From Multilingual to Multimodal: The Evolution of CLEF over Two Decades

Nicola Ferro and Carol Peters

Abstract This introductory chapter begins by explaining briefly what is intended by experimental evaluation in information retrieval in order to provide the necessary background for the rest of this volume. The major international evaluation initiatives that have adopted and implemented in various ways this common framework are then presented and their relationship to CLEF indicated. The second part of the chapter details how the experimental evaluation paradigm has been implemented in CLEF by providing a brief overview of the main activities and results obtained over the last two decades. The aim has been to build a strong multidisciplinary research community and to create a sustainable technical framework that would not simply support but would also empower both research and development and evaluation activities, while meeting and at times anticipating the demands of a rapidly evolving information society.

1 Introduction

CLEF - the Cross-Language Evaluation Forum for the first ten years, and the Conference and Labs of the Evaluation Forum since - is an international initiative whose main mission is to promote research, innovation, and development of information retrieval systems.

CLEF currently promotes research and development by providing an infrastructure for:

Nicola Ferro
Department of Information Engineering, University of Padua, Via G. Gradenigo, 6/B, 35131 Padova, Italy, e-mail: ferro@dei.unipd.it

Carol Peters
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), National Research Council (CNR), Via G. Moruzzi 1, 56124 Pisa, Italy, e-mail: carol.peters@isti.cnr.it

- multilingual and multimodal system testing, tuning and evaluation;
- investigation of the use of unstructured, semi-structured, highly-structured, and semantically enriched data in information access;
- creation of reusable test collections for benchmarking;
- exploration of new evaluation methodologies and innovative ways of using experimental data;
- discussion of results, comparison of approaches, exchange of ideas, and transfer of knowledge.

This activity is conducted by providing a platform for experimental system evaluation and then holding workshops and organizing an annual conference where researchers and developers can get together to discuss results and exchange ideas and experiences.

This aim of this chapter is to present the activity and results of CLEF over the last two decades. In Section 1, we begin by explaining briefly what is intended by experimental evaluation in information retrieval, providing pointers to more detailed discussions, in particular to the other two chapters in this first part of the book, in order to provide the necessary context. We then present the major international evaluation initiatives that have adopted this common framework and indicate their relationship to CLEF.

Sections 2 and 3 detail how the experimental evaluation paradigm has been implemented in CLEF by providing a brief overview of the main activities and results obtained in these first twenty years. The evolution and shift in focus can be seen as a reflection of the development of the information retrieval scene in this span of time. While the activities of CLEF in the first ten years (2000 - 2009) were very much focused on the evaluation of systems developed to run on multiple languages, since 2010 the scope has been widened to embrace many different types of multimodal retrieval. For convenience, in this chapter we refer to these two distinct, but not separate, phases of CLEF as CLEF 1.0 and CLEF 2.0. Figure 1 shows clearly the evolution of CLEF over the last two decades, and the shift from mainly text retrieval in the early years of CLEF 1.0 to all kinds of multimedia retrieval, with increasing attention being given to dynamic and user-oriented activities in CLEF 2.0. Many of the main CLEF activities are described in separate chapters in Parts III, IV and V of this volume; however, full details on all experiments, including methodologies adopted, test collections employed, evaluation measures used and results obtained, can be found in the CLEF Working Notes[1] and the CLEF Proceedings[2].

Section 4 provides valuable information on the test collections that have been created as a result of the evaluation activities in CLEF and on their availability. The final two Sections describe the CLEF Association, established in 2013 to support CLEF activities (Section 5), and the impact that we feel that CLEF has had on research into information access and evaluation both in Europe and globally (Section 6).

[1] Published annually in the CEUR Workshop Proceedings series (CEUR-WS.org).

[2] Published by Springer in their Lecture Notes for Computer Science series.

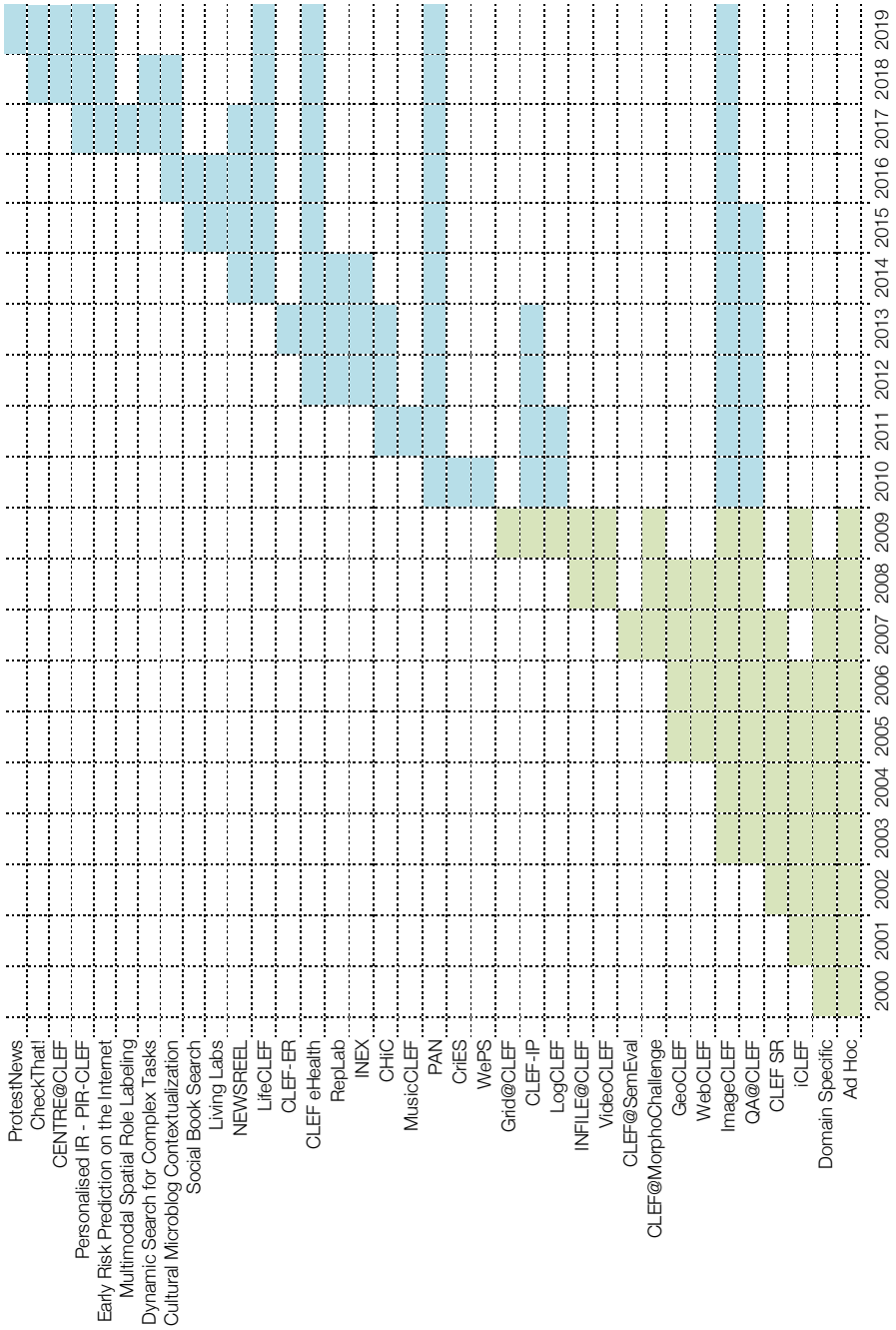

Fig. 1 Evolution of CLEF activities over time.

