Report on the 13th Conference and Labs of the Evaluation Forum (CLEF 2022): Experimental IR Meets Multilinguality, Multimodality, and Interaction

Alberto Barrón-Cedeño
University of Bologna
Italy
a.barron@unibo.it

Giovanni Da San Martino
University of Padua
Italy
giovanni.dasanmartino@unipd.it

Mirko Degli Esposti
University of Bologna
Italy
mirko.degliesposti@unibo.it

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

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

Allan Hanbury
Vienna Univ. of Technology
Austria
allan.hanbury@tuwien.ac.at

Craig Macdonald
University of Glasgow
UK
craig.macdonald@glasgow.ac.uk

Gabriella Pasi
University of Milan Bicocca
Italy
gabriella.pasi@unimib.it

Martin Potthast
Leipzig University
Germany
martin.potthast@uni-leipzig.de

Fabrizio Sebastiani
National Council of Research, ISTI CNR
Italy
fabrizio.sebastiani@isti.cnr.it

Abstract

This is a report on the thirteenth edition of the *Conference and Labs of the Evaluation Forum* (CLEF 2022), held on September 5–8, 2022, in Bologna, Italy. CLEF was a four-day hybrid event combining a conference and an evaluation forum. The conference featured keynotes by Benno Stein and Rita Cucchiara, and presentation of peer-reviewed research papers covering a wide range of topics, in addition to many posters. The evaluation forum consisted of fourteen labs: ARQMath, BioASQ, CheckThat!, ChEMU, eRisk, HIPE, iDPP, ImageCLEF, JokeR, LeQua, LifeCLEF, PAN, SimpleText, and Touché, addressing a wide range of tasks, media, languages, and ways to go beyond standard test collections.

Date: 5–8 September, 2022.

Website: https://clef2022.clef-initiative.eu/.

1 Introduction

The 2022 edition of the *Conference and Labs of the Evaluation Forum* (CLEF) was hosted by the University of Bologna, Italy from September 5 to September 8, 2022. 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 2022 was the 13th year of the CLEF Conference and the 23rd 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 2022 continued the initiative introduced in the 2019 edition, during which the *European Conference for Information Retrieval (ECIR)* and CLEF joined forces: ECIR 2022 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 2022 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 2022 was attended by 251 participants, out of which 143 in-presence and 108 remotely, denoting a vibrant community, from different academic institutions and industrial organizations. Although the majority (73%) of the participants came from different European countries, there

was also considerable worldwide interest in CLEF 2022, with 12% participants from Asia, 12% from the Americas, and 3% from Oceania.

2 The CLEF Conference

CLEF 2022 continued the focus of the CLEF conference on "experimental IR", as carried out at evaluation forums (CLEF Labs, TREC, NTCIR, FIRE, MediaEval, RomIP, 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 [Barrón-Cedeño et al., 2022].

2.1 Keynotes

The following scholars were invited to give a keynote talk at the CLEF 2022 conference:

Benno Stein (Bauhaus-Universität Weimar, Germany) delivered a talk entitled "[Perceived] Limits in Information Retrieval". Here is the abstract of this talk: "Can information retrieval systems satisfy the desire for unbiased and unframed information? Moreover, should information retrieval systems satisfy the desire for unbiased and unframed information? And finally, do users of information retrieval systems have such a desire in the first place? We start with these questions to discuss and shed light on recent and related developments in information retrieval, such as conversational search, argumentative search, direct answers, information quality labels, and ideas to quantify bias."

Rita Cucchiara (Università degli Studi di Modena e Reggio Emilia, Italy) gave a speech on "A journey into image captioning research". Here is the abstract of her talk: "Image captioning is a fashionable research field in AI: as in Neuroscience, only recently the link between human vision and language generation has been clarified. Also in Deep Learning recent architectures, both recurrent and self-attentive, have shown their capabilities in multimodal understanding and generation. Image captioning is indeed the task of describing the visual content of an image in natural language, employing a visual understanding system and a language model capable of generating meaningful and syntactically correct sentences. Research is explored in different directions: on the one hand, it is going toward very large-scale foundation models, with large-scale parameters and very large web-based supervised datasets; on the other hand, new paradigms mixing generative self-attentive architectures and information retrieval explore new problems as for instance zero-shot learning, long-tail concept description enriched by proper names of persons, places and events. This talk presented a brief overview of recent research results and some architecture models developed at AImagelab, University of Modena and Reggio Emilia, for controllable captioning, universal captioning and retrieval-augmented captioning and discussed possible applications in e-commerce, robotics and web mining, supported by Italian, European and PNRR cofounded projects."

