldt@unibas.ch
University of Basel
Basel, Switzerland

Luca Rossetto
rossetto@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

ABSTRACT

Evaluation campaigns for interactive multimedia retrieval, such as the Video Browser Shodown (VBS) or the Lifelog Search Challenge (LSC), so far imposed constraints on both simultaneity and locality of all participants, requiring them to solve the same tasks in the same place, at the same time and under the same conditions. These constraints are in contrast to other evaluation campaigns that do not focus on interactivity, where participants can process the tasks in any place at any time. The recent travel restrictions necessitated the relaxation of the locality constraint of interactive campaigns, enabling participants to take place from an arbitrary location. Born out of necessity, this relaxation turned out to be a boon since it greatly simplified the evaluation process and enabled organisation of ad-hoc evaluations outside of the large campaigns. However, it also introduced an additional complication in cases where participants were spread over several time zones. In this paper, we introduce an evaluation scheme for interactive retrieval evaluation that relaxes both the simultaneity and locality constraints, enabling participation from any place at any time within a predefined time frame. This scheme, as implemented in the Distributed Retrieval Evaluation Server (DRES), enables novel ways of conducting interactive retrieval evaluation and bridged the gap between interactive campaigns and non-interactive ones.

CCS CONCEPTS

• **Information systems** → *Retrieval tasks and goals*; *Users and interactive retrieval*; **Evaluation of retrieval results**.

KEYWORDS

Interactive Retrieval Evaluation, Distributed Evaluation, Asynchronous Evaluation

IMuR '22, October 14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9497-0/22/10.
https://doi.org/10.1145/3552467.3554797

ACM Reference Format:
Loris Sauter, Ralph Gasser, Abraham Bernstein, Heiko Schuldt, and Luca Rossetto. 2022. An Asynchronous Scheme for the Distributed Evaluation of Interactive Multimedia Retrieval. In *Proceedings of the 2nd International Workshop on Interactive Multimedia Retrieval (IMuR '22), October 14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3552467.3554797

1 INTRODUCTION

Since interactive multimedia retrieval involves the active participation of human actors, special considerations for the benchmarking setup are necessary in order to obtain comparable results. Evaluation campaigns such as the Video Browser Showdown (VBS) [20] or the Lifelog Search Challenge (LSC) [1] address this by performing an annual competition-style event, wherein all participants gather at the same physical location to solve a series of previously unknown retrieval tasks at the same time. While this setup equalizes the conditions for all participants as much as possible, it comes with a substantial organizational overhead. During the past three years, this has been exacerbated by the restrictions in international travel. Since it was not possible for multiple instances of these campaigns [1–3, 5] to be hosted at one common location that all participants could reach, the requirement on common locality had to be relaxed and the evaluation had to be conducted in a distributed setting. To enable such distributed campaigns, we introduced the 'Distributed Retrieval Evaluation Server' (DRES)[1] [17], which provides the underlying basic infrastructure. DRES manages all relevant aspects of an interactive multimedia retrieval evaluation, such as the presentation of query information to participants and the assessment and scoring of submitted results. DRES does this in both traditional local settings, where all participants gather in the same location, as well as fully distributed ones, where no two participants share a location, and any combination thereof. Since its introduction, DRES has been used not only for established international campaigns but also for various other experiments, such as those presented in [16], which could benefit from a distributed setting. In order to maximize its usefulness to the community, DRES is made freely available as open-source software.

The relaxation of the locality requirement enabled the continuation of established campaigns and, now that travelling becomes increasingly possible again, simplifies organisation of campaigns in

[1] https://dres.dev

hybrid settings. Furthermore, it provides new opportunities for experimentation, since smaller evaluations can be performed without any of the participant leaving their offices. However, this relaxation of the locality requirement comes with new challenges, especially when participants are spread over several time zones, since thus far, simultaneity has still been a requirement. In this paper, we introduce an extension to DRES, which relaxes this simultaneity requirement by enabling the *asynchronous* evaluation of retrieval tasks allowing for interactive multimedia retrieval evaluation campaigns that can be decoupled from not only location but also time.

After surveying related work in the context of multimedia retrieval evaluation in Section 2, we outline the mechanisms that have been used for synchronous interactive retrieval evaluations in Section 3. In Section 4, we then detail how these mechanisms can be adapted to an asynchronous setting and what consequences are to be expected. Section 5 offers some details on how the discussed concepts are implemented in DRES and Section 6 outlines some possibilities that are offered by this distributed asynchronous evaluation scheme, which prior to DRES were not possible. We close with some concluding remarks and an outlook towards future opportunities and challenges in Section 7.

