Report on the 14th Conference and Labs of the Evaluation Forum (CLEF 2023): Experimental IR Meets Multilinguality, Multimodality, and Interaction

Mohammad Aliannejadi
University of Amsterdam
The Netherlands
m.aliannejadi@uva.nl

Avi Arampatzis
Democritus University of Thrace
Greece
avi@ee.duth.gr

Guglielmo Faggioli
University of Padua
Italy
faggioli@dei.unipd.it

Nicola Ferro
University of Padua
Italy
ferro@dei.unipd.it

Anastasia Giachanou
Utrecht University
The Netherlands
a.giachanou@uu.nl

Evangelos Kanoulas
University of Amsterdam
The Netherlands
e.kanoulas@uva.nl

Dan Li
Elsevier
The Netherlands
d.li1@elsevier.com

Theodora Tsikrika
Information Technologies Institute, CERTH
Greece
theodora.tsikrika@iti.gr

Michalis Vlachos
University of Lausanne
Switzerland
michalis.vlachos@unil.ch

Stefanos Vrochidis
Information Technologies Institute, CERTH
Greece
stefanos@iti.gr

Abstract

This is a report on the fourteenth edition of the *Conference and Labs of the Evaluation Forum* (CLEF 2023), held on September 18–21, 2023, in Thessaloniki, Greece. CLEF was a four-day hybrid event combining a conference and an evaluation forum. The conference featured keynotes by Barbara Plank and Claudia Hauff, and presentation of peer-reviewed research papers covering a wide range of topics, in addition to many posters. The evaluation forum consisted of thirteen labs: BioASQ, CheckThat!, DocILE, eRisk, EXIST, iDPP, ImageCLEF, JokeR, LifeCLEF, LongEval, PAN, SimpleText, and Touché, addressing a wide range of tasks, media, languages, and ways to go beyond standard test collections.

Date: 18–21 September, 2023.

Website: https://clef2023.clef-initiative.eu/.

1 Introduction

The 2023 edition of the *Conference and Labs of the Evaluation Forum* (CLEF) was co-organized by the Information Technologies Institute of the Centre for Research and Technology Hellas, Greece, the University of Amsterdam, The Netherlands, and the Democritus University of Thrace, Greece, and was held in Thessaloniki, Greece from September 18 to September 21, 2023. The conference format remained the same as in previous years, and consisted of keynotes, contributed papers, lab sessions, and poster sessions, including reports from other benchmarking initiatives from around the world. All sessions were organized and run in hybrid mode, allowing for both in-presence and remote attendance. CLEF 2023 was the 14th year of the CLEF Conference and the 24th year of the CLEF initiative as a forum for IR Evaluation.

CLEF was established in 2000 as a spin-off of the TREC Cross-Language Track, with a focus on stimulating research and innovation in multimodal and multilingual information access and retrieval [Ferro, 2019; Ferro and Peters, 2019]. Over the years, CLEF has fostered the creation of language resources in many European and non-European languages, promoted the growth of a vibrant and multidisciplinary research community, provided sizable improvements in the performance of monolingual, bilingual, and multilingual information access systems [Ferro and Silvello, 2017], and achieved a substantial scholarly impact [Larsen, 2019; Tsikrika et al., 2011, 2013].

In its first 10 years, CLEF hosted a series of experimental labs that reported their results at an annual workshop held in conjunction with the European Conference on Digital Libraries (ECDL, now TPDL). In 2010, by then a mature and well-respected evaluation forum, CLEF was expanded to include a complementary peer-reviewed conference, focused on discussing the advancement of evaluation methodologies and on reporting evaluations of information access and retrieval systems regardless of data type, format, language, and others. Moreover, the scope of the evaluation labs was broadened, to include not only multilinguality but also multimodality in information access. Multimodality is here intended as the ability to deal with information not only conveyed by multiple media, but also coming in different modalities, e.g. the Web, social media, news streams, specific domains, and so on. Since 2010, the CLEF conference has established a format which includes keynotes, contributed papers, lab sessions, and poster sessions, including reports from other benchmarking initiatives from around the world. Since 2013, CLEF has been supported by an association, a lightweight non-for-profit legal entity that, thanks to the financial support of the CLEF community, takes care of the small central coordination needed to operate CLEF on an ongoing basis and makes it a self-sustaining activity [Ferro, 2019].

CLEF 2023 continued the initiative introduced in the 2019 edition, during which the *European Conference for Information Retrieval (ECIR)* and CLEF joined forces: ECIR 2023 hosted a special session dedicated to CLEF Labs where lab organizers presented the major outcomes of their Labs and their plans for ongoing activities, followed by a poster session to favour discussion during the conference. This was reflected in the ECIR 2023 proceedings, where CLEF Lab activities and results were reported as short papers. The goal was not only to engage the ECIR community in CLEF activities but also to disseminate the research results achieved during CLEF evaluation cycles as submission of papers to ECIR.

CLEF 2023 was attended by 231 participants, out of which 161 in-presence and 70 remotely, denoting a vibrant community, from different academic institutions and industrial organizations.

Although the majority (72%) of the participants came from different European countries, there was also considerable worldwide interest in CLEF 2023, with 12% participants from Asia, 13% from the Americas, 2% from Oceania, and 1% from Africa.