2.2 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 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. Finally, *Gareth Jones* (Dublin City University, Ireland) presented MediaEval[4], the benchmarking initiative for the evaluation of multimedia retrieval, including speech, audio, visual content, tags, users, and context.

2.3 Technical Program

CLEF 2022 received a total of 14 scientific submissions, of which a total of 10 papers (7 long & 3 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, fake news detection and news tracking, noise detection in automatically transferred relevance judgments, impact of online education on children's conversational search behaviour, analysis of multi-modal social media content, knowledge graphs for sensitivity identification, a fusion of deep learning and logic rules for sentiment analysis, medical concept normalization, and domain-specific information extraction.

Similarly to what happened in the previous editions from 2015 onwards, CLEF 2022 invited CLEF 2021 lab organizers to nominate a "best of the labs" paper that was reviewed as a full paper submission to the CLEF 2022 conference, according to the same review criteria and PC. 7 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.

[1] https://trec.nist.gov/
[2] http://research.nii.ac.jp/ntcir/
[3] http://fire.irsi.res.in/
[4] http://multimediaeval.org/

The 14 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 14 labs running as part of CLEF 2022 comprised mainly labs that continued from previous editions at CLEF (ARQMath, BioASQ, CheckThat!, CheMU, eRisk, ImageCLEF, LifeCLEF, PAN, SimpleText, and Touché) and new pilot/workshop activities (HIPE, iDPP, JOKER, and LeQua). Details of the individual labs are described by the lab organizers in the CLEF Working Notes [Faggioli et al., 2022]. We only provide a brief overview of them here (in alphabetical order).

ARQMath: Answer Retrieval for Mathematical Questions[5] [Mansouri et al., 2022] aims to advance math-aware search and the semantic analysis of mathematical notation and texts. It offered the following tasks. Task 1: Answer Retrieval, given a math question post, return relevant answer posts. Task 2: Formula Retrieval; given a formula in a math question post, return relevant formulas from both question and answer posts. Task 3: Open Domain Question Answering, given a math question post, return an automatically generated answer that comprises excerpts from arbitrary sources, and/or machine generated.

BioASQ: Large-scale biomedical semantic indexing and question answering[6] [Nentidis et al., 2022] 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: Large-Scale Online Biomedical Semantic Indexing, it classifies new PubMed documents, before PubMed curators annotate (in effect, classify) them manually into classes from the MeSH hierarchy. Task 2: Biomedical Semantic Question Answering, it uses 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 3 - DisTEMIST: Disease Text Mining and Indexing Shared Task, it focuses on the recognition and indexing of diseases in medical documents in Spanish, by posing subtasks on (1) indexing medical documents with controlled terminologies; (2) automatic detection indexing textual evidence (i.e. disease entity mentions in text); and (3) normalization of these disease mentions to terminologies. Task 4 - Task Synergy: Question Answering for developing problems, biomedical experts pose unanswered questions for the developing problem of 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 COVID-19.

CheckThat!: Lab on Fighting the COVID-19 Infodemic and Fake News Detection[7] [Nakov et al., 2022] aims at fighting misinformation and disinformation in social media,

[5] https://www.cs.rit.edu/~dprl/ARQMath
[6] http://www.bioasq.org/workshop2022
[7] https://sites.google.com/view/clef2022-checkthat

in political debates and in the news, with focus on three tasks (in seven languages: Arabic, Bulgarian, Dutch, English, German, Spanish, and Turkish). It offered the following tasks. Task 1: Identifying Relevant Claims in Tweets, it focuses on disinformation related to the ongoing COVID-19 infodemic politics. It asks to identify which posts in a Twitter stream are worth fact-checking, contain a verifiable factual claim, are harmful to the society, and why. This task was offered in Arabic, Bulgarian, Dutch, English, Spanish, and Turkish. Task 2: Detecting Previously Fact-Checked Claims, given a check-worthy claim, and a set of previously-checked claims, determine whether the claim has been previously fact-checked with respect to a collection of fact-checked claims. The text can be a tweet or a sentence from a political debate. The task is offered in Arabic and English. Task 3: Fake News Detection, given the text and the title of a news article, determine whether the main claim made in the article is true, partially true, false, or other (e.g., articles in dispute and unproven articles). This task is offered in English and German.