2 RELATED WORK

The evaluation of information retrieval techniques and systems has been an important pillar for research in that domain for many years now. The Text Retrieval Conference (TREC) initiative dates back to 1991, when it was started to incentivize cooperation and coordination among researchers and provide standards, such as collections and task definitions, to establish a common baseline that often did not exists before [15]. Since its conception, TREC gave rise to standardised benchmarks for ad-hoc search and retrieval in text documents and many other areas that—over the years – developed into independent tracks such as TRECVid for retrieval in video collections, which started in 2001 [21]. Similar initiatives include but are in no way limited to CLEF [13], ImageCLEF [7] or MediaEVAL [9], all of which exhibit a slightly different theme and focus. What all of these grand challenge campaigns have in common, however, is that participation takes place asynchronously. Tasks are usually published at a given point in time. Teams are then given a time window spanning several days or even weeks during which they can solve these tasks and submit their results, which are subsequently validated to produce a final score. Interactivity during task execution plays little to no role at all.

This stands in opposition to more recent, interactive, real-time retrieval campaigns such as the VideOlympics [22], the Video Browser Showdown (VBS) [11, 19, 20] or the Lifelog Search Challenge (LSC) [1–3], which regard multimedia retrieval as a collaboration between a system and an operator [19, 22]. These campaigns are inherently synchronous as they typically take place on-site during conferences, where a certain number of teams gather to solve different tasks in a time-boxed setting. Every year, the best systems are identified [4, 6, 8], which provides a continuous incentive to improve upon previous iterations. Many interesting insights have been gained from these events, including but not limited to the role of deep learning in video retrieval [5] or comparison between searching and browsing techniques [12], leaving this type of evaluation as a promising path to new discoveries and to novel trends in multimedia retrieval.

Over the years, a lot of effort went into the curation of standardised multimedia collections upon which evaluations campaigns can be based [2, 3, 14, 18], which arguably is a critical factor for success. Furthermore, some attention was given to the proper definition of benchmark workloads [15] and task settings [10] as well as approaches to scoring the performance of teams and their systems [11]. However, very little attention has been given thus far to a proper formalisation of the setup that enables both synchronous and asynchronous types of evaluation campaigns, which we argue is at least as important to generate reproducible and meaningful results. To the best of our knowledge, the first attempt at formalising the entire process of the campaign end-to-end from task definition to execution was attempted with the conception of DRES [17], which was built out of necessity during the COVID-19 pandemic, when classical, on-site evaluation campaigns could not be conducted and distributed settings were required.

3 SYNCHRONOUS EVALUATIONS

Prior to mid-2020 [11], state-of-the-art interactive multimedia retrieval evaluations had to be organised on-site; participants had to gather in the same location and had to concurrently solve evaluation tasks. Particularly, on a practical note, they had to be directly connected to the evaluation system[2] used at the time. Furthermore, in cases where human judging was required to determine the correctness of retrieved results, participants did have to idle until the judging process completed and the next task was available.

Conducting such an evaluation campaign involves a lot of organisational effort: Gathering a large number of participants in a single location and synchronising the start of tasks in order to have the full attention of participants adds to this complexity.

With the introduction of DRES [17], four distinct entities are involved in an evaluation: The *administrator* orchestrates the evaluation as it unfolds and takes care of setting up the evaluation beforehand. This setup involves defining categories and creating tasks. Such tasks fundamentally consist of instructions and solutions, i.e., hints on how the task is to be solved and a ground truth to compare results to. However, for some categories, administrators might omit the pre-definition of a ground truth. Such tasks are later manually assessed by the entity called *judge*. The remaining two entities are *participants* and *viewers*. There is only one difference between these two: Participants actively take part in the evaluation by examining tasks and submitting solutions, whereas viewers only display information and cannot interact with the evaluation directly. Participants are human agents that interact with the system during interactive evaluation campaigns such as LSC or VBS. However, for most of the time of such a campaign, participants operate their multimedia system under evaluation and the interplay with DRES is reduced to gather information about the current task and the eventual submission. Campaigns such as LSC and VBS are set up as competitions and thus, in order to support an audience, the viewer entity was designed.

[2] https://github.com/klschoef/vbsserver/

Courtesy of its distributed architecture, DRES lifts the restriction of a local evaluation. One prerequisite for distribution is the decoupling of the definition of a task from its state within an evaluation and separation of the management from the presentation plane. For example, this allows for repetition of tasks, e.g., if there was a problem with the run, and it prevents the judgement process from halting the continuation of the evaluation campaign.