2 The CLEF Conference

CLEF 2023 continued the focus of the CLEF conference on "experimental IR", as carried out at evaluation forums (CLEF Labs, TREC, NTCIR, FIRE, MediaEval, TAC, etc.), with special attention to the challenges of multimodality, multilinguality, and interactive search. We invited submissions on significant new insights demonstrated on IR test collections, on analyses of IR test collections and evaluation measures, and on concrete proposals to push the boundaries of the Cranfield/TREC/CLEF paradigm [Arampatzis et al., 2023].

Keynotes The following scholars were invited to give a keynote talk at the CLEF 2023 conference:

Barbara Plank (LMU Munich, Germany) delivered a talk entitled "Human-centric Natural Language Processing". Here is the abstract of her talk: "Despite the recent success of Natural Language Processing (NLP), driven by advances in large language models (LLMs) trained on enormous amounts of data, there are many challenges ahead to make NLP more human-facing and inclusive. For instance, low-resource languages, non-standard data and dialects pose particular challenges, due to the high variability in language, paired with low availability of data. Moreover, while language varies along many dimensions, evaluation today largely focuses on standard splits, and assumes the existence of a single correct answer. In this talk, I survey some of the challenges, and outline some potential solutions, discussing work on cross-lingual transfer learning, NLP for dialects, and data-centric NLP, which includes learning in light of human label variation."

Claudia Hauff (Spotify, The Netherlands) gave a speech "On the Challenges of Podcast Search at Spotify". Here is the abstract of her talk: "Online music streaming is enjoying ever-growing popularity over the past decades, enabled by the abundance of music content in digital format and online streaming services. In recent years, podcasts (a talk-focused media format) have witnessed a rapid growth among listeners. Podcasts come in many forms and sizes. They range from 20-minute daily meditation sessions, weekly recaps of global news, to interviews with celebrities, and hosts bantering with each other for hours. More and more streaming services are now expanding their catalogs to support both music and podcasts on the same platform. This setup requires an effective aggregated search system to assemble information from heterogeneous information sources and content types into one result interface in order to support diverse information needs. In this talk I present a number of open challenges in this domain."

Other Evaluation Initiatives *Ian Soboroff* (NIST, USA) briefly introduced TREC[1] (Text REtrieval Conference), whose purpose is to support research within the Information Retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. *Noriko Kando* (NII, Japan) presented NTCIR[2] (NII Testbeds and Community for

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

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

Information access Research), which promotes research in information access technologies with a special focus on East-Asian languages and English. *Surupendu Gangopadhyay* (DA-IICT, India) introduced FIRE[3], which fosters the development of multilingual information access systems for the Indian subcontinent and explores new domains like plagiarism detection, legal information access, mixed-script information retrieval, and spoken document retrieval.

Technical Program CLEF 2023 received a total of 14 scientific submissions, of which a total of 11 papers (10 long & 1 short) were accepted. Each submission was reviewed by three program committee members, and the program chairs oversaw the reviewing and follow-up discussions. Ten countries are represented in the accepted papers, several of them being products of international collaboration. This year, researchers addressed the following important challenges in the community: authorship attribution, abusive language detection, machine-generated text detection, medical concept normalization, trend detection in time series data, fashion image captioning, the use of quantum annealing in IR and recommender systems, the impact of evolving evaluation environments on retrieval performance, reproducibility studies, automatic variant interpretation in clinical genetics literature, and multilingual podcast access research.

Similarly to what happened in the previous editions from 2015 onwards, CLEF 2023 invited CLEF 2022 lab organizers to nominate a "best of the labs" paper that was reviewed as a full paper submission to the CLEF 2023 conference, according to the same review criteria and PC. Seven full papers were accepted for this "best of the labs" section.

3 The CLEF Lab Sessions

A total of 15 lab proposals were received and evaluated in peer review based on their innovation potential and the quality of the resources created. To identify the best proposals, well-established criteria from previous editions of CLEF were applied, like, for example, topical relevance, novelty, potential impact on future world affairs, likely number of participants, and the quality of the organizing consortium. This year we further stressed the connection to real-life usage scenarios, and we tried to avoid, as much as possible, overlaps among labs, in order to promote synergies and integration.

The 13 selected labs represented scientific challenges based on new datasets and real-world problems in multimodal and multilingual information access. These datasets provide unique opportunities for scientists to explore collections, to develop solutions for these problems, to receive feedback on the performance of their solutions, and to discuss related challenges with peers at the workshops. In addition to these workshops, the labs reported results of their year-long activities in overview talks and lab sessions.

The 13 labs running as part of CLEF 2023 comprised mainly labs that continued from previous editions at CLEF (BioASQ, CheckThat!, eRisk, iDPP, ImageCLEF, JOKER, LifeCLEF, PAN, SimpleText, and Touché) and new pilot/workshop activities (DocILE, EXIST, and LongEval). Details of the individual labs are described by the lab organizers in the CLEF Working Notes [Aliannejadi et al., 2023]. We only provide a brief overview of them here (in alphabetical order).