1.1 Experimental Evaluation

Information Retrieval (IR) is concerned with developing methods, algorithms, and systems which allow users to retrieve and access digitally stored information, in whatever language and media, relevant to their needs.

In IR users express their needs by means of queries – typically keyword-based queries expressed in natural language – that are often vague and imprecise formulations of their actual information needs, and systems retrieve items – generally termed documents – that match the user query and rank them by an estimation of their relevance to the query.

Since user queries and documents can be somewhat ambiguous, since a lot of contextual and task information is often left implicit, and since the notion of relevance itself is very complex and can change as the user progresses in the search (Saracevic, 1975; Mizzaro, 1997), IR systems adopt a *best match* approach, where results are ranked according to how well queries can be matched against documents, but always knowing that there will be some sort of inaccuracy and fuzziness.

IR system performance can be evaluated from two different standpoints, *efficiency* and *effectiveness*. Efficiency is concerned with the algorithmic costs of IR systems, i.e. how fast they are in processing the needed information and how demanding they are in terms of the computational resources required. Effectiveness, instead, is concerned with the ability of IR systems to retrieve and properly rank relevant documents while at the same time suppressing the retrieval of non relevant ones. The ultimate goal is to satisfy the user's information needs.

While efficiency could also be assessed formally, e.g. by proving the computational complexity of the adopted algorithms, effectiveness can be assessed only experimentally and this is why IR is a discipline strongly rooted in experimentation since its inception (Harman, 2011; Spärck Jones, 1981). Over the years, experimental evaluation has thus represented a main driver of progress and innovation in the IR field, providing the means to assess, understand, and improve the performance of IR systems from the viewpoint of effectiveness.

Experimental evaluation addresses a very wide spectrum of cases, ranging from system-oriented evaluation (Sanderson, 2010) to user-oriented evaluation (Kelly, 2009). In this volume, we will mainly focus on system-oriented evaluation which is performed according to the *Cranfield paradigm* (Cleverdon, 1967).

Figure 2 summarizes the Cranfield paradigm which is based on experimental collections $\mathscr{C} = (D, T, RJ)$ where: a corpus of documents D represents the domain of interest; a set of topics T represents the user information needs; and human-made relevance judgments RJ are the "correct" answers, or ground-truth, determining, for each topic, the relevant documents. Relevance judgments are typically expressed as either binary relevance, i.e. relevant or not relevant, or as graded relevance (Kekäläinen and Järvelin, 2002), e.g. not relevant, partially relevant, highly relevant. The ranked result lists, i.e. the IR system outputs, are then scored with respect to the ground-truth using several evaluation measures (Sakai, 2014a). The

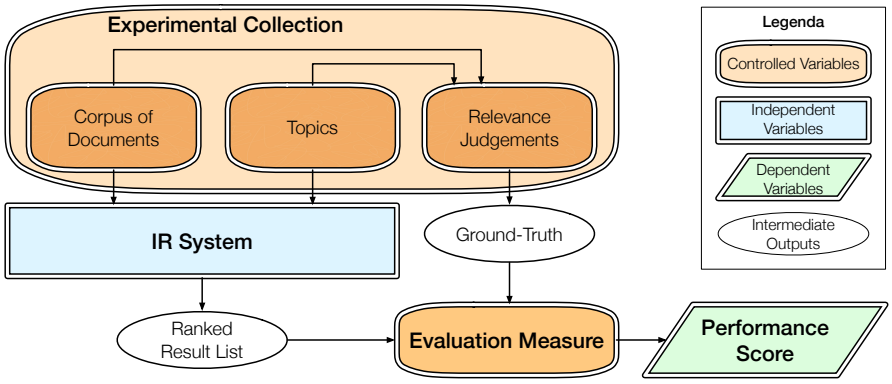

Fig. 2 The Cranfield paradigm for experimental evaluation.

evolution of Cranfield in IR system evaluation is discussed in detail in a following chapter by Ellen Voorhees.

The main goal of this experimental setup is to be able to compare the performance of different IR systems in a robust and repeatable way, as they are all scored with respect to the same experimental collection. Experimental collections and evaluation measures are *controlled variables*, since they are kept fixed during experimentation; IR systems are *independent variables*, since they are the object of experimentation, compared one against the other; and, performance scores are the *dependent variables*, since their observed value changes as IR systems change (Fuhr, 2012).

Carrying out experimental evaluation according to the Cranfield paradigm is very demanding in terms of both the time and the effort required to prepare the experimental collection. Therefore, it is usually carried out in publicly open and large-scale evaluation campaigns, often at international level, as exemplified in the next section, to share the effort, compare state-of-the-art systems and algorithms on a common and reproducible ground, and maximize the impact. Tetsuya Sakai, in another chapter in this first part of the book, provides a detailed description on how to setup a Cranfield style evaluation task, create experimental collections, and use evaluation measures.

In fact, IR evaluation adopts a whole breadth of evaluation measures (Sakai, 2014a) because different evaluation measures embed different user models in scanning the result list and thus represent different angles on the effectiveness of an IR system. *Average Precision (AP)* (Buckley and Voorhees, 2005), *Precision at Ten (P@10)* (Büttcher et al, 2007), *Rank-Biased Precision (RBP)* (Moffat and Zobel, 2008), *Normalized Discounted Cumulated Gain (nDCG)* (Järvelin and Kekäläinen, 2002), and *Expected Reciprocal Rank (ERR)* (Chapelle et al, 2009) are among the most commonly adopted measures. Evaluation measures are typically studied in an empirical way, e.g. by using correlation analysis (Voorhees and Harman, 1998), discriminative power (Sakai, 2006, 2012), or robustness to pool downsampling (Buckley and Voorhees, 2004; Yilmaz and Aslam, 2006). On the other hand, few studies have been undertaken to understand the formal properties of evaluation measures

and they have just scratched the surface of the problem: (Bollmann, 1984; Busin and Mizzaro, 2013; Amigó et al, 2013; van Rijsbergen, 1974; Ferrante et al, 2015, 2017, 2019). The following chapters by Voorhees and Sakai both discuss evaluation measures in more detail.

Finally, statistical analyses and statistical significance testing play a fundamental role in experimental evaluation (Carterette, 2012; Hull, 1993; Sakai, 2014b; Savoy, 1997) since they provide us with the means to properly assess differences among compared systems and to understand when they actually matter.

The activities of CLEF, as described in the rest of this book, have been conducted within this context of theory and practice, with the results helping to stimulate progress in the field of IR system experimentation and evaluation.

1.2 International Evaluation Initiatives

There are a number of evaluation initiatives around the world that follow the Cranfield paradigm, extending and adapting it to meet local requirements. In this section, we list the major ones, indicating their relationship with CLEF.

As is described in the chapter by Voorhees, IR experimental evaluation was initiated in 1992 by the US National Institute of Standards and Technology, in the *Text REtrieval Conference (TREC)* (Harman and Voorhees, 2005). The TREC conference series has constituted the blueprint for the organization of evaluation campaigns, providing guidelines and paving the way for others to follow[3]. In 1997, TREC included a track for *Cross Language Information Retrieval (CLIR)*. The aim was to provide researchers with an infrastructure for evaluation that would enable them to test their systems and compare the results achieved using different cross-language strategies (Harman et al, 2001).

However, after three years within TREC, it was decided that Europe with its diversity of languages was better suited for the coordination of an activity that focused on multilingual aspects of IR. Not only was it far easier in Europe to find the people and groups with the necessary linguistic competence to handle the language-dependent issues involved in creating test collections in different languages, but European researchers, both in academia and industry, were particularly motivated to study the problems involved in searching over languages other than English. Consequently, with the support of TREC, the *Cross-Language Evaluation Forum (CLEF)* was launched in 2000 by a consortium with members from several different European countries, and test collections were created in four languages (English, French, German and Italian).