ChEMU: Cheminformatics Elsevier Melbourne University[8] [Li et al., 2022] focuses on information extraction in chemical patents, including five tasks ranging from document- to expression-level. It offered the following tasks. Task 1a: Named entity recognition, it aims to identify chemical compounds, their specific types, temperatures, reaction times, yields, and the label of the reaction. Task 1b: Event extraction, a chemical reaction leading to an end product often consists of a sequence of individual event steps. The task is to identify those steps which involve chemical entities recognized from Task 1a. Task 1c: Anaphora resolution, it requires the resolution of anaphoric dependencies between expressions in chemical patents. The participants are required to find five types of anaphoric relationships in chemical patents: coreference, reaction-associated, work-up, contained, and transform. Task 2a: Chemical reaction reference resolution, given a reaction description, this task requires identifying references to other reactions that the reaction relates to, and to the general conditions that it depends on. Task 2b: Table semantic classification, it is about classifying tables in chemical patents into 8 categories based on their contents.

eRisk: Early Risk Prediction on the Internet[9] [Parapar et al., 2022] explores the evaluation methodology, effectiveness metrics, and practical applications (particularly those related to health and safety) of early risk detection on the Internet. The 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: 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 2: Early Detection of Depression, the challenge consists of sequentially processing pieces of evidence and detect early traces of depression as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates

[8] http://chemu2022.eng.unimelb.edu.au/
[9] https://erisk.irlab.org/

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 are given a history of postings and the participants have to fill a standard eating disorder questionnaire (based on the evidence found in the history of postings).

HIPE: Named Entity Recognition and Linking in Multilingual Historical Documents[10] [Ehrmann et al., 2022] focuses on named entity recognition and linking in historical documents, with the objective of assessing and advancing the development of robust, adaptable, and transferable named entity processing systems. Compared to the first HIPE edition in 2020, HIPE 2022 confronts systems with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation schemas. It offered the following tasks. Task 1: Named Entity Recognition and Classification (NERC), with two subtasks: NERC-coarse on high-level entity types, for all languages and NERC-fine on finer-grained entity types, for English, French, and German only. Task 2: Named Entity Linking (EL), Or the linking of named entity mentions to a unique referent in a knowledge base (Wikidata) or to a NIL node if the mention does not have a referent in the KB.

iDPP: Intelligent Disease Progression Prediction[11] [Guazzo et al., 2022] 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: Ranking Risk of Impairment, it focuses on ranking of patients based on the risk of impairment in specific domains. It uses the ALSFRS-R scale to monitor speech, swallowing, handwriting, dressing/hygiene, walking and respiratory ability in time and asks participants to rank patients based on time to event risk of experiencing impairment in each specific domain. Task 2: Predicting Time of Impairment, it refines Task 1 asking participants to predict when specific impairments will occur (i.e. in the correct time-window) by assessing model calibration in terms of the ability of the proposed algorithms to estimate a probability of an event close to the true probability within a specified time-window. Task 3: Explainability of AI algorithms, it calls for position papers to start a discussion on AI explainability including proposals on how the single patient data can be visualized in a multivariate fashion contextualizing its dynamic nature and the model predictions together with information on the predictive variables that most influence the prediction.

ImageCLEF: Multimedia Retrieval[12] [Ionescu et al., 2022] is set to promote the evaluation of technologies for annotation, indexing, classification and retrieval of multi-modal data, with the objective of providing information access to large collections of images in various usage scenarios and domains. It offered the following tasks. Task 1: ImageCLEFmedical, it focuses on interpreting and summarizing the insights gained from radiology images, i.e. develop systems that are able to predict the UMLS concepts from visual image content, and

[10] https://hipe-eval.github.io/HIPE-2022/
[11] https://brainteaser.health/open-evaluation-challenges/idpp-2022/
[12] https://www.imageclef.org/2022

implementing models to predict captions for given radiology images. The tuberculosis task fosters systems that are expected to detect cavern regions localization rather than simply provide a label for the CT images. Task 2: ImageCLEFcoral, it fosters tools for creating 3-dimensional models of underwater coral environments. It requires participants to label coral underwater images with types of benthic substrate together with their bounding box, and to segment and parse each coral image into different image regions associated with benthic substrate types. Task 3: ImageCLEFaware, the online disclosure of personal data often has effects which go beyond the initial context in which data were shared. Participants are required to provide automatic rankings of photographic user profiles in a series of real-life situations such as searching for a bank loan, an accommodation, a waiter job or a job in IT. The ranking is based on an automatic analysis of profile images and the aggregation of individual results. Task 4: ImageCLEFfusion, system fusion allows exploiting the complementary nature of individual systems to boost performance. Participants are tasked with creating novel ensemble methods that are able to significantly increase the performance of precomputed inducers in various use-case scenarios, such as visual interestingness and video memorability prediction.