[3] http://fire.irsi.res.in/

BioASQ: Large-scale biomedical semantic indexing and question answering[4] [Nentidis et al., 2023] aims to push the research frontier towards systems that use the diverse and voluminous information available online to respond directly to the information needs of biomedical scientists. It offered the following tasks. *Task 1 - b: Biomedical Semantic Question Answering*: benchmark datasets of biomedical questions, in English, along with gold standard (reference) answers constructed by a team of biomedical experts. The participants have to respond with relevant articles, and snippets from designated resources, as well as exact and "ideal" answers. *Task 2 - Synergy: Question Answering for developing problems*: biomedical experts pose unanswered questions for developing problems, such as COVID-19, receive the responses provided by the participating systems, and provide feedback, together with updated questions in an iterative procedure that aims to facilitate the incremental understanding of developing problems in biomedicine and public health. *Task 3 - MedProcNER: Medical Procedure Text Mining and Indexing Shared Task*: focuses on the recognition and indexing of medical procedures in clinical documents in Spanish posing subtasks on (1) indexing medical documents with controlled terminologies, (2) automatic detection indexing textual evidence, i.e. medical procedure entity mentions in text, and (3) normalization of these medical procedure mentions to terminologies.

CheckThat!: Check-Worthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and their Sources[5] [Barrón-Cedeño et al., 2023] aims at producing technology to support the fight against misinformation and disinformation in social media, in political debates and in the news with a focus on check-worthiness, subjectivity, bias, factuality, and authority of the claim. It offered the following tasks. *Task 1 - Check-worthiness in textual and multimodal content*: determine whether an item, be it a text alone or a text plus an image deserves the attention of a journalist to be fact-checked.*Task 2 - Subjectivity in News Articles*: assess whether a text snippet within a news article is subjective or objective. *Task 3 - Political Bias of News Articles and News Media*: identify the political leaning of an article or media source: left, centre or right. *Task 4 - Factuality of Reporting of News Media*: determine the level of factuality of both a document and a medium. *Task 5 - Authority Finding in Twitter*: identify authorities that should be trusted to verify a contended claim expressed in an Arabic tweet.

DocILE: Document Information Localization and Extraction[6] [Šimsa et al., 2023] runs the largest benchmark for the tasks of Key Information Localization and Extraction (KILE) and Line Item Recognition (LIR) from business documents like invoices. It offered the following tasks. *Task 1 - Key Information Localization and Extraction (KILE)*: localize fields of each pre-defined category and read out their values. *Task 2 - Line Item Recognition (LIR)*: find all line items, e.g., a billed item in a table, and localize their corresponding fields in the document as in Task 1.

[4] http://www.bioasq.org/workshop2023

[5] http://checkthat.gitlab.io/

[6] https://docile.rossum.ai/

eRisk: Early Risk Prediction on the Internet[7] [Parapar et al., 2023] explores the evaluation methodology, effectiveness metrics, and practical applications (particularly those related to health and safety) of early risk detection on the Internet. Early detection technologies can be employed in different areas, particularly those related to health and safety. For instance, early alerts could be sent when a predator starts interacting with a child for sexual purposes, or when a potential offender starts publishing antisocial threats on a blog, forum or social network. Our main goal is to pioneer a new interdisciplinary research area that would be potentially applicable to a wide variety of situations and to many different personal profiles. Examples include potential paedophiles, stalkers, individuals that could fall into the hands of criminal organisations, people with suicidal inclinations, or people susceptible to depression. It offered the following tasks. *Task 1 - Search for symptoms of depression*: the challenge consists of ranking sentences from a collection of user writings according to their relevance to a depression symptom. The participants had to provide rankings for the 21 symptoms of depression from the BDI Questionnaire. A sentence is deemed relevant to a BDI symptom when it conveys information about the user's state concerning the symptom. That is, it may be relevant even when it indicates that the user is OK with the symptom. *Task 2 - Early Detection of Signs of Pathological Gambling*: the challenge consists of sequentially processing pieces of evidence and detect early traces of pathological gambling (also known as compulsive gambling or disordered gambling), as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates on texts written in Social Media *Task 3 - Measuring the severity of the signs of Eating Disorders*: the task consists of estimating the level of features associated with a diagnosis of eating disorders from a thread of user submissions. For each user, the participants were given a history of postings and the participants had to fill a standard eating disorder questionnaire (based on the evidence found in the history of postings).

EXIST: sEXism Identification in Social neTworks[8] [Plaza et al., 2023] aims to capture and categorize sexism, from explicit misogyny to other subtle behaviors, in social networks. Participants were asked to classify tweets in English and Spanish according to the type of sexism they enclose and the intention of the persons that writes the tweets. It offered the following tasks. *Task 1 - Sexism Identification*: is a binary classification tasks. The systems have to decide whether or not a given tweet contains or describes sexist expressions or behaviors (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behavior). *Task 2 - Source Intention*: aims to categorize the sexist messages according to the intention of the author in one of the following categories: (i) direct sexist message, (ii) reported sexist message and (iii) judgemental message. *Task 3 - Sexism Categorization*: is a multiclass task that aims to categorize the sexist messages according to the type or types of sexism they contain (according to the categorization proposed by experts and that takes into account the different facets of women that are undermined): (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence.