3.1 Synchronous Sequence of Events

Figure 1 depicts a sequence diagram of the events taking place during a synchronous evaluation over time. In this example, only two participants are shown, but any further participant would simply generate more but equivalent events. Additionally, the perspectives of the evaluation server, its administrator, as well as a single judge are depicted. The judge is responsible for the assessment of submissions for tasks where no ground truth exists. And again, the example only shows one judge since additional judges would not materially alter the flow of events. All participants as well as judges and administrators have user accounts that authenticate them and assign them their respective roles.

At the beginning, the administrator sends a command to the server, instructing it to create a new instance of an evaluation based on a template created prior to the evaluation. This causes the server to handle the required pre-processing, after which it informs the administrator as well as all participants about the new evaluation's availability (Figure 1, Phase I). At any point after that, the administrator can start the evaluation, which causes further preparations on the server side followed by a notification to all involved parties. After this point, the evaluation is active and tasks can be started. The administrator can select a task for the participants to solve, which again causes the server to send an update to all participants. Then, the task is ready to be evaluated. To do this, the administrator instructs the server to start the task. Subsequently, the server makes all task information available to the participants and informs them about its availability. As a consequence, the participants request all the relevant information from the server. Since this information can be comparatively large, especially for tasks that involve video-based target descriptions, the transfer from the server to the participants can take some time. Once a participant has received all information, it acknowledges this. As soon as all participants have acknowledged the receipt and thereby their readiness, the task starts. This synchronization step is important to ensure that all participants see the same information about the task at the same time. It is therefore unavoidable to introduce some delays, as the time between the administrator requesting the task to start and the actual start of a task is determined by the slowest participant (Figure 1, Phase II).

Once the task has started, the server begins accepting submissions (of answers) for this task from any of the registered participants (Figure 1, Phase III). In our example, the first task has a known ground-truth. Therefore, a submission (ω) is evaluated immediately upon it being received and an acknowledgement together with the information about the submission's correctness is returned to the participant.

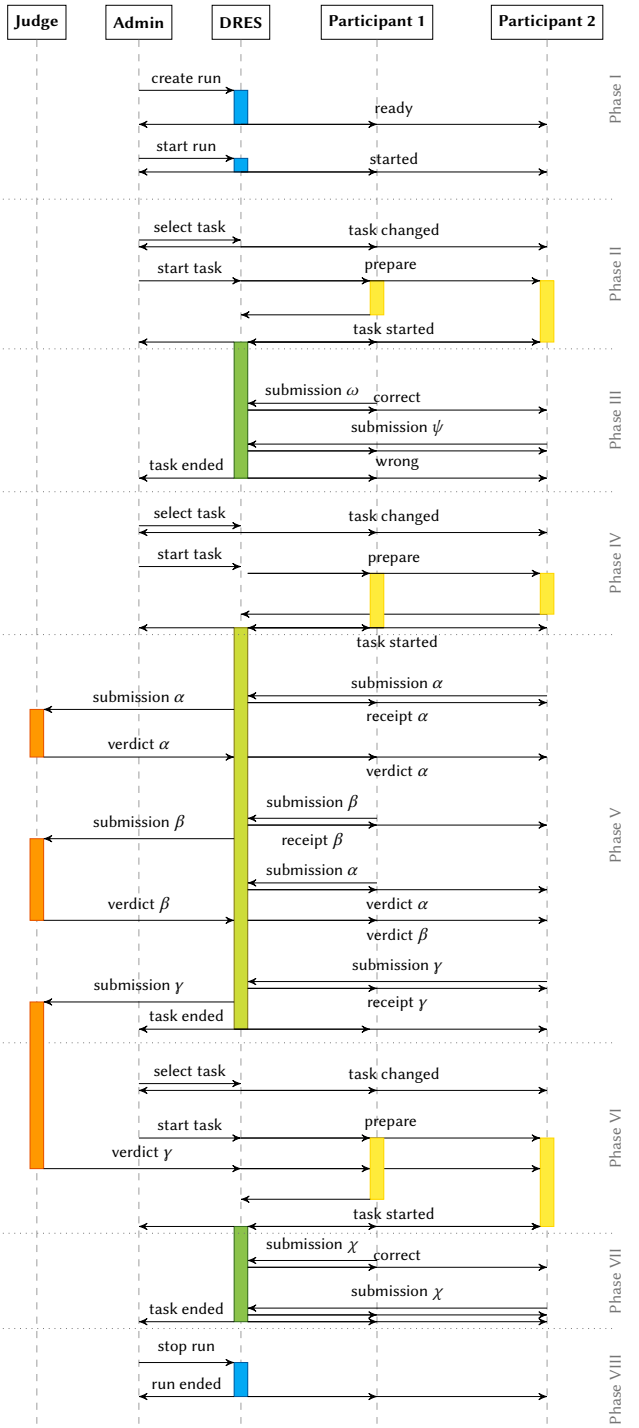

Figure 1: Pseudo sequence diagram of the events over time for an example synchronous evaluation. DRES denotes the DRES system, while Judge, Admin, Participant 1 and Participant 2 are human agents.