JokeR: Automatic Wordplay and Humour Translation Workshop[13] [Ermakova et al., 2022a] aims to bring together translators and computer scientists to work on an evaluation framework for creative language, including data and metric development, and to foster work on automatic methods for wordplay translation. It offered the following tasks. Pilot task 1: Classify and interpret wordplay, classify single words containing wordplay according to a given typology, and provide lexical-semantic interpretations. Pilot task 2: Translate single term wordplay, translate single words containing wordplay. Pilot task 3: Translate phrase wordplay, translate entire phrases that subsume or contain wordplay. Task 4: Unshared Task, open to submissions that use the provided data in other ways.

LeQua: Learning to Quantify[14] [Esuli et al., 2022] aims to allow the comparative evaluation of methods for "learning to quantify" in textual datasets; i.e. methods for training predictors of the relative frequencies of the classes of interest in sets of unlabelled textual documents. These predictors (called "quantifiers") are required to issue predictions for several such sets, some of them characterized by class frequencies radically different from the ones of the training set. It offered the following tasks. Task 1: participants are provided with documents already converted into vector form; the task is thus suitable for participants who do not wish to engage in generating representations for the textual documents, but want instead to concentrate on optimizing the methods for learning to quantify. Task 2: participants are provided with the raw text of the documents; the task is thus suitable for participants who also wish to engage in generating suitable representations for the textual documents, or to train end-to-end systems.

LifeCLEF: Biodiversity identification and prediction[15] [Joly et al., 2022] aims to stimulate research in data science and machine learning for biodiversity monitoring. It offered the

[13] http://joker-project.com/
[14] https://lequa2022.github.io/
[15] https://www.imageclef.org/LifeCLEF2022

following tasks. Task 1: BirdCLEF, bird species recognition in audio soundscapes. Task 2: PlantCLEF, image-based plant identification at global scale (300K classes). Task 3: GeoLifeCLEF, location-based prediction of species based on environmental and occurrence data. Task 4: SnakeCLEF, snake species identification in medically important scenarios. Task 5: FungiCLEF, fungi Recognition from image and metadata.

PAN: Digital Text Forensics and Stylometry[16] [Bevendorff et al., 2022] focuses on digital text forensics and stylometry, studying how to quantify writing style and improve authorship technology. It offered the following tasks. Task 1: Authorship Verification, given two texts, determine if they are written by the same author. Task 2: IROSTEREO, profiling Irony and Stereotype Spreaders on Twitter, given a Twitter feed, determine whether its author spreads Irony and Stereotypes. Task 3: Style Change Detection, given a document, determine the number of authors and at which positions the author changes. Task 4: Trigger Warning Prediction, given a document, determine whether its content warrants a warning of potential negative emotional responses in readers.

SimpleText: Automatic Simplification of Scientific Texts[17] [Ermakova et al., 2022b] addresses the challenges of text simplification approaches in the context of promoting scientific information access, by providing appropriate data and benchmarks, and creating a community of NLP and IR researchers working together to resolve one of the greatest challenges of today. It offered the following tasks. Task 1: What is in (or out)? Select passages to include in a simplified summary, given a query. Task 2: What is unclear? Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications, ...). Task 3: Rewrite this! Given a query, simplify passages from scientific abstracts. Task 4: Unshared task, open to any submission that uses the provided data.

Touché: Argument Retrieval[18] [Bondarenko et al., 2022] focuses on decision-making processes, be it at the societal or at the personal level, often come to a point where one side challenges the other with a why-question, which is a prompt to justify some stance based on arguments. Since technologies for argument mining are maturing at a rapid pace, also ad-hoc argument retrieval becomes a feasible task in reach. It offered the following tasks. Task 1: Argument Retrieval for Controversial Questions, given a controversial topic and a collection of argumentative documents, the task is to retrieve and rank sentences (the main claim and its most important premise in the document) that convey key points pertinent to the controversial topic. Task 2: Argument Retrieval for Comparative Questions, given a comparative topic and a collection of documents, the task is to retrieve relevant argumentative passages for either compared object or for both and to detect their respective stances with respect to the object they talk about. 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.