The decision to launch CLEF in Europe came just one year after the first *NII-NACSIS Test Collection for IR Systems (NTCIR)* workshop was held in Asia[4]. NTCIR also saw the creation of test collections in languages other than English, i.e.

[3] See http://trec.nist.gov/

[4] See http://research.nii.ac.jp/ntcir/

in this case Asian languages, as strategic. NTCIR-1 thus included a task for cross-language Japanese to English IR and since then NTCIR has offered test collections and tasks for Chinese and Korean as well as Japanese and English. Organized on an eighteen monthly cycle, NTCIR has grown steadily over the years, covering many diverse information access tasks including, but not limited to, information retrieval, question answering, text summarisation and text mining, always with an emphasis on East Asian languages. In 2017, with its twelth conference, NTCIR celebrated its 20th birthday

In 2006 and 2007, in response to requests from colleagues in India, CLEF organized mono- and cross-language text retrieval tasks dedicated to Indian languages. Descriptions of this activity can be found in the CLEF Workshop Proceedings for those years (Nardi et al, 2006, 2007). This preliminary action helped to lead to the birth of a new evaluation initiative in India: the *Forum for Information Retrieval Evaluation (FIRE)*[5] in 2008. The objective of FIRE is to stimulate the development of IR systems capable of handling the specific needs of the languages of the Indian sub-continent. When FIRE began, Indian language information retrieval research was in a relatively primitive stage (especially with regard to large-scale quantitative evaluation). FIRE has had a significant impact on the growth of this discipline by providing test collections in many Indian languages (e.g. Bengali, Gujarati, Hindi, Marathi, Tamil, Telugu) and a forum where beginners can meet with and learn from experts in the field. Over the years, FIRE has evolved to include new domains like plagiarism detection, legal information access, mixed script information retrieval and spoken document retrieval.

Another important activity, which was first launched in CLEF before becoming an independent IR evaluation initiative in 2010, is MediaEval[6]. MediaEval attracts participants interested in multimodal approaches to multimedia involving, e.g., speech recognition, multimedia content analysis, music and audio analysis, viewer affective response, and social networks. In particular, it focuses on the human and social aspects of multimedia tasks. MediaEval began life as VideoCLEF, a track offered in CLEF in 2008 and 2009. Relations between the two activities have been maintained and, in 2017, the MediaEval workshop and the CLEF conference were co-located and run in close collaboration. More details on MediaEval can be found in the chapter by Gareth Jones in this volume.

On the other hand, the *INitiative for the Evaluation of XML Retrieval (INEX)*, run as a separate evaluation initiative from 2002 to 2011, decided in 2012 to run as a Lab under the CLEF umbrella. This Lab ran in CLEF until 2016. INEX promoted the evaluation of search engines for focused retrieval, i.e. the identification of the relevant parts of a relevant document. This can take many forms, e.g. passage retrieval from a long document, element retrieval from an XML document, page retrieval from books, as well as question answering. The chapter by Kamps et al. describes the important contribution made by INEX to experimental evaluation.

---

[5] See http://fire.irsi.res.in/

[6] See http://www.multimediaeval.org/

Each of the initiatives mentioned has been studied to meet the perceived needs of a specific community, reflecting linguistic, cultural and resource differences, while being designed within a common theoretical framework. This common background has facilitated discussion and exchange of ideas between the different groups and, at times, tasks run in collaboration. The aim is to avoid the duplication of effort and to provide complementary challenges, thus achieving a synergy of ideas and activities. An example of this is the CLEF/NTCIR/TREC task focused on Reproducibility, first experimented at CLEF in 2018[7] (Ferro et al, 2018). The objectives are to (i) reproduce the best - or most interesting - results achieved in previous editions of CLEF, NTCIR and TREC by using standard open source IR systems; and then (ii) to offer the additional components and resources developed in this activity to the IR community with the aim of improving existing open source systems.

2 CLEF 1.0: Cross-Language Evaluation Forum (2000–2009)

When CLEF began in 2000, cross language IR had only just started to be recognized as a separate sub-discipline[8], there were very few research prototypes in existence and work was almost entirely concentrated on text retrieval systems running on at most two languages. Thus, when CLEF was launched, the declared objectives were to develop and maintain an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, in both monolingual and cross-language contexts, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes (Peters, 2001). The aim was to promote the development of IR systems and tools in languages other than English and to stimulate the growth of the European research community in this area. However, while the first three editions of CLEF were dedicated to mono- and multilingual ad-hoc text retrieval, gradually the scope of activity was extended to include other kinds of text retrieval across languages (i.e., not just document retrieval but question answering and geographic IR as well) and on other media (i.e., collections containing images and speech). The goal was not only to meet but also to anticipate the emerging needs of the R & D community and to encourage the development of next generation multilingual IR systems.

In this section, dedicated to CLEF 1.0, we outline the main activities undertaken in the first ten years.

[7] See http://www.centre-eval.org/

[8] The first workshop on Cross-Lingual Information Retrieval was held at the Nineteenth ACM-SIGIR Conference on Research and Development in Information Retrieval in 1996. At this meeting there was considerable discussion aimed at establishing the scope of this area of research and defining the core terminology. The first ten years of CLEF did much to consolidate this field of study.

2.1 Tracks and Tasks in CLEF 1.0

Initially CLEF was very much influenced by its origins as a track within TREC. We not only adopted the same experimental paradigm that had been studied and implemented within TREC, but also inherited much of the vocabulary and the organizational framework. Therefore, for the first ten years the different activities were run under the heading of *Tracks*. Each track was run by a coordinating group with specific expertise in the area covered[9]. The coordinators were responsible for the definition and organization of the evaluation activity of their *Track* throughout the year. The results were presented and discussed at the annual CLEF Workshop held in conjunction with the European Conference for Digital Libraries. Most tracks offered several different tasks and these tasks normally varied each year according to the interests of the track coordinators and participants. This meant that the number of tracks offered by CLEF 1.0 increased over the years from just two in 2000 to ten separate tracks in 2009. Activities were mostly divided into two groups: tracks concerned with text retrieval and those which studied retrieval in other media: image, speech and video. The focus was always on collections in languages other than English. In this section we present the main tracks.

Of course, some of the CLEF 1.0 tracks continued as Labs in CLEF 2.0. This is the case, for example, of ImageCLEF and CLEF-QA, two of the most popular activities, in terms both of participation and diversity of tasks. For this reason, they are presented both as tracks in this section and as Labs in Section 3. On the other hand, the descriptions of LogCLEF and CLEF-IP, pilot experiments at the very end of CLEF 1.0 and Labs in the following years, appear in the CLEF 2.0 section.

2.1.1 Multilingual Text Retrieval (2000–2009)

Ad-Hoc document retrieval was the core track in CLEF 1.0. It was the one track that was offered every year and was considered of strategic importance. For this reason, we describe it is some detail here. The aim of the track was to promote the development of monolingual and cross-language text retrieval systems. From 2000-2007, the track exclusively used target collections of European newspaper and news agency documents and worked hard at offering increasingly complex and diverse tasks, adding new languages every year. Up until 2005, European languages were also used for the queries. In 2006 and 2007, in a collaboration with colleagues from the *Information Retrieval Society of India (IRSI)* which would lead in 2008 to the launching of FIRE, we added the possibility to query the English document collection with queries in a number of Indian languages. In 2008 and 2009, as a result of a joint activity with the Database Research Group of Tehran University, we included a test collection in Farsi, the Hamshahri corpus of 1996-2002 newspapers.