Simultaneously, all participants are informed about there being a new submission, in order to ensure a consistent state across all participants. The amount of details being revealed to the participants at this time depends on the type of the task and is therefore freely configurable by the administrator. The first task ends due to exhaustion of total time allotted for the task (Figure 1, Phase III). Depending on how the task is configured, other criteria for task termination are that all participants have one correct submission (see Figure 1 Phase VII) or deliberate intervention by the administrator.

The process for starting the second task is the same as for the first (Figure 1, Phase IV). In this example, the second task (Figure 1, Phase V) has no pre-defined ground-truth and accepts an arbitrary number of possibly correct submissions. For the first submission (α), the server has no information about its correctness. It therefore acknowledges its receipt to the participant without any further information, and informs all participants as it did before. The received submission is then sent to a judgement queue from which it can be retrieved by a judge. The judge inspects the submission and renders a verdict about its correctness, which is then sent back to the server. Once the server receives such a verdict, it updates the state of the corresponding submission, recomputes the scores of the relevant task and notifies all participants about the updates. If a submission (α) is received from a participant that has been previously received from another and has already been assessed by a judge, the server directly assigns the previously assigned verdict (Figure 1, Phase IV).

The judgement process is decoupled from the participant-facing task life-cycle. It is therefore possible for a task to end and no longer accept further submissions before all received submissions have been assessed by a judge. This is also illustrated in our example, that shows submissions from the second task still being sent to the judge and verdicts being received while participants are already involved with the third task (Figure 1, Phase VII). Once all tasks have been completed, the administrator can end the evaluation (Figure 1, Phase VIII).

4 ASYNCHRONOUS EVALUATIONS

In order to enable synchronous distributed evaluations, the requirements of common locality had to be relaxed while the simultaneity of tasks was kept in place. This resulted in the trade-off that a controlled environment shared by all participants was given up in exchange for greatly simplified logistics and a lower barrier of entry for participants. Distributed, synchronous evaluations come with an additional logistical difficulty in cases where participants are spread over a wide geographic area, as the differences in time zones can limit the time frames of common availability.

Motivated by this limitation, we aim to explore another trade-off to see what can be gained by also relaxing the simultaneity requirement by the introduction of asynchronous evaluations. In an asynchronous evaluation, all participating teams solve the same tasks independently of each other at any point in time during which the evaluation is active. From the perspective of any participant, the setting looks almost identical to a synchronous distributed evaluation where they are the only participant. The only actionable difference is that, in contrast to synchronous evaluations where the tasks are controlled by the evaluation administrator, the participants request the start of a next task whenever they are ready. Since there is no dependency between participants, no synchronization step upon the start of a task is necessary and tasks can start immediately upon request.

4.1 Asynchronous Sequence of Events

Figure 2 illustrates a sample sequence diagram of events during an asynchronous evaluation, analogously to the example shown in Section 3.1. This example shows the same two participants and uses the same tasks.

The example begins analogously to what is shown in Section 3.1 with the administrator creating and launching an evaluation. Once the evaluation is started (Figure 2, Phase I), the administrator does not need to manage anything, since all participants have control over their own view. In our example, this is shown by the first participant who requests the launch of the first task assigned to them. Once the server receives this request, the task can start immediately as no synchronization with other participants is necessary. In this example, the first task shown to participant 1 is a task with a known ground-truth. As soon as the participant makes a correct submission (φ), the task terminates, since there is no other possible action remaining (Figure 2, Phase II-1). Once the task is over, the participant can request the next task. In contrast to the administrator in the synchronous setting, the participant cannot request an arbitrary task from the list but has to solve the tasks in the order presented by the server. This order can either be fixed for all participants or randomized, ensuring that no two participants will see the tasks in the same order. In this example, we show the tasks thus randomized. As soon as the server acknowledges the availability of the next task, the participant can request the start of the task at any time (Figure 2, Phase III-1). Once the task is started, it runs until either a success condition is reached or the allotted time is exhausted.