[16] http://pan.webis.de/
[17] http://simpletext-project.com/
[18] https://touche.webis.de/

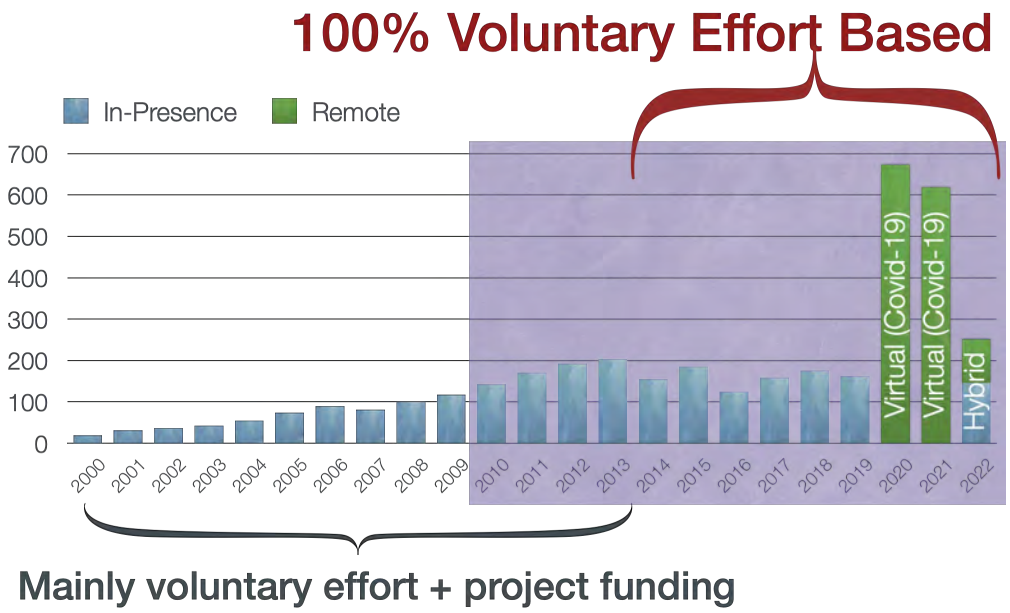

Figure 1: Attendance to CLEF over the years: x-axis reports CLEF editions; y-axis the number of attendees; the shading indicates the change from CLEF as a workshop, co-located with ECDL/TPDL, to CLEF as an independent conference and labs.

More information on the CLEF 2022 conference, the CLEF initiative and the CLEF Association is provided on the Web:

- CLEF 2022: https://clef2022.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 a substantial growth over the years, especially since when it is backed by the CLEF Association. We can also note how CLEF 2020 and CLEF 2021, which were online only and almost free registration due to COVID-19, represent a spike in the attendance. The in-presence attendance for CLEF 2022 has been substantially comparable to the pre-COVID editions, but the overall one has also increased 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.

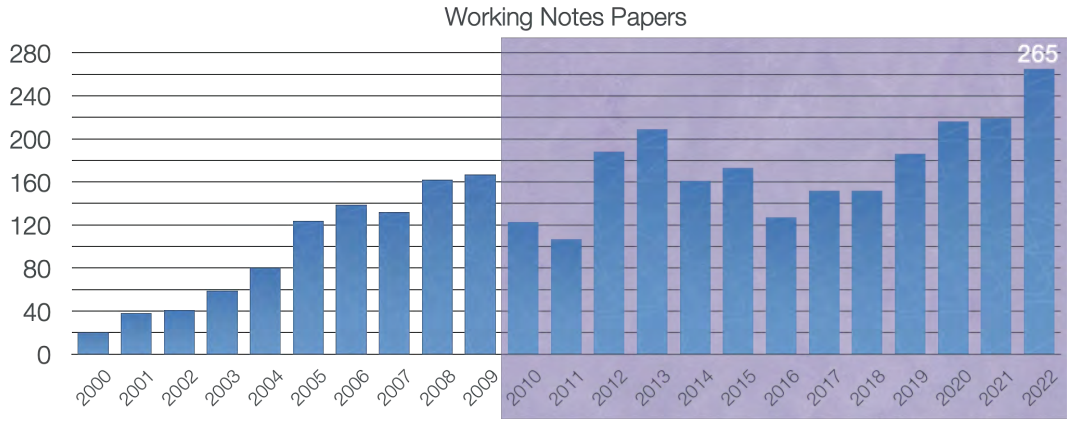

Figure 2: Papers published in the working notes over the years: x-axis reports CLEF editions; y-axis the number of papers in the working notes, highlighting 265, the papers in the CLEF 2022 working notes; the shading indicates the change from CLEF as a workshop co-located with ECDL/TPDL to CLEF as an independent conference and labs.