[7] https://erisk.irlab.org/

[8] http://nlp.uned.es/exist2023/

iDPP: Intelligent Disease Progression Prediction[9] [Faggioli et al., 2023] aims to design and develop an evaluation infrastructure for AI algorithms able to: (1) better describe mechanism of the Amyotrophic Lateral Sclerosis (ALS) disease; (2) stratify patients according to their phenotype assessed all over the disease evolution; and (3) predict ALS progression in a probabilistic, time dependent fashion. It offered the following tasks. *Task 1 – Predicting Risk of Disease Worsening (Multiple Sclerosis)*: focuses on ranking subjects based on the risk of worsening, setting the problem as a survival analysis task. More specifically the risk of worsening predicted by the algorithm should reflect how early a patient experience the event ”worsening”. Worsening is defined based on the Expanded Disability Status Scale (EDSS), accordingly to clinical standards. *Task 2 – Predicting Probability of Worsening (Multiple Sclerosis)*: refines Task 1 asking participants to explicitly assign a probability of worsening at different time windows (e.g. between years 4 and 6, 6 and 8, 8 and 10 etc.). *Task 3 – Impact of Exposition to Pollutants (Amyotrophic Lateral Sclerosis)*: evaluates proposals of different approaches to assess if exposure to different pollutants is a useful variable to predict time to Percutaneous Endoscopic Gastrostomy (PEG), Non-Invasive Ventilation (NIV) and death in ALS patients.

ImageCLEF: Multimedia Retrieval[10] [Ionescu et al., 2023] is set to promote the evaluation of technologies for annotation, indexing, classification and retrieval of multimodal data, with the objective of providing information access to large collections in various usage scenarios and domains. It offered the following tasks. *Task 1 - ImageCLEFmedical*: continues the tradition of bringing together several initiatives for medical applications fostering cross-exchanges, namely: medical concept detection and caption prediction, synthetic medical images generated with GANs, Visual Question Answering and generation, and doctor-patient conversation summarization. *Task 2 - ImageCLEFaware*: the images available on social networks can be exploited in ways users are unaware of when initially shared, including situations that have serious consequences for the users' real lives. The task addresses the development of algorithms which raise the users' awareness about real-life impact of online image sharing. *Task 3 - ImageCLEFfusion*: despite the current advances in knowledge discovery, single learners do not produce satisfactory performances when dealing with complex data, such as class imbalance, high-dimensionality, concept drift, noise, multimodality, subjective annotations, etc. This task aims to fill this gap by exploiting novel and innovative late fusion techniques for producing a powerful learner based on the expertise of a pool of classifiers. *Task 4 - ImageCLEFrecommendation*: focuses on content-recommendation for cultural heritage content in 15 broad themes that have been curated by experts in the Europeana Platform. Despite current advances, there is limited understanding how well these perform and how relevant they are for the final end-users.

JOKER: Automatic Wordplay Analysis[11] [Ermakova et al., 2023a] aims to create reusable test collections for benchmarking and to explore new methods and evaluation metrics for the automatic processing of wordplay. It offered the following tasks. *Task 1 - Pun detection*:

[9] https://brainteaser.health/open-evaluation-challenges/idpp-2023/

[10] https://www.imageclef.org/2023

[11] http://joker-project.com/

detection of puns in English, French, and Spanish. *Task 2 - Pun interpretation*: interpretation of puns in English, French, and Spanish. *Task 3 - Pun translation*: translation of puns from English to French and Spanish.

LifeCLEF: Multimedia Retrieval in Nature[12] [Joly et al., 2023] is dedicated to the large-scale evaluation of biodiversity identification and prediction methods based on artificial intelligence. It offered the following tasks. *Task 1 - BirdCLEF*: bird species recognition in audio soundscapes. *Task 2 - FungiCLEF*: fungi recognition from images and metadata. *Task 3 - GeoLifeCLEF*: remote sensing based prediction of species. *Task 4 - PlantCLEF*: global-scale plant identification from images. *Task 5 - SnakeCLEF*: snake species identification in medically important scenarios.

LongEval: Longitudinal Evaluation of Model Performance[13] [Alkhalifa et al., 2023] is focused on evaluating the temporal persistence of information retrieval systems and text classifiers. The goal is to develop temporal information retrieval systems and longitudinal text classifiers that survive through dynamic temporal text changes, introducing time as a new dimension for ranking models performance. It offered the following tasks. *Task 1 - LongEval-Retrieval*: aims to propose a temporal information retrieval system which can handle changes over the time. The proposed retrieval system should follow the temporal persistence on Web documents. This task had 2 sub-tasks focusing on short-term and long-term persistence. *Task 2 - LongEval-Classification* aims to propose a temporal persistence classifier which can mitigate performance drop over short and long periods of time compared to a test set from the same time frame as training. This task had 2 sub-tasks focusing on short-term and long-term persistence.