[9] It is impossible to acknowledge all the researchers and institutions that have been involved in the coordination of CLEF. Many, but certainly not all, are represented by the authors of the papers in this volume

Monolingual and cross-language (English to Persian) tasks were offered. As was to be expected, many of the eight participants focused their attention on problems of stemming. Only three submitted cross-language runs.

The addition of queries and a document collection in non European languages was important as it provided the opportunity to test retrieval systems on languages with very different scripts and syntactic structures. For example, the decision to offer a Persian target collection was motivated by several reasons: the challenging script (a modified version of Arabic with elision of short vowels) written from right to left; the complex morphology (extensive use of suffixes and compounding); the political and cultural importance.

In 2006 we added a task designed for more experienced participants, the so-called robust task, which used test collections from previous years in six languages (Dutch, English, French, German, Italian and Spanish) with the objective of rewarding experiments which achieve good stable performance over all queries rather than high average performance.

In 2008 we also introduced a task offering monolingual and cross-language search on library catalog records. It was organized in collaboration with The European Library (TEL)[10] and used three collections from the catalogs of the British Library, the Bibliothéque Nationale de France and the Austrian National Library. The underlying aim was to identify the most effective retrieval technologies for searching this type of very sparse multilingual data. In fact, the collections contained records in many languages in addition to English, French or German. The task presumed a user with a working knowledge of these three languages who wants to find documents that can be useful for them in one of the three target catalogs. Records in other languages were counted irrelevant. This was a challenging task but proved popular; participants tried various strategies to handle the multilinguality of the catalogs. The fact that the best results were not always obtained by experienced CLEF participants shows that the traditional approaches used for newspaper document retrieval are not necessarily the most effective for this type of data. The task was offered for two years.

Another task, offered for just two years, was designed to attract participation from groups interested in *Natural Language Processing (NLP)*. English test data from previous years was used but the organizers provided *Word Sense Disambiguation (WSD)* for documents and queries. Both monolingual and bilingual (Spanish to English) tasks were activated. This task ran for two years, however, the results were inconclusive. Overall, little or no improvement in performance was achieved by groups attempting to exploit the WSD information.

The focus of the Ad-Hoc track on multilingual IR implied considering and understanding the challenges posed to information access technology by variation between languages in their writing systems, and in their morphological, syntactic and lexical properties. This problematic is investigated in the chapter by Karlgren at al. in Part III of this volume.

[10] See http://www.theeuropeanlibrary.org/

Table 1 CLEF 2000–2009 Ad-Hoc Tasks. The following ISO 639-1 language codes have been used: am = Amharic; bg = Bulgarian; bn = Bengali; de = German; en = English; es = Spanish; fa = Farsi; fi = Finnish; fr = French; hi = Hindi; hu = Hungarian; id = Indonesian; it = Italian; mr = Marathi; nl = Dutch; or = Oromo; pt = Portuguese; ru = Russian; sv = Swedish; ta = Tamil; te = Telugu. TEL = data from The European Library

Edition	Monolingual	Bilingual	Multilingual
CLEF 2000	de; fr; it	x → en	x → de; en; fr; it
CLEF 2001	de; es; fr; it; nl	x → en x → nl	x → de; en; es; fr; it
CLEF 2002	de; es; fi; fr; it; nl; sv	x → de; es; fi; fr; it; nl; sv x → en (newcomers only)	x → de; en; es; fr; it
CLEF 2003	de; es; fi; fr; it; nl; ru; sv	it → es de → it fr → nl fi → de x → ru x → en (newcomers only)	x → de; en; es; fr x → de; en; es; fi; fr; it; nl; sv
CLEF 2004	fi; fr; ru; pt	es; fr; it ;ru → fi de; fi; nl; sv → fr x →ru x → en (newcomers only)	x → fi; fr; ru; pt
CLEF 2005	bg; fr; hu; pt	x→ bg; fr; hu; pt	Multi8 2yrson (as in CLEF 2003) Multi8 Merge (as in CLEF 2003)
CLEF 2006	bg; fr; hu; pt Robust de; en; es; fr; it; nl	x → bg; fr; hu; pt am; hi; id; te; or → en Robust it → es fr → nl en → de	 Robust x → de; en; es; fr; it; nl
CLEF 2007	bg; cz; hu Robust en; fr; pt	x → bg; cz; hu am; id; or; zh → en bn; hi; mr; ta; te → en Robust x → en; fr; pt	
CLEF 2008	fa TEL de; en; fr Robust WSD en	en → fa TEL x → de; en; fr Robust WSD es → en	
CLEF 2009	fa TEL de; en; fr Robust WSD en	en → fa TEL x → de; en; fr Robust WSD es → en	

Table 1 gives a detailed breakdown of the collections and tasks offered for Ad-Hoc in each of these ten years. It can be seen that bilingual tasks were often proposed for unusual pairs of languages, such as Finnish to German, or French to Dutch. In addition multilingual tasks were offered in which queries in one language were posed to target collections in a varying number of languages. x as the query language in the bilingual and multilingual tasks denotes any of the languages offered for the monolingual task of that year.

The results of this track were considerable. It is probably true to say that it has done much to foster the creation of a strong European research community in the CLIR area. It provided the resources, the test collections and also the forum for discussion and comparison of ideas and results. Groups submitting experiments over several years showed flexibility in advancing to more complex tasks, from monolingual to bilingual and multilingual experiments. Much work was done on fine tuning for individual languages while other efforts concentrated on developing language independent strategies (McNamee and Mayfield, 2004). Over the years, there was substantial proof of significant increase in retrieval effectiveness in multilingual settings by systems of CLEF participants (Braschler, 2004).

The paper by Savoy and Braschler in this volume discusses some of the lessons learnt from this track.

2.1.2 The Domain-Specific Track (2001–2008)

Another text retrieval track offered for many years in CLEF 1.0 was the Domain-Specific track which was organised by a group with specific expertise in the area covered.[11] Mono- and cross-language retrieval was investigated using structured data (e.g. bibliographic data, keywords and abstracts) from scientific reference databases. The track used German, English and Russian target collections in the social science domain. A multilingual controlled vocabulary was also provided. A main finding was that metadata-based search can achieve similar results as those obtained using full-text. The results of the mono- and cross-language experiments were very similar in terms of performance to those achieved in the ad-hoc track.

In CLEF 2.0, domain-specific activities acquired a multimedia/multimodal perspective and included tasks involving patent retrieval, heath management and biodiversity.

2.1.3 Interactive Cross-Language Retrieval (2002–2009)

In the iCLEF track, cross-language search capabilities were studied from a user-inclusive perspective. A central research question was how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language.

[11] This track was coordinated by Michael Kluck, Informationszentrum Sozialwissenschaften (IZ), Germany.

In 2006, iCLEF moved from the news collections used in the ad-hoc tasks in order to explore user behaviour in a collection where the cross-language search necessity arises more naturally for average users. The choice fell on Flickr, a large-scale, online image database based on an extensive social network of WWW users, with the potential for offering both challenging and realistic multilingual search tasks for interactive experiments. The search interface provided by the iCLEF organizers was a basic cross-language retrieval system for the Flickr image database[12] presented as an online game: the user was given an image, and had to find it again without any a priori knowledge of the language(s) in which the image is annotated. The game was publicized on the CLEF mailing list and prizes were offered for the best results in order to encourage participation. The main novelty of the iCLEF 2008 experiments was the shared analysis of a search log from a single search interface provided by the organizers (i.e. the focus was on log analysis, rather than on system design).

The 2008 experiments resulted in a truly reusable data set (the first time in iCLEF!), with 5,000 complete search sessions recorded and 5,000 post-search and post-experience questionnaires. 200 users from 40 countries played an active role in these experiments which covered six target languages. A main observation was that, in addition to better CLIR algorithms, more research was needed on interactive features to help users bridge the language gap.