Finally, Figure 3 shows the Google Scholar metrics for CLEF[19] 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"[20], together with other important venues for the IR community, like SIGIR, CIKM, and WWW.

5 CLEF 2023 and Beyond

CLEF 2023 will be hosted by the Center for Research and Technology Hellas (CERTH), Thessaloniki, Greece, on 18–21 September 2023.

More information on CLEF 2023, the call for papers and the ongoing labs is available at:

- http://clef2023.clef-initiative.eu/

As far as labs are concerned, CLEF 2023 will run 13 evaluation activities out of 15 proposals received: 10 will be a continuation of the labs running during CLEF 2022

- BioASQ – Large-scale biomedical semantic indexing and question answering[21];
- CheckThat! – Check-Worthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Sources[22];
- eRisk – Early risk prediction on the Internet[23];
- iDPP – Intelligent Disease Progression Prediction[24];

[19] 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.

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

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

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

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

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

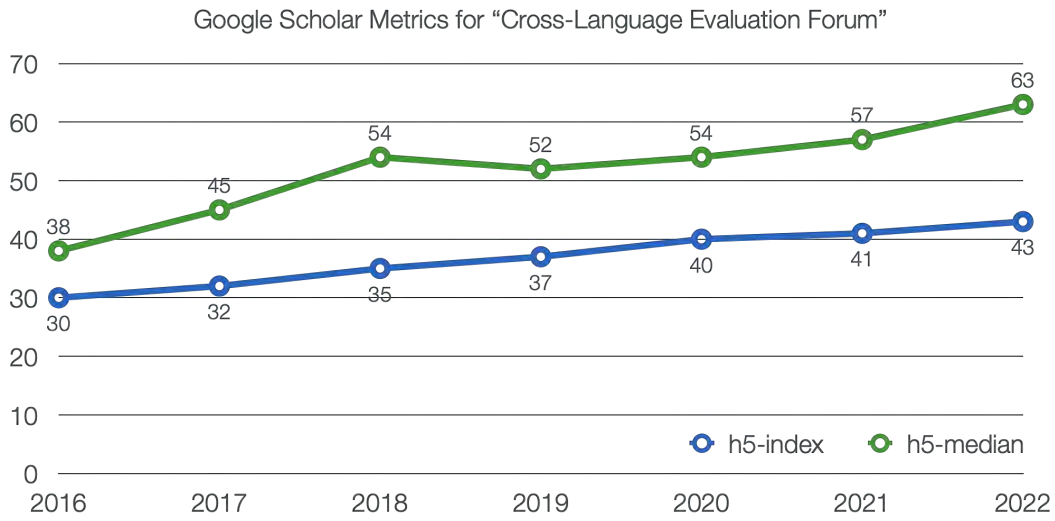

Figure 3: Google Scholar metrics for "Cross-Language Evaluation Forum" since 2016: the x-axis reports years, the y-axis the value for the h5-index (the largest number h such that at least h articles in that publication were cited at least h times each, only those of its articles that were published in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

- ImageCLEF – Multimedia Retrieval Challenge[25];
- JokeR – Automatic Wordplay Analysis[26];
- LifeCLEF – Multimedia Retrieval in Nature[27];
- PAN – Digital Text Forensics and Stylometry[28];
- SimpleText – Automatic Simplification of Scientific Texts[29];
- Touché – Argument and Causal Retrieval[30];

and 3 will be new pilot labs:

- DocILE – Document Information Localization and Extraction[31];
- EXIST – sEXism Identification in Social neTworks[32];
- LongEval – Longitudinal Evaluation of Model Performance[33].

Finally, bids for hosting CLEF 2024 are now open and will close around December 2022. Proposals can be sent to the CLEF Steering Committee Chair at chair@clef-initiative.eu.