While participant 1 is engaged with their second task, participant 2 starts their evaluation by requesting the start of their first task (Figure 2, Phase II-2). The task presented first to participant 2 is one without a ground-truth and accepting an arbitrary number of submissions. These submissions are queued for assessment until a judge is available.

While participant 2 is still engaged with their first task, participant 1 has run out of time for their second one. Depending on the task configuration, the correct solution might be displayed at this point in time. Subsequently, participant 1 requested and started their third task (Figure 2, Phase IV-1). In this example, this is the same task as the one still active for participant 2. While participant 1 is still engaged with their last task in this example, the judge becomes active and starts to request submissions to be assessed (see within Figure 2, Phase IV-1). In this example, this is a reasonable point in time for the judge to begin with the assessment, since all tasks without a ground-truth are either active or already completed. Where this not the case, the judge would need to come back at a later point in time to handle any future submissions.

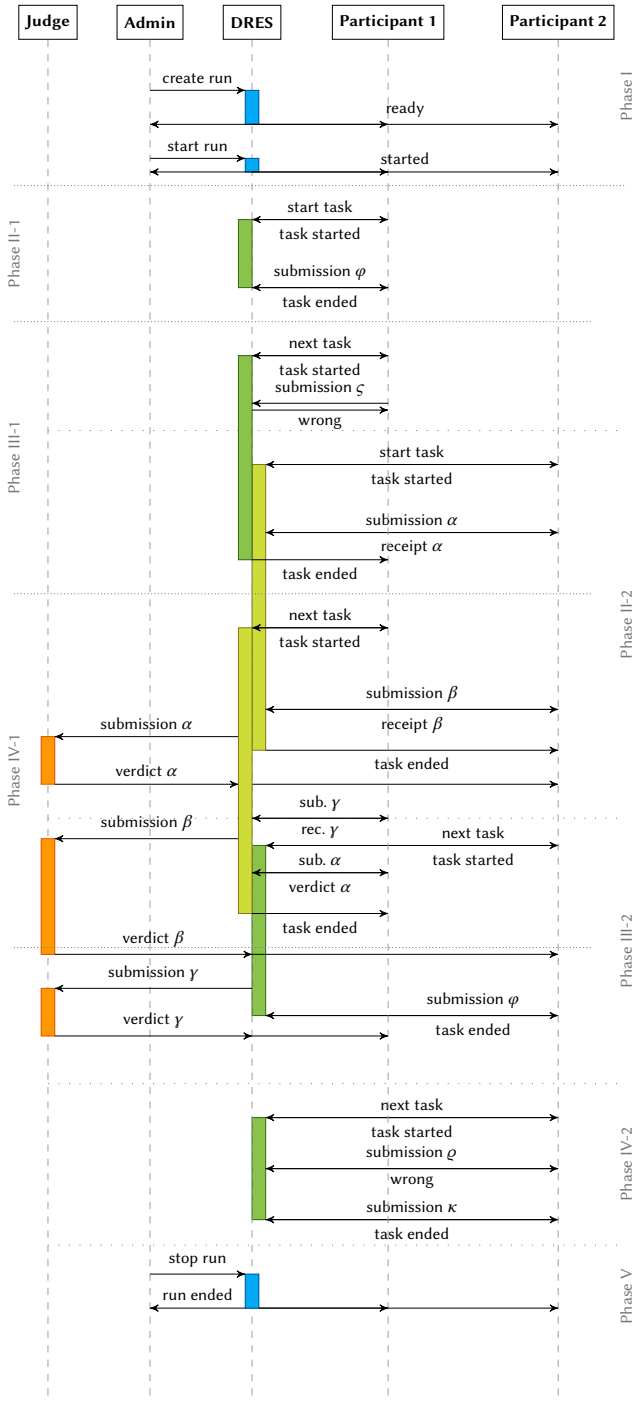

Figure 2: Pseudo sequence diagram of the events over time for an example asynchronous evaluation. As in Figure 1, human agents are Judge, Admin, Participant 1 and 2.

As for the synchronous example, when a submission is sent that was previously judged (see Figure 2, submission α), the system immediately replies with the verdict. Additionally, hinted by the submission numbering, judges cannot derive from which run or participant submissions are coming.

In the meantime, participant 2 completed their first task and is free to request the next. For their two remaining tasks, they can continue in an equivalent manner.

As soon as the time of the last task of participant 1 has elapsed, they have completed all their tasks and are done with the evaluation. There are no more actions they can take within this evaluation. However, they can check back at any point in time while the overall evaluation is still active to see the aggregated scores of all participants.