PAN: Digital Text Forensics and Stylometry[14] [Bevendorff et al., 2023] aims to advance the state of the art and provide for an objective evaluation on newly developed benchmark datasets in those areas. It offered the following tasks. *Task 1 - Cross-Discourse Type Authorship Verification*: focuses on (cross-discourse type) authorship verification where both written (e.g., essays, emails) and oral language (e.g., interviews, speech transcriptions) are represented in the set of discourse types. *Task 2 - Profiling Cryptocurrency Influencers with Few-shot Learning*: aims to profile cryptocurrency influencers in social media (Twitter) from a low-resource perspective. *Task 3 - Multi-Author Writing Style Analysis*: addresses multi-authored documents whose authorship cannot be easily determined by exploiting topic changes alone. *Task 4 - Trigger Detection*: addresses the task of assigning a single trigger warning label (violence) to narratives in a corpus of fanfiction.

SimpleText: Automatic Simplification of Scientific Texts[15] [Ermakova et al., 2023b] aims to create a simplified summary of multiple scientific documents based on a popular science query which provides a user with an instant accessible overview on this specific topic. It offered the following tasks. *Task 1 - What is in, or out?*: selecting passages to include

[12] http://www.lifeclef.org/

[13] https://clef-longeval.github.io/

[14] http://pan.webis.de/

[15] http://simpletext-project.com/

in a simplified summary. *Task 2 - What is unclear?*: difficult concept identification and explanation. *ask 3 - Rewrite this!*: rewriting scientific text.

Touché: Argument and Causal Retrieval[16] [Bondarenko et al., 2023] aims to foster and support the development of technologies for argument and causal retrieval and analysis that includes argument quality estimation, stance detection, image retrieval, and causal evidence retrieval. It offered the following tasks. *Task 1 - Argument Retrieval for Controversial Questions*: given a controversial topic and a collection of web documents, the task is to retrieve and rank documents by relevance to the topic, by argument quality, and to detect the document stance. *Task 2 - Evidence Retrieval for Causal Questions*: given a causality-related topic and a collection of web documents, the task is to retrieve and rank documents by relevance to the topic and detect the document "causal" stance (i.e., whether a causal relationship from the title of the topic holds). *Task 3 - Image Retrieval for Arguments*: given a controversial topic, the task is to retrieve images (from web pages) for each stance (pro/con) that show support for that stance. *Task 4 - Intra-Multilingual Multi-Target Stance Classification*: given a proposal on a socially important issue, its title, and topic in different languages, the task is to classify whether a comment is in favor, against, or neutral towards the proposal.

More information on the CLEF 2023 conference, the CLEF initiative and the CLEF Association is provided on the Web:

- CLEF 2023: https://clef2023.clef-initiative.eu/
- CLEF initiative: https://www.clef-initiative.eu/
- CLEF Association: https://www.clef-initiative.eu/#association

4 Overall Trends for CLEF

Figure 1 shows the attendance trends to CLEF since its inception. We can observe that there has been a substantial growth over the years, especially since when it was backed by the CLEF Association. We can also note that CLEF 2020 and CLEF 2021, which were online only and with almost free registration due to COVID-19, represent a spike in the attendance. The in-presence attendance for CLEF 2023 has been substantially comparable to the pre-COVID editions, slightly growing with respect to CLEF 2022 (161 vs 143 in-presence participants), while the overall participation has increased compared to the pre-COVID editions, thanks to the remote participants.

Figure 2 shows the number of papers published in the Working Notes over the years; we report the Working Notes because they contain both the labs overviews and all the participant papers. We can observe how the increase in participation to CLEF has been accompanied by an increase in the publication output. Note that both the Working Notes and the Conference Proceedings are fully peer-reviewed venues.

[16] https://touche.webis.de/

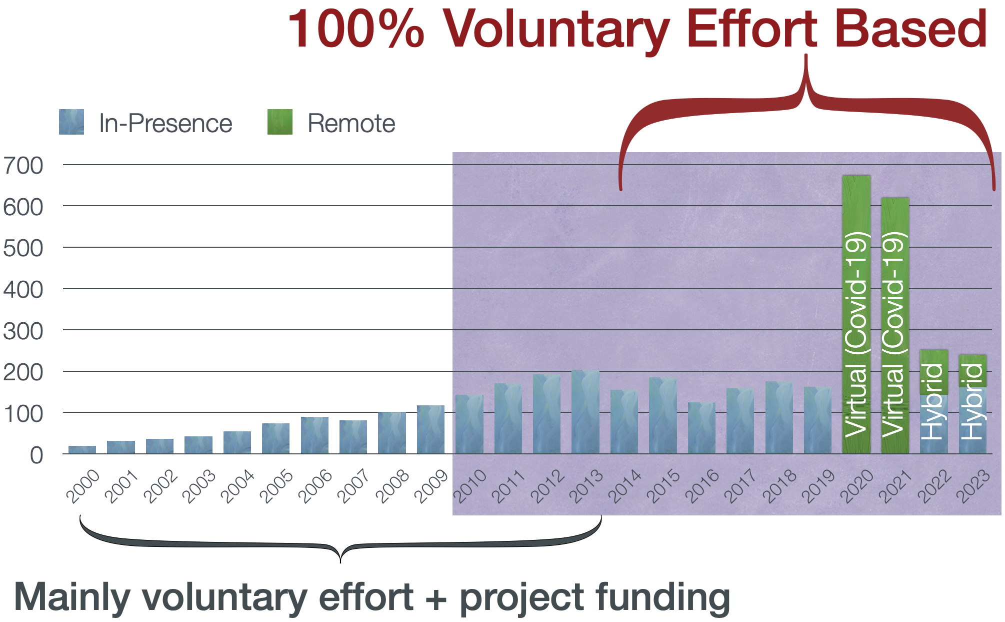

Figure 1. Attendance to CLEF over the years: x-axis reports CLEF editions; y-axis the number of attendees; the shading indicates the change from CLEF as a workshop, co-located with ECDL/TPDL, to CLEF as an independent conference and labs.