The track was organised in a similar way in 2009. The organizers provided a default multilingual search system which accessed images from Flickr, with the whole iCLEF experiment run as an online game. Interaction by users with the system was recorded in log files which were shared with participants for further analyses, and provide a future resource for studying various effects on user-orientated cross-language search.

2.1.4 The Question-Answering Track (2003–2015)

From 2003 on, CLEF also offered mono- and cross-language question answering tasks. The QA track was instrumental in encouraging researchers working in the natural language processing field to participate in CLEF. The main scenario in the early years was event targeted QA on a heterogeneous document collection. Besides the usual news collections used in the ad-hoc track, articles from Wikipedia were also considered as sources of answers and parallel aligned European legislative documents were included from 2009.

This track was inspired by the work in TREC on question answering but in CLEF the focus was on multilinguality. Many monolingual and cross-language sub-tasks were offered: Basque, Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish were proposed as both query and target languages; not all were used in the end. This track proved very popular in CLEF 1.0 and was, in fact, continued in CLEF 2.0. Over the years, a lot of resources and know-how were accumulated. One important lesson learnt was that offering so many language

[12] See http://www.flickr.com/

possibilities meant that there were always only a few systems participating in the same task, with the same languages. This meant that comparative analysis was often problematic. The chapter by Peñas et al. in this volume discusses in detail the design, experience and results of question answering activities in CLEF.

2.1.5 Cross-Language Retrieval in Image Collections (2003–2019)

Although at the beginning CLEF was very much focused on text retrieval, in 2003 it was decided to offer a track testing the retrieval of images from multilingual collections. ImageCLEF was thus launched with the goal of providing support for the evaluation of 1) multilingual image retrieval methods, to compare the effect of retrieval of image annotations and query formulations in several languages, 2) multimodal information retrieval methods based on the combination of visual and textual features, and 3) language-independent methods for the automatic annotation of images with concepts. The initial activity in this track is described in the chapter by Clough and Tsikrika in this volume. However, over the years, the track became increasingly complex. With the introduction of search on medical images in CLEF 2004, it also became very oriented towards the needs of an important user community (see the chapter by Müller et al.).

ImageCLEF rapidly became the most popular track in CLEF 1.0, even though (or maybe because) it was the track that deals the least with language and linguistic issues. This interest was to continue and diversify in CLEF 2.0. This is also exemplified in the chapters by Wang et al. and Piras et al.

2.1.6 Spoken Document/Speech Retrieval (2003–2007)

Following a preliminary investigation carried out as part of the CLEF 2002 campaign, a Cross-Language Spoken Document Retrieval (CLSDR) track was organized in CLEF 2003 and 2004. The track took as its starting point automatic transcripts prepared by NIST for the TREC 8-9 SDR tracks and generated using different speech recognition systems. The task consisted of retrieving news stories within a repository of about 550 hours of transcripts of American English news. The original English short search topics were formulated in French and German, to provide a CL-SDR task.

The CLEF 2005 Cross-Language Speech Retrieval (CL-SR) track followed these two years of experimentation but used audio data from the MALACH (Multilingual Access to Large Spoken Archives) collection which is based on interviews with Holocaust survivors from the archives of the Shoah Visual History Foundation. Spontaneous, conversational speech lacks clear topic boundaries and is considerably more challenging for the Automatic Speech Recognition, (or ASR), techniques on which fully-automatic content-based search systems are based. Although, advances in ASR had made it possible to contemplate the design of systems that would provide a useful degree of support for searching large collections of spontaneous con-

versational speech, no representative test collection that could be used to support the development of such systems was widely available for research use at that time. The principal goal of the CLEF CL-SR track was thus to create such a test collection. The data used was mainly in English and Czech. Topics were developed in several languages. Additional goals included benchmarking the current state of the art for ranked retrieval of spontaneous conversational speech and fostering interaction among a community of researchers with interest in that challenge.

Those goals were achieved. Over 3 years, research teams from 14 universities in 6 countries submitted runs for official scoring. The resulting English and Czech collections are the first information retrieval test collections of substantial size for spontaneous conversational speech. Unique characteristics of the English collection fostered research comparing searches based on automatic speech recognition and manually assigned metadata, and unique characteristics of the Czech collection inspired research on evaluation of information retrieval from unsegmented speech.

The CLEF spoken document and speech retrieval activities are described in more detail in the chapter by Gareth Jones.

2.1.7 Multilingual Web Retrieval (2005–2008)

The WebCLEF track focused on evaluation of systems providing multi- and cross-lingual access to web data. In the final year, a multilingual information synthesis task was offered, where, for a given topic, participating systems were asked to extract important snippets from web pages (fetched from the live web and provided by the task organizers). The systems had to focus on extracting, summarizing, filtering and presenting information relevant to the topic, rather than on large scale web search and retrieval per se. The aim was to refine the assessment procedure and evaluation measures. WebCLEF 2008 had lots of similarities with (topic-oriented) multidocument summarization and with answering complex questions. An important difference was that at WebCLEF, topics could come with extensive descriptions and with many thousands of documents from which important facts had to be mined. In addition, WebCLEF worked with web documents, which can be very noisy and redundant.

Although the Internet would seem to be the obvious application scenario for a CLIR system, WebCLEF had a rather disappointing participation. For this reason, the track was dropped.

2.1.8 Geographical Retrieval (2005–2008)

The purpose of GeoCLEF was to test and evaluate cross-language geographic information retrieval for topics with a geographic specification. How best to transform into a machine readable format the imprecise description of a geographic area found in many user queries was considered an open research problem. This track was run for four years in CLEF, examining geographic search of a text corpus. Some topics

simulated the situation of a user who poses a query when looking at a map on the screen. In GeoCLEF 2006 and 2007, it was found that keyword based systems often do well on the task and the best systems worked without any specific geographic resource. In 2008 the best monolingual systems used specific geo reasoning; there was much named-entity recognition (often using Wikipedia) and NER topic parsing. Geographic ontologies were also used (such as GeoNames and World Gazeteer), in particular for query expansion.

The track was coordinated by Frederic Gey and Ray Larson of UC Berkeley, School of Information. In 2009, they decided to move this activity from Europe to Asia and initiated a geotemporal retrieval task at NTCIR-8. However, in CLEF 2009, a new track, LogCLEF, continued to study information retrieval problems from the geographical perspective (see Section 3.1.9).

2.1.9 Multilingual Information Filtering (2008–2009)

The purpose of the INFILE (INformation FILtering & Evaluation) track, sponsored by the French National Research Agency, was to evaluate cross-language adaptive filtering systems. The goal of these systems is to successfully separate relevant and non-relevant documents in an incoming stream of textual information with respect to a given profile. The document and profile may be written in different languages.

INFILE extended the last filtering track of TREC 2002 in the following ways:

- Monolingual and cross-language tasks were offered using a corpus of 100,000 Agence France Press (AFP) comparable newswire stories for Arabic, English and French;
- Evaluation was performed by an automatic querying of test systems with a simulated user feedback. A curve of the evolution of efficiency was computed along with more classical measures already tested in TREC.

Unfortunately, the innovative crosslingual aspect of the task was not really explored, since most of the runs were monolingual English and no participant used the Arabic topics or documents.

2.1.10 Cross-Language Video Retrieval (2008–2009)

The aim of the VideoCLEF track was to develop and evaluate tasks related to analysis of and access to multilingual multimedia content. Participants used a video corpus containing episodes of a dual language television program in Dutch and English, accompanied by speech recognition transcripts. The dual language programming of Dutch TV offered a unique scientific opportunity, presenting the challenge of how to exploit speech features from both languages.