When all participants are done with all their tasks and all submissions have been assessed by the judge(s), there are no more actions that would influence the outcome of the evaluation as a whole and the scores will not change anymore. At this point, the evaluation is done and the administrator can close it.

5 IMPLEMENTATION

In order to provide the distinction of the entities administrator, judge, viewer and participant, we have implemented a role-based user management system in which each entity corresponds to a role. Users only have access to the components of DRES their role requires. For instance, participants cannot judge submissions or create evaluations, while administrators have full access. However, the role participant usually also represents a multimedia retrieval system to be evaluated and thus, a link between the DRES user and the system has to be established. In contrast to the previous, manual, error-prone identification that had to be sent upon submission, DRES provides standardised token-based authentication, as well as dedicated OpenAPI[3] client specifications[4] for submission endpoints. This approach minimises human errors that, in the past, could have lead to submissions without corresponding participants and simplifies the evaluation setup due to the ability to generate code from the OpenAPI specifications. At its core, DRES follows the client-server architecture, which is essential to enable our distributed approach.

Furthermore, for evaluations as a whole, as well as for individual tasks within an evaluation, we follow an instantiation approach. Once an evaluation or a task is started, an instance is created which is based on an existing template. An instance of an evaluation or a task is immutable with respect to its template.

Internally, we employ an event loop that processes incoming REST and websocket messages, which is a proven and robust approach. Every cycle in the loop updates the internal state through operators that filter and process requests and generate notifications that are sent to interested entities. All state changes are persisted immediately.

To have an additional layer of data security, we also log each incoming request persistently and thus create an audit trail, which

[3] https://swagger.io/specification/
[4] https://editor.swagger.io/?url=https://raw.githubusercontent.com/dres-dev/DRES/master/doc/oas-client.json

allows complete post-hoc reconstruction of an evaluation state at any point in time.

DRES' open-source platform-independent implementation was first demonstrated in [17] and is available at https://dres.dev. The application is self-contained and provides a CLI as well as a front-end, accessible from the web.

6 NOVEL EVALUATION TYPES

The asynchronous, distributed and interactive evaluation scheme introduced in this paper can be seen as bridging the gap between the distributed non-interactive evaluations that are commonly used in a grand challenge format in various disciplines as discussed in Section 2 and the interactive, synchronous format that has so far most prominently been employed by VBS and LSC. The combination of both in a unified framework offers new possibilities for evaluation settings that were previously not feasible, as illustrated in the follwing examples:

Higher-frequency Evaluations: The interactive campaigns described so far take place once a year, partly because organisation and participation incurs a certain overhead and cost. These annual evaluations can now easily be complemented by more frequent iterations that may focus on specific questions and can benefit from the same, standardised environment. Both the relaxed locality and distribution requirement contribute to this. A first glimpse at what is possible with such rather ad-hoc evaluations is given in [16].

Supplement Existing Campaigns: Several of the types of tasks evaluated regularly in existing and well-established campaigns that follow a more traditional setting could arguably also be solved in an interactive manner. So far, this has not been feasible due to the general structure of these campaigns as participants would only meet at the end of the campaign to discuss results if at all, and would hence not have the opportunity for a synchronous interactive evaluation. Now it becomes feasible to supplement such campaigns with interactive aspects without the need to interfere with their overall setup and workflow.

Experimenting with New Tasks: So far, the introduction of new types of tasks into interactive evaluation campaigns has been rather expensive and hence only happens rarely. This can be explained partially by the understandable reluctance of the challenge organizers to risk a substantial amount of the limited available time for experimentation on the challenge format itself. The asynchronous distributed setting offers a much cheaper alternative for the experimentation with novel task types and can serve as a test-bed for such experimentation independently of the actual challenges themselves. This has the potential to catalyze further innovation not only in the challenges but also in the area of research they aim to analyze.

These examples are just a selection of the potential applications of the framework offered by DRES. However, despite the various new abilities DRES offers, there is still an argument for the traditional, on-site competition format. Firstly, it allows for exchange on a professional level about technical details, systems, and strategies, which leads to fruitful discussions. Secondly, it allows for variations such as the novice sessions, where system operators are recruited randomly from the audience. And finally, only an on-site competition allows for the control and thus standardisation of environmental parameters such as lighting, sound quality and the visibility of query hints.

7 CONCLUSIONS AND OUTLOOK

In this paper, we introduced an asynchronous scheme for distributed interactive evaluations of retrieval tasks. The scheme is implemented in the Distributed Retrieval Evaluation Server (DRES) that is freely available as open-source software. This novel way of conducting evaluations of interactive multimedia retrieval approaches comes with substantially less coordination overhead compared to traditional synchronous schemes and offers new opportunities for comparative evaluations that would not have been feasible previously. It is intended to serve as a complementary setting to the more established interactive retrieval evaluations that are conducted with all participants being at the same place at the same time. Thereby, it aims at making such evaluations more accessible and increasing their overall frequency.