[25] https://www.imageclef.org/2023
[26] https://www.joker-project.com/
[27] http://www.lifeclef.org
[28] https://pan.webis.de/
[29] https://simpletext-project.com/
[30] https://touche.webis.de/
[31] https://docile.rossum.ai/
[32] http://nlp.uned.es/exist2023/
[33] https://clef-longeval.github.io/

Acknowledgments

The success of CLEF 2022 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 Bologna, 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: University of Bologna (with special mention to the departments DIT, DIFA and DISI), the Department of Mathematics of the University of Padova, as well as the AI4media[34] and SoBigData[35] H2020 projects for their invaluable support. We thank the Friends of SIGIR program[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

A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C: Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, and N. Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)*, 2022. Lecture Notes in Computer Science (LNCS) 13390, Springer, Heidelberg, Germany.

J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, and E. Zangerle. Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection. In Barrón-Cedeño et al. [2022], pages 382–394.

A. Bondarenko, M. Fröbe, J. Kiesel, S. Syed, T. Gurcke, M. Beloucif, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, and M. Hagen. Overview of Touché 2022: Argument Retrieval. In Barrón-Cedeño et al. [2022], pages 311–336.

M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, and S. Clematide. Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In Barrón-Cedeño et al. [2022], pages 423–446.

L. Ermakova, T. Miller, F. Regattin, A.-G. Bosser, C. Borg, E. Mathurin, G. Le Corre, S. Araújo, R. Hannachi, J. Boccou, A. Digue, A. Damoy, and B. Jeanjean. Overview of JOKER@CLEF

[34] https://www.ai4media.eu/
[35] http://project.sobigdata.eu/
[36] http://sigir.org/general-information/funding-for-sigir-related-events/

2022: Automatic Wordplay and Humour Translation Workshop. In Barrón-Cedeño et al. [2022], pages 447–469.

L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, E. Mathurin, and P. Bellot. Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts. In Barrón-Cedeño et al. [2022], pages 470–494.

A. Esuli, A. Moreo, F. Sebastiani, and G. Sperduti. A Concise Overview of LeQua@CLEF 2022: Learning to Quantify. In Barrón-Cedeño et al. [2022], pages 362–381.

G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast, editors. *CLEF 2022 Working Notes*, 2022. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073. http://ceur-ws.org/Vol-3180/.

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.

A. Guazzo, I. Trescato, E. Longato, E. Hazizaj, D. Dosso, G. Faggioli, G. M. Di Nunzio, G. Silvello, M. Vettoretti, E. Tavazzi, C. Roversi, P. Fariselli, S. C. Madeira, M. de Carvalho, M. Gromicho, A. Chiò, U. Manera, A. Dagliati, G. Birolo, H. Aidos, B. Di Camillo, and N. Ferro. Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2022. In Barrón-Cedeño et al. [2022], pages 395–422.

B. Ionescu, H. Müller, R. Péteri, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. Dicente Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, and A. Clark. Overview of the ImageCLEF 2022: Multimedia Retrieval in Medical, Social Media and Nature Applications. In Barrón-Cedeño et al. [2022], pages 541–564.

A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet, M. Šulc, and M. Hrúz. Overview of LifeCLEF 2022: An Evaluation of Machine-Learning Based Species Identification and Species Distribution Prediction. In Barrón-Cedeño et al. [2022], pages 257–285.

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

Y. Li, B. Fang, J. He, H. Yoshikawa, S. A. Akhondi, C. Druckenbrodt, C. Thorne, Z. Afzal, Z. Zhai, T. Baldwin, and K. Verspoor. Overview of ChEMU 2022 Evaluation Campaign: Information Extraction in Chemical Patents. In Barrón-Cedeño et al. [2022], pages 521–540.

B. Mansouri, V. Novotný, A. Agarwal, D. W. Oard, and R. Zanibbi. Overview of ARQMath-3 (2022): Third CLEF Lab on Answer Retrieval for Questions on Math. In Barrón-Cedeño et al. [2022], pages 286–310.

P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. Kishore Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. Selim Kartal, M. Wiegand, M. Siegel, and J. Köhler. Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection. In Barrón-Cedeño et al. [2022], pages 495–520.

A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, A. Miranda-Escalada, L. Gasco, M. Krallinger, and G. Paliouras. Overview of BioASQ 2022: The Tenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In Barrón-Cedeño et al. [2022], pages 337–361.

J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani. Overview of eRisk 2022: Early Risk Prediction on the Internet. In Barrón-Cedeño et al. [2022], pages 233–256.

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.