In 2010, the VideoCLEF organisers decided to set up an independent benchmarking initiative, known as MediaEval[13]. MediaEval attracts participants who are

[13] See http://www.multimediaeval.org/

interested in multimodal approaches to multimedia involving, e.g., speech recognition, multimedia content analysis, music and audio analysis, user-contributed information (tags, tweets), viewer affective response, social networks, temporal and geo-coordinates. This initiative is having a lot of success with a very active participation. Results are presented in an annual workshop.

More information on VideoCLEF and MediaEval is given in the chapter by Gareth Jones.

2.1.11 Component-based Evaluation (2009)

Grid@CLEF was a pilot experiment focused on *component-based evaluation* and aimed at establishing a long term activity comprising a series of systematic experiments in order to improve the comprehension of *MultiLingual Information Access (MLIA)* systems and gain an exhaustive picture of their behaviour with respect to languages. To this end, Grid@CLEF introduced the notion of *Grid of Points (GoP)* (Ferro and Harman, 2010), i.e. a set of IR systems originated by all the possible combinations of components under experimentation.

Grid@CLEF 2009 offered traditional monolingual ad-hoc tasks in 5 different languages (Dutch, English, French, German, and Italian) and used consolidated and very well known collections from CLEF 2001 and 2002 with a set of 84 topics. Participants had to conduct experiments according to the *Coordinated Information Retrieval Components Orchestration (CIRCO)* framework, an XML-based protocol which allows for a distributed, loosely coupled, and asynchronous experimental evaluation of IR systems. A Java library was provided which could be exploited to implement CIRCO together with an example implementation with the Lucene IR system. The task proved to be particularly challenging. Of the 9 original participants, only 2 were able to submit runs. They used different IR systems or combination of them, namely Lucene, Terrier, and Cheshire II. Partly because it was seen as overly complex, the activity was suspended.

Even if only run for one year, Grid@CLEF seeded some follow-up research lines. The interest in component-based evaluation was continued by Hanbury and Müller (2010) and embedded in the idea of evaluation-as-a-service (Hopfgartner et al, 2018), as discussed in a chapter in this book by Hanbury and Müller. The idea of GoP was taken up by Ferro and Silvello (2016, 2017) to develop *ANalysis Of VAriance (ANOVA)* models able to break-down overall system performance into those of the constituting components. GoP have also been exploited by Angelini et al (2018) to develop a *Visual Analytics (VA)* system to explore and intuitively make sense of them, as is described in a chapter in this book by Ferro and Santucci.

3 CLEF 2.0: Conference and Labs of the Evaluation Forum (2010–2019)

The second period of CLEF started with a clear and compelling question: after a successful decade studying multilinguality for European languages, what were the main unresolved issues currently facing us? To answer this question, we turned to the CLEF community to identify the most pressing challenges and to list the steps to be taken to meet them.

The discussion led to the definition and establishment of the *CLEF Initiative*, whose main mission is to promote research, innovation, and the development of information access systems with an emphasis on multilingual and *multimodal* information with various levels of structure.

In the CLEF Initiative an increased focus is on the *multimodal* aspect, intended not only as the ability to deal with information coming in multiple media but also in different modalities, e.g. the Web, social media, news streams, specific domains and so on. These different modalities should, ideally, be addressed in an integrated way; rather than building vertical search systems for each domain/modality the interaction between the different modalities, languages, and user tasks needs to be exploited to provide comprehensive and aggregated search systems. Thus, multimodality became a major theme of CLEF 2.0.

The new challenges for CLEF also called for a renewal of its structure and organization. The annual CLEF meeting is no longer a Workshop, held in conjunction with the European Digital Library Conference (ECDL, now TPDL), but has become an independent event, held over 3.5-4 days and made up of two interrelated activities: the *Conference* and the workshops of the *Labs*.

The *Conference* is a peer-reviewed conference, open to the IR community as a whole and not just to *Lab* participants, and aims at stimulating discussion on innovative evaluation methodologies and fostering a deeper analysis and understanding of experimental results. The *Labs* replace the *Tracks* of CLEF 1.0 and are organised on a yearly basis, culminating with the annual meeting where the results are discussed. Lab coordinators are responsible for the organization of the IR system evaluation activities of their Lab throughout the year and for their annual Lab workshop. They also give plenary Lab "overview presentations" during the conference to allow non-participants to get a sense of the direction of the research frontiers.The *Conference* and the *Labs* are expected to interact, bringing new interests and new expertise into CLEF.

Moreover, in order to favour participation and the introduction of new perspectives, CLEF now has an open-bid process which allows research groups and institutions to bid to host the annual CLEF event and to propose new themes, characterizing each edition.

The new challenges and new organizational structure motivated the change in name for CLEF: from the *Cross-Language Evaluation Forum* to *Conference and Labs of the Evaluation Forum*, in order to reflect the widened scope.

3.1 Workshops and Labs in CLEF 2.0

The move from the Tracks of CLEF 1.0 to the Labs of CLEF 2.0 was first made in CLEF 2010. A procedure was set up for the selection of the Labs to be held each year. A Lab Selection Committee launches a Call for Proposals in the Fall of the previous year. Proposals are accepted for two different types of Labs:

- Benchmarking Labs, providing a "campaign-style" evaluation for specific information access problems, similar in nature to the traditional CLEF campaign "Tracks" of CLEF 1.0. Topics covered by campaign-style labs can be inspired by any information access-related domain or task.
- Workshop-style Labs, following a more classical "workshop" pattern, exploring issues of evaluation methodology, metrics, processes etc. in information access and closely related fields, such as natural language processing, machine translation, and human-computer interaction.

For first time proposers, it is highly recommended that a lab workshop be first organised to discuss the format, the problem space, and the practicalities of the shared task. At the annual meeting, Labs are organised so that they contain ample time for general discussion and engagement by all participants - not just those presenting campaign results and papers. The criteria adopted for selection of Lab proposals include: importance of problem, innovation, soundness of methodology, clear movement along a growth path, likelihood that the outcome would constitute a significant contribution to the field. Additional factors are minimal overlap with other evaluation initiatives and events, vision for a potential continuation, and possible interdisciplinary character.

In this section, we provide a brief description of the Workshops and the Labs held in the second decade of CLEF, and shown in Figure 1 at page 3. For completeness, we have also included indication of the activities underway in 2019. We begin by describing the one-year experimental Workshops and continue with presentations of the fully-fledged Labs.

3.1.1 Web People Search (2010)

The WePS workshop focused on person name ambiguity and person attribute extraction from Web pages and on online reputation management for organizations. The first edition of this workshop, WePS-1, was run as a Semeval 1 task in 2007, whereas WePS-2 was a workshop at the WWW 2009 Conference. WePS-1 addressed only the name co-reference problem, defining the task as clustering of web search results for a given person name. In WePS-2 the evaluation metrics were refined and an attribute extraction task for web documents returned by the search engine for a given person name was added.

In the edition of WePS at CLEF both problems were merged into a single task, where the system must return both the documents and the attributes for each of a number of people sharing a given name. This was not a trivial step from the point of

view of evaluation: a system may correctly extract attribute profiles from different URLs but then incorrectly merge these profiles. While WePS-1 and WePS-2 had focused on consolidating a research community around the problem and developing an appropriate evaluation methodology, in WePS-3 the focus was on involving industrial stakeholders in the evaluation campaign, as providers of input to the task design phase and also as providers of realistic scale datasets. Intelius, Inc. - one of the main Web People Search services, providing advanced people attribute extraction and profile matching from web pages – collaborated in the activity. The discussions at this workshop resulted in the setting up of RepLab, described in Section 3.1.14.