Future work on DRES and the formalisation of interactive retrieval evaluation campaigns could address the judgement process for tasks that do not come with a ground-truth. The current process has several opportunities for improvement:

(1) Experience has shown, that different judges sometimes arrive at different verdicts for the same item. This is not optimal and could be addressed by a consensus algorithm involving multiple judges.
(2) Currently, side-channels information could be exploited in order to associate submissions with teams. This can be problematic, especially in the asynchronous model due to the potentially larger delay between submissions of different teams. We could address this by making affiliation of judges with teams explicit and taking this information into account during the assignment process. Furthermore, one could defer the judgement process as a whole.
(3) Judgement as a whole could be designed as a multi-step process with the option for participants to challenge verdicts reached by the judges.

It must be mentioned, however, that the aforementioned proposals cannot be addressed by technical means alone but also require policy changes that must be implemented and enforced by the evaluation organisers.

ACKNOWLEDGMENTS

This work has been partially supported by the Swiss National Science Foundation, Project MediaGraph, Grant Number 202125.

REFERENCES

[1] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoc, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *ICMR '21: International Conference on Multimedia Retrieval, Taipei, Taiwan, August 21-24, 2021*. ACM, New York, NY, USA, 690–691. https://doi.org/10.1145/3460426.3470945

[2] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy.

2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *International Conference on Multimedia Retrieval (ICMR)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3512527.3531439

[3] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoc, Wolfgang Hürst, Minh-Triet Tran, and Klaus Schöffmann. 2020. Introduction to the Third Annual Lifelog Search Challenge (LSC'20). In *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*. ACM, New York, NY, USA, 584–585. https://doi.org/10.1145/3372278.3388043

[4] Silvan Heller, Ralph Gasser, Cristina Illi, Maurizio Pasquinelli, Loris Sauter, Florian Spiess, and Heiko Schuldt. 2021. Towards Explainable Interactive Multimodal Video Retrieval with Vitrivr. In *MultiMedia Modeling - 27th International Conference, MMM 2021, Prague, Czech Republic, June 22-24, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12573)*. Springer, 435–440. https://doi.org/10.1007/978-3-030-67835-7_41

[5] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoc, Andreas Leibetseder, Frantisek Mejzlík, Ladislav Peska, Luca Rossetto, Konstantin Schall, Klaus Schoeffmann, Heiko Schuldt, Florian Spiess, Ly-Duyen Tran, Lucia Vadicamo, Patrik Veselý, Stefanos Vrochidis, and Jiaxin Wu. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *International Journal on Multimedia Information Retrieval* 11, 1 (2022), 1–18. https://doi.org/10.1007/s13735-021-00225-2

[6] Nico Hezel, Konstantin Schall, Klaus Jung, and Kai Uwe Barthel. 2022. Efficient Search and Browsing of Large-Scale Video Collections with Vibro. In *MultiMedia Modeling - 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6-10, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13142)*. Springer, 487–492. https://doi.org/10.1007/978-3-030-98355-0_43

[7] Bogdan Ionescu, Henning Müller, Renaud Péteri, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Dzmitri Klimuk, Aleh Tarasau, Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, Obioma Pelka, Christoph M. Friedrich, Alba Garcia Seco de Herrera, Narciso García, Ergina Kavallieratou, Carlos Roberto del-Blanco, Carlos Cuevas, Nikos Vasilopoulos, Konstantinos Karampidis, Jon Chamberlain, Adrian F. Clark, and Antonio Campello. 2019. ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9-12, 2019, Proceedings (Lecture Notes in Computer Science, Vol. 11696)*. Springer, 358–386. https://doi.org/10.1007/978-3-030-28577-7_28

[8] Miroslav Kratochvíl, Patrik Veselý, Frantisek Mejzlík, and Jakub Lokoc. 2020. SOM-Hunter: Video Browsing with Relevance-to-SOM Feedback Loop. In *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11962)*. Springer, 790–795. https://doi.org/10.1007/978-3-030-37734-2_71

[9] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. 2017. The benchmarking initiative for multimedia evaluation: MediaEval 2016. *IEEE MultiMedia* 24, 1 (2017), 93–96. https://doi.org/10.1109/MMUL.2017.9