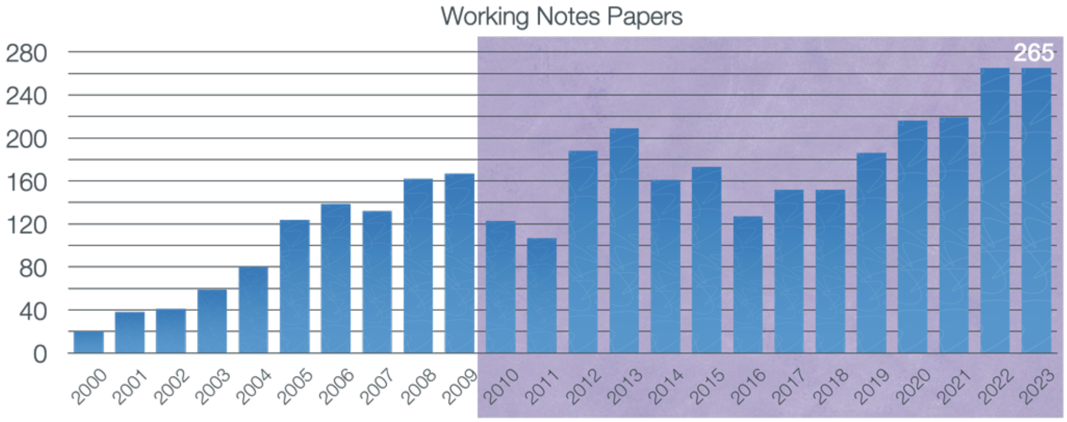

Figure 2. Papers published in the working notes over the years: x-axis reports CLEF editions; y-axis the number of papers in the working notes, highlighting 265, the papers in the CLEF 2023 working notes; the shading indicates the change from CLEF as a workshop co-located with ECDL/TPDL to CLEF as an independent conference and labs.

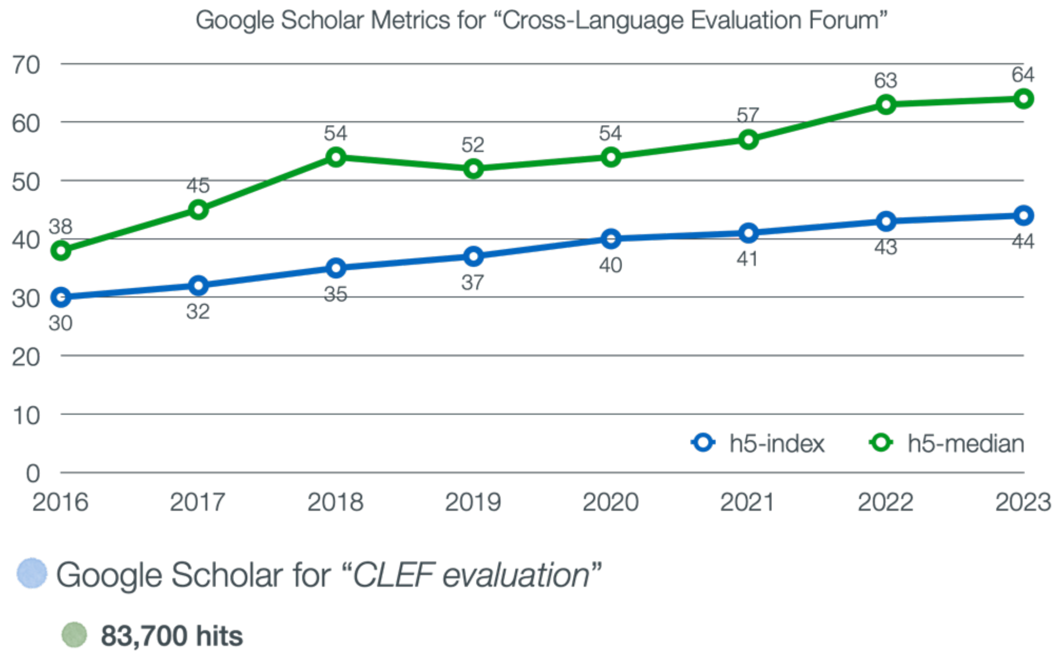

Figure 3. Google Scholar metrics for “Cross-Language Evaluation Forum” since 2016: the x-axis reports years, the y-axis the value for the h5-index (the largest number h such that at least h articles in that publication were cited at least h times each, only those of its articles that were published in the last five complete calendar years) and h5-median (the median number of citations for the articles that make up the h5-index).

Finally, Figure 3 shows the Google Scholar metrics for CLEF[17] since 2016; also in this case we can observe a positive growth trend, giving an idea of the impact of CLEF. In particular, CLEF is listed among the top-20 venues for the sub-category "Databases & Information Systems"[18], together with other important venues for the IR community, like SIGIR, CIKM, RecSys, and WWW.