3.1.2 Cross-lingual Expert Search (2010)

CriES was run as a brainstorming workshop and addressed the problem of multilingual expert search in social media environments. The main topics were multilingual expert retrieval methods, social media analysis with respect to expert search, selection of datasets and evaluation of expert search results. Online communities generate major economic value and form pivotal parts of corporate expertise management, marketing, product support, product innovation and advertising. In many cases, large-scale online communities are multilingual by nature (e.g. developer networks, corporate knowledge bases, blogospheres, Web 2.0 portals). Nowadays, novel solutions are required to deal with both the complexity of large-scale social networks and the complexity of multilingual user behavior. It thus becomes more important to efficiently identify and connect the right experts for a given task across locations, organizational units and languages. The key objective of the workshop was to consider the problem of multilingual retrieval in the novel setting of modern social media leveraging the expertise of individual users.

3.1.3 Music Information Retrieval (2011)

MusiCLEF was run as a brainstorming workshop promoting the development of new methodologies for music access and retrieval on real public music collections. A major focus was on multimodal retrieval achieved by combining content-based information, automatically extracted from music files, with contextual information, provided by users via tags, comments, or reviews. MusiCLEF aimed at maintaining a tight connection with real world application scenarios, focusing on issues related to music access and retrieval that are faced by professional users. Two benchmarking tasks were studied: the automatic categorization of music to be used as soundtrack for TV shows; the automatic identification of the pieces in a music digital library. In 2012, this activity continued as part of the MediaEval Initiative[14], described in Section 2.1.10.

[14] See http://www.multimediaeval.org/

3.1.4 Entity Recognition (2013)

The identification and normalisation of biomedical entities in scientific literature has a long tradition and a number of challenges have contributed to the development of reliable solutions. Increasingly, patient records are processed to align their content with other biomedical data resources, but this approach requires analysing documents in different languages across Europe.

CLEF-ER was a brainstorming workshop on the multilingual annotation of named entities and terminology resource acquisition with a focus on entity recognition in biomedical text in different languages and on a large scale. Several corpora in different languages, i.e. Medline titles, European Medicines Agency documents and patent claims, were provided to enable ER in parallel documents. Participants were asked to annotate entity mentions with concept unique identifiers (CUIs) in the documents of their preferred non-English language. The evaluation determined the number of correctly identified mentions against a silver standard and performance measures for the identification of CUIs in the non-English corpora. Participants could make use of the prepared terminological resources for entity normalisation and the English silver standard corpora (SSCs) as input for concept candidates in the non-English documents. Participants used different approaches including translation techniques and word or phrase alignments as well as lexical look-up and other text mining techniques.

3.1.5 Multimodal Spatial Role Labeling (2017)

The extraction of spatial semantics is important in many real-world applications such as geographical information systems, robotics and navigation, semantic search, etc. This workshop studied how spatial information could be best extracted from free text while exploiting accompanying images. The task investigated was a multimodal extension of a spatial role labeling task previously introduced in the SemEval series. The multimodal aspect of the task made it appropriate for CLEF 2.0.

3.1.6 Extracting Protests from News (2019)

ProtestNews aimed at testing and improving state-of-the-art generalizable machine learning and natural language processing methods for text classification and information extraction on English news from multiple countries such as India and China in order to create comparative databases of contentious political events (riots, social movements), i.e. the repertoire of contention that can enable large scale comparative social and political science studies. Three tasks were investigated: *Task 1 - News article classification as protest vs. non-protest*: given a random news article, to what extent can we predict whether it is reporting a contentious politics event that has happened or is happening? *Task 2 - Event sentence detection*: given a news article that is classified as positive in Task 1, to what extent can we identify the sentence(s)

that contain the event information? *Task 3 - Event extraction*: given the event sentence that is identified in Task 2, to what extent can we extract key event information such as place, time, participants, etc.?

3.1.7 Question Answering (2003–2015)

As described in the previous section, question answering was an important activity in CLEF from 2003. The QA@CLEF track, which became a Lab in 2010, examined several aspects of question answering in a multilingual setting on document collections ranging from news, legal documents, medical documents, and linked data. From 2010 on, it was decided that if progress was to be made a substantial change was needed in the design of the QA system architecture, with particular regard to answer validation and selection technologies. For this reason, the new formulation of the task after 2010 left the retrieval step aside to focus on the development of technologies able to work with a single document, answering questions about it and using the reference collections as sources of background knowledge that help the answering process. See the chapter by Peñas et al. in this volume for a more exhaustive description.

3.1.8 Image Retrieval (2003–2019)

As has already been stated, since its beginnings, ImageCLEF has been one of the most popular activities at CLEF. It has had the important merit of helping to make CLEF truly multidisciplinary by bringing the image processing community into close contact with researchers working on all kinds of text retrieval and in natural language processing. The main goal of the ImageCLEF Labs in CLEF 2.0 is to support multilingual users from a global community accessing an ever growing body of visual information. The objective is to promote the advancement of the fields of visual media analysis, indexing, classification, and retrieval, by developing the necessary infrastructure for the evaluation of visual information retrieval systems operating in monolingual, crosslanguage and language-independent contexts. ImageCLEF aims at providing reusable resources for such benchmarking purposes.

The chapters by Wang et al., Piras et al., and Müller et al. in this volume give an account of the wide range of ImageCLEF activities in CLEF 2.0.

3.1.9 Log File Analysis (2009–2011)

Search logs are a means to study user information needs and preferences. Interactions between users and information access systems can be analyzed and studied to gather user preferences and to learn what the user likes the most, and to use this information to personalize the presentation of results. The literature of log analysis of information systems shows a wide variety of approaches to learning user prefer-

ences by looking at implicit or explicit interaction. However, there has always been a lack of availability and use of log data for research experiments which makes the verifiability and repeatability of experiments very limited. LogCLEF investigated the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems by offering openly-accessible query logs from search engines and digital libraries. An important long-term aim of the LogCLEF activity was to stimulate research on user behavior in multilingual environments and promote standard evaluation collections of log data.

Between 2009 and 2011, LogCLEF released collections of log data with the aim of verifiability and repeatability of experiments. During the three editions of LogCLEF, different collections of log datasets were distributed to the participants together with manually annotated query records to be used as a training or test set. In the final edition, a Web based interface to annotate log data was designed and created on the basis of the experience of past participants for different tasks: language identification, query classification, and query drift. The public distribution of the datasets and results and the exchange of system components aimed at advancing the state of the art in this research area (Di Nunzio et al, 2011).

3.1.10 Intellectual Property in the Patent Domain (2009–2013)

The patent system is designed to encourage disclosure of new technologies and novel ideas by granting exclusive rights on the use of inventions to their inventors, for a limited period of time. An important requirement for a patent to be granted is that the invention it describes is novel. That is, there is no earlier patent, publication or public communication of a similar idea. To ensure the novelty of an invention, patent offices as well as other Intellectual Property (IP) service providers perform thorough searches called prior art searches or validity searches. Since the number of patents in a company's patent portfolio affects the company market value, well-performed prior art searches that lead to solid, difficult to challenge patents are of high importance.

The CLEF-IP Lab, which began as an experimental track at the end of CLEF 1.0, focused on various aspects of patent search and intellectual property search in a multilingual context using the MAREC collection of patents, gathered from the European Patent Office. In its first year, CLEF-IP organized one task only, a text oriented retrieval that modeled the Search for Prior Art done by experts at patent offices. In terms of retrieval effectiveness the results of this initial study were hard to evaluate: it appeared that the effective combination of a wide range of indexing methods produced the best results. It was agreed that further studies were needed to understand what methodology maps best to what makes a good (or better) system from the point of view of patent searchers. In the following years, the types of CLEF-IP tasks broadened to include patent text classification, patent image retrieval and classification, and (formal) structure recognition. With each task, the test collection was extended to accommodate the additional tasks.