[10] Jakub Lokoc, Werner Bailer, Kai Uwe Barthel, Cathal Gurrin, Silvan Heller, Björn Þór Jónsson, Ladislav Peska, Luca Rossetto, Klaus Schoeffmann, Lucia Vadicamo, Stefanos Vrochidis, and Jiaxin Wu. 2022. A Task Category Space for User-Centric Comparative Multimedia Search Evaluations. In *MultiMedia Modeling - 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6-10, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13141)*. Springer, 193–204. https://doi.org/10.1007/978-3-030-98358-1_16

[11] Jakub Lokoč, Werner Bailer, Klaus Schoeffmann, Bernd Münzer, and George Awad. 2018. On influential trends in interactive video retrieval: video browser showdown 2015–2017. *IEEE Transactions on Multimedia* 20, 12 (2018), 3361–3376. https://doi.org/10.1109/TMM.2018.2830110

[12] Jakub Lokoč, Gregor Kovalčík, Bernd Münzer, Klaus Schöffmann, Werner Bailer, Ralph Gasser, Stefanos Vrochidis, Phuong Anh Nguyen, Sitapa Rujikietgumjorn, and Kai Uwe Barthel. 2019. Interactive search or sequential browsing? A detailed analysis of the video browser showdown 2018. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1 (2019), 1–18. https://doi.org/10.1145/3295663

[13] Giorgio Maria Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. 2007. CLEF 2007: Ad Hoc Track Overview. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 5152)*. Springer, 13–32. https://doi.org/10.1007/978-3-540-85760-0_2

[14] Paul Over, George Awad, Alan F. Smeaton, Colum Foley, and James Lanagan. 2009. Creating a Web-Scale Video Collection for Research *(WSMC '09)*. Association for Computing Machinery, New York, NY, USA, 25–32. https://doi.org/10.1145/1631135.1631141

[15] Paul Over, Clement H. C. Leung, Horace Ho-Shing Ip, and Michael Grubinger. 2004. Multimedia Retrieval Benchmarks. *IEEE Multimedia* 11, 2 (2004), 80–84. https://doi.org/10.1109/MMUL.2004.1289045

[16] Luca Rossetto, Ralph Gasser, Silvan Heller, Mahnaz Parian-Scherb, Loris Sauter, Florian Spiess, Heiko Schuldt, Ladislav Peska, Tomás Soucek, Miroslav Kratochvíl, Frantisek Mejzlík, Patrik Veselý, and Jakub Lokoc. 2021. On the User-Centric Comparative Remote Evaluation of Interactive Video Search Systems. *IEEE Multimedia* 28, 4 (2021), 18–28. https://doi.org/10.1109/MMUL.2021.3066779

[17] Luca Rossetto, Ralph Gasser, Loris Sauter, Abraham Bernstein, and Heiko Schuldt. 2021. A System for Interactive Multimedia Retrieval Evaluations. In *MultiMedia Modeling - 27th International Conference, MMM 2021, Prague, Czech Republic, June 22-24, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12573)*. Springer, Cham, 385–390. https://doi.org/10.1007/978-3-030-67835-7_33

[18] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. 2019. V3C - A Research Video Collection. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 11295)*. Springer, 349–360. https://doi.org/10.1007/978-3-030-05710-7_29

[19] Klaus Schoeffmann. 2014. A user-centric media retrieval competition: The video browser showdown 2012-2014. *IEEE MultiMedia* 21, 4 (2014), 8–13. https://doi.org/10.1109/MMUL.2014.56

[20] Klaus Schoeffmann. 2019. Video Browser Showdown 2012-2019: A Review. In *2019 International Conference on Content-Based Multimedia Indexing, CBMI 2019, Dublin, Ireland, September 4-6, 2019*, Cathal Gurrin, Björn Þór Jónsson, Renaud Péteri, Stevan Rudinac, Stéphane Marchand-Maillet, Georges Quénot, Kevin McGuinness, Gylfi Þór Guðmundsson, Suzanne Little, Marie Katsurai, and Graham Healy (Eds.). IEEE, 1–4. https://doi.org/10.1109/CBMI.2019.8877397

[21] Alan F. Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation Campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval* (Santa Barbara, California, USA) *(MIR '06)*. Association for Computing Machinery, New York, NY, USA, 321–330. https://doi.org/10.1145/1178677.1178722

[22] Cees GM Snoek, Marcel Worring, Ork de Rooij, Koen EA van de Sande, Rong Yan, and Alexander G Hauptmann. 2008. VideOlympics: real-time evaluation of multimedia retrieval systems. *IEEE Multimedia* 15, 1 (2008), 86–91. https://doi.org/10.1109/MMUL.2008.21