5 CLEF 2024

CLEF 2024 will be hosted by University of Grenoble Alpes, Grenoble, France, on 9–12 September 2024.

More information on CLEF 2024, the call for papers and the ongoing labs is available at:

- https://clef2024.clef-initiative.eu/

As far as labs are concerned, CLEF 2024 will run 15 evaluation activities out of 25 proposals received: twelve will be a continuation of the labs running during CLEF 2023:

- BioASQ – Large-scale biomedical semantic indexing and question answering[19];
- CheckThat! – Check-Worthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Sources[20];
- eRisk – Early risk prediction on the Internet[21];
- EXIST – sEXism Identification in Social neTworks[22];
- iDPP – Intelligent Disease Progression Prediction[23];
- ImageCLEF – Multimedia Retrieval Challenge[24];
- JokeR – Automatic Wordplay Analysis[25];
- LifeCLEF – Multimedia Retrieval in Nature[26];
- LongEval – Longitudinal Evaluation of Model Performance[27].
- PAN – Digital Text Forensics and Stylometry[28];
- SimpleText – Automatic Simplification of Scientific Texts[29];
- Touché – Argument and Causal Retrieval[30];

and three will be new pilot labs:

[17] Note that Google Scholar still indexes CLEF as "Cross-Language Evaluation Forum", even if the name has changed to "Conference and Labs of the Evaluation Forum" since 2010.

[18] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_databasesinformationsystems

[19] http://www.bioasq.org/workshop2024

[20] http://checkthat.gitlab.io

[21] https://erisk.irlab.org/

[22] http://nlp.uned.es/exist2024/

[23] https://brainteaser.health/open-evaluation-challenges/idpp-2024/

[24] https://www.imageclef.org/2024

[25] https://www.joker-project.com/

[26] http://www.lifeclef.org

[27] https://clef-longeval.github.io/

[28] https://pan.webis.de/

[29] https://simpletext-project.com/

[30] https://touche.webis.de/

- ELOQUENT – Quality of generative language models[31].
  ELOQUENT provides a set a of tasks for evaluating the quality of generative language models. It focuses on several aspects: (i) topical competence, Can an LLM assess itself if it is capable to process data in some application domain of interest? (ii) hallucination, Can an LLM be used to evaluate the output of other LLMs to detect hallucinated or factually incorrect information? (iii) robustness: Will an LLM output the same content independent of input variation which is equivalent in content but non-identical in form or style; (iv) Voight-Kampff test, Can an LLM be used to detect if some piece of text is written by a humn author or generated by an LLM.
- MonsterCLEF – One Lab to Rule Them All[32].
  The MonsterCLEF lab is organized as a meta-challenge across a selection of tasks chosen from the other labs running in CLEF 2024 and participants are asked to develop a generative AI/LLM-based system that will be run against all the selected tasks with no or minimal adaptation. For each targeted task we rely on the same dataset, experimental setting, and evaluation measures adopted for that specific task. In this way, the LLM-based systems participating in the MonsterCLEF lab are directly comparable with the specialized systems participating in each targeted task.
- QuantumCLEF – Quantum Computing at CLEF[33].
  The objective of QuantumCLEF is to investigate how algorithms for Information Retrieval and Recommender Systems can be formulated for and executed on a quantum computer. QuantumCLEF offers an infrastructure for developing, executing, and evaluating quantum computing algorithms and, in particular, Quantum Annealing (QA) algorithms, in order to compare their performance – both efficiency and effectiveness – with respect to traditional solutions.

6 Bids for CLEF 2026

Bids for hosting CLEF 2026 are now open and close on December 2023. Proposals can be sent to the CLEF Steering Committee Chair at chair@clef-initiative.eu and a template for bids is available here https://www.clef-initiative.eu/assets/CLEF-Template_for_bids.docx.

Acknowledgments

The success of CLEF 2023 would not have been possible without the huge effort of several people and organizations, including the CLEF Association, the program committee, the lab organizing committee, the local organization committee in Thessaloniki, the reviewers, and the many students and volunteers who contributed along the way.

We gratefully acknowledge the support we received from our supporters and sponsors: We thank the Information Technologies Institute (ITI) of the Centre for Research and Technology

[31] https://eloquent-lab.github.io/

[32] https://monsterclef.dei.unipd.it/

[33] https://qclef.dei.unipd.it/

Figure 4. CLEF group photo taken at the closing session of CLEF 2023

Hellas (CERTH) in Thessaloniki, Greece (with special mention to the Multimodal Data Fusion & Analytics Group of the Multimedia Knowledge and Social Media Analytics Laboratory)[34] for their invaluable support. We also thank the Friends of SIGIR program[35] and the Information Retrieval Specialist Group (IRSG) of the BCS[36] for covering the registration fees for a number of student delegates.

Last but not least, without the important and tireless effort of the enthusiastic and creative authors, the organizers of the selected labs, the colleagues and friends involved in running them, and the participants who contribute their time to making the labs and the conference a success, as well as financially supporting them through the CLEF Association, CLEF would not be possible.