The activity of this Lab and the results achieved are described in the chapter by Piroi and Hanbury in this volume.

3.1.11 Digital Text Forensics (2010–2019)

Since its first introduction in 2010, the PAN Lab has been extremely popular with a large participation. Over the years, the Lab has offered a range of tasks focusing on the general area of "Uncovering Plagiarism, Authorship and Social Software Misuse" in a multilingual context. In 2016, the Lab changed its name to the more general "Digital Text Forensics". PAN is also a good example of the cooperation between the different international evaluation initiatives listed in Section 1.2. The Lab coordinators have collaborated for a number of years in the organization of evaluation tasks at *Forum for Information Retrieval Evaluation (FIRE)*, organized by the Information Retrieval Society of India, in Indian languages, Arabic and Persian.

Details on the diverse activities of this Lab are presented in the chapter by Rosso et al. in this volume.

3.1.12 Cultural Heritage in CLEF (2011–2013)

Cultural heritage collections preserved by archives, libraries, museums and other institutions are often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity. Cultural heritage institutions have different approaches to managing information and serve diverse user communities, often with specialized needs. The targeted audience of the CHiC lab and its tasks were developers of cultural heritage information systems, information retrieval researchers specializing in domain-specific (cultural heritage) and / or structured information retrieval on sparse text (metadata), and semantic web researchers specializing in semantic enrichment with LOD data. Evaluation approaches (particularly system-oriented evaluation) in this domain have been fragmentary and often non-standardized.

CHiC began with a brainstorming workshop in 2011 aimed at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems. In a pilot lab in 2012, a standard ad-hoc information retrieval scenario was tested together with two use-case-based scenarios (diversity task and semantic enrichment task). The 2013 lab diversified and became more realistic in its task organization. The pilot lab in 2012 demonstrated that in cultural heritage information systems ad-hoc searching might not be the prevalent form of access to this type of content. The 2013 CHiC lab focused on multilinguality in the retrieval tasks (up to 13 languages) and added an interactive task, where different usage scenarios were tested. CHiC teamed up with Europeana[15], Europe's largest digital library, mu-

[15] http://www.europeana.eu

seum and archive for cultural heritage objects to provide a realistic environment for experiments. Europeana provided the document collection (digital representations of cultural heritage objects) and queries from their query logs.

3.1.13 Retrieval on Structured Datasets (2012–2014)

Traditional IR focuses on pure text retrieval over "bags of words" but the use of structure – such as document structure, semantic metadata, entities, or genre/topical structure is of increasing importance on the Web and in professional search. INEX was founded as the INitiative for the Evaluation of XML Retrieval and has been pioneering the use of structure for focused retrieval since 2002. It joined forces with CLEF in 2012 and continued this activity. From 2015 it merged into the Social Book Search Lab (see Section 3.1.19). A chapter by Kamps et al. in this volume discusses INEX activities.

3.1.14 Online Reputation Management (2012–2014)

Reputation management is an essential part of corporate communication. It comprises activities aiming at building, protecting and repairing the images of people, organizations, products, or services. It is vital for companies (and public figures) to maintain their good name and preserve their reputation capital. Current technology applications provide users with a wide access to information, enabling them to share it instantly and 24 hours a day due to constant connectivity. Information, including users' opinions about people, companies or products, is quickly spread over large communities. In this setting, every move of a company, every act of a public figure are subject, at all times, to the scrutiny of a powerful global audience. The control of information about public figures and organizations at least partly has moved from them to the users and consumers. For effective Online Reputation Management (ORM) this constant flow of online opinions needs to be watched. While traditional reputation analysis is mostly manual, online media allow to process, understand and aggregate large streams of facts and opinions about a company or individual. In this context, Natural Language Processing and text mining software play key, enabling roles. Although opinion mining has made significant advances in the last few years, most of the work has been focused on products. However, mining and interpreting opinions about companies and individuals is, in general, a much harder and less understood problem, since unlike products or services, opinions about people and organizations cannot be structured around any fixed set of features or aspects, requiring a more complex modeling of these entities.

RepLab was an initiative promoted by the EU project LiMoSINe, which aimed at studying reputation management as a living lab: a series of evaluation campaigns in which task design and evaluation methodologies are jointly carried out by researchers and the target user communities (reputation management experts). Given the novelty of the topic (as compared with opinion mining on product reviews and

mainstream topic tracking), it was felt that an evaluation campaign would maximize the use of the data collections built within LiMoSINe, encourage the academic interest in tasks with practical relevance, and promote the standardization of evaluation methodologies and practices in the field. RepLab, therefore, set out to bring together the Information Access research community with representatives from the ORM industry, aiming at: establishing a roadmap that included a description of the language technologies required in terms of resources, algorithms, and applications; specifying suitable evaluation methodologies and metrics; developing test collections that enable systematic comparison of algorithms and reliable benchmarking of commercial systems (Amigó et al, 2012).

The activities of RepLab are described in a chapter by Carrillo-de-Albornoz et al. in this volume.

3.1.15 eHealth (2012–2019)

Medical content is becoming increasingly available electronically in a variety of forms ranging from patient records and medical dossiers, scientific publications and health-related websites to medical-related topics shared across social networks. Laypeople, clinicians and policy-makers need to be able to easily retrieve, and make sense of this content to support their decision making. Information retrieval systems have been commonly used as a means to access health information available online. However, the reliability, quality, and suitability of the information for the target audience varies greatly while high recall or coverage, that is finding all relevant information about a topic, is often as important as high precision, if not more. Furthermore, information seekers in the health domain also experience difficulties in expressing their information needs as search queries[16].

The main objective of CLEF eHealth is thus to promote the development of information processing techniques that will assist the information provider and seeker to manage and retrieve electronically archived medical documents. The activities of this Lab are described in a chapter by Suominen et al. in this volume.

3.1.16 Biodiversity Identification and Prediction (2014–2019)

The LifeCLEF Lab aims at boosting research on the identification and prediction of living organisms in order to solve the taxonomic gap and improve our knowledge of biodiversity.

Building accurate knowledge of the identity, the geographic distribution and the evolution of living species is essential for a sustainable development of humanity as well as for biodiversity conservation. Unfortunately, such basic information is often only partially available for professional stakeholders, teachers, scientists and citizens, and is often incomplete for ecosystems that possess the highest diversity.

[16] See https://sites.google.com/view/clef-ehealth-2018/home

A noticeable consequence of this sparse knowledge is that the precise identification of living plants or animals is usually impossible for the general public, and often difficult for professionals, such as farmers, fish farmers or foresters and even also for the naturalists and specialists themselves. This taxonomic impediment was actually identified as one of the main ecological challenges to be solved during the United Nations Conference in Rio de Janeiro in 1992. In this context, an ultimate ambition is to set up innovative information systems relying on the automated identification and understanding of living organisms as a means to engage massive crowds of observers and boost the production of biodiversity and agro-biodiversity data[17].

Through its biodiversity informatics related challenges, LifeCLEF aims at pushing the boundaries of the state-of-the-art in several research directions at the frontier of multimedia information retrieval, machine learning and knowledge engineering with a focus on species identification using images for plants, audio for birds, and video for fishes.

In 2019 the LifeCLEF Lab proposes three data-oriented challenges related to this vision, in continuity with previous editions of the Lab:

- PlantCLEF aims at evaluating image-based plant identification on 10K species;
- BirdCLEF aims at evaluating bird species detection in audio soundscapes;
- GeoLifeCLEF aims at evaluating location-based prediction of species based on environmental and occurrence data.

The chapter by Joly et