Thank you all very much!

References

M. Aliannejadi, G. Faggioli, N. Ferro, and M. Vlachos, editors. *CLEF 2023 Working Notes*, 2023. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, https://ceur-ws.org/Vol-3497/.

R. Alkhalifa, I. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, G. Gonzalez-Saez, P. Galuščáková, L. Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, P. Mulhem, F. Piroi, M. Popel, C. Servan, H. Tayyar Madabushi, and A. Zubiaga. Overview of the CLEF-2023 LongEval Lab on Longitudinal Evaluation of Model Performance. In Arampatzis et al. [2023], pages 440–458.

[34] https://m4d.iti.gr/; https://mklab.iti.gr/; https://www.iti.gr/; https://www.certh.gr/
[35] http://sigir.org/general-information/funding-for-sigir-related-events/
[36] http://irsg.bcs.org/

A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, A. Aliannejadi, M. Vlachos, G. Faggioli, and N. Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, 2023. Lecture Notes in Computer Science (LNCS) 14163, Springer, Heidelberg, Germany.

A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, T. Elsayed, D. Azizov, T. Caselli, G. S. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, and W. Zaghouani. Overview of the CLEF–2023 CheckThat! Lab on Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Source. In Arampatzis et al. [2023], pages 251–275.

J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, and E. Zangerle. Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection. In Arampatzis et al. [2023], pages 459–481.

A. Bondarenko, M. Fröbe, J. Kiesel, F. Schlatt, V. Barriere, B. Ravenet, L. Hemamou, S. Luck, J. H. Reimer, B. Stein, M. Potthast, and M. Hagen. Overview of Touché 2023: Argument and Causal Retrieval. In Arampatzis et al. [2023], pages 507–530.

L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, and A. Jatowt. Overview of JOKER – CLEF-2023 Track on Automatic Wordplay Analysis. In Arampatzis et al. [2023], pages 397–415.

L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, O. Augereau, and J. Kamps. Overview of the CLEF 2023 SimpleText Lab: Automatic Simplification of Scientific Texts. In Arampatzis et al. [2023], pages 482–506.

G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Dominguez, M. Gromicho, E. Longato, S. C. Madeira, U. Manera, G. Silvello, E. Tavazzi, E. Tavazzi, M. Vettoretti, B. Di Camillo, and N. Ferro. Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023. In Arampatzis et al. [2023], pages 343–369.

N. Ferro. What Happened in CLEF... For a While? In F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)*, pages 3–45. Lecture Notes in Computer Science (LNCS) 11696, Springer, Heidelberg, Germany, 2019.

N. Ferro and C. Peters, editors. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, 2019. Springer International Publishing, Germany.

N. Ferro and G. Silvello. 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc-*ish* CLEF? *Information Processing & Management*, 53(1):175–202, January 2017.

B. Ionescu, H. Müller, A.-M. Drăgulinescu, W.-W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. M. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A.-G. Andrei, I. Coman, V. Kovalev, A. Radzhabov, Y. Prokopchuk, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, J. Deshayes, and A. Popescu. Overview of the ImageCLEF 2023: Multimedia Retrieval in Medical, Social Media and Internet Applications. In Arampatzis et al. [2023], pages 370–396.

A. Joly, C. Botella, L. Picek, S. Kahl, H. Goëau, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, R. Chamidullin, M. Šulc, M. Hrúz, M. Servajean, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, and H. Müller. Overview of LifeCLEF 2023: Evaluation of AI Models for the Identification and Prediction of Birds, Plants, Snakes and Fungi. In Arampatzis et al. [2023], pages 416–439.

B. Larsen. The Scholarly Impact of CLEF 2010-2017. In Ferro and Peters [2019], pages 547–554.

A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gascó, M. Krallinger, and G. Paliouras. Overview of BioASQ 2023: The Eleventh BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In Arampatzis et al. [2023], pages 227–250.

J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani. Overview of eRisk 2023: Early Risk Prediction on the Internet. In Arampatzis et al. [2023], pages 294–315.

L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. In Arampatzis et al. [2023], pages 316–342.

S. Šimsa, M. Uřičář, M. Šulc, Y. Patel, A. Hamdi, M. Kocián, M. Skalický, J. Matas, A. Doucet, M. Coustaty, and D. Karatzas. Overview of DocILE 2023: Document Information Localization and Extraction. In Arampatzis et al. [2023], pages 276–293.

T. Tsikrika, A. Garcia Seco de Herrera, and H. Müller. Assessing the Scholarly Impact of ImageCLEF. In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, and M. de Rijke, editors, *Multilingual and Multimodal Information Access Evaluation. Proceedings of the Second International Conference of the Cross-Language Evaluation Forum (CLEF 2011)*, pages 95–106. Lecture Notes in Computer Science (LNCS) 6941, Springer, Heidelberg, Germany, 2011.

T. Tsikrika, B. Larsen, H. Müller, S. Endrullis, and E. Rahm. The Scholarly Impact of CLEF (2000–2009). In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the Fourth International Conference of the CLEF Initiative (CLEF 2013)*, pages 1–12. Lecture Notes in Computer Science (LNCS) 8138, Springer, Heidelberg, Germany, 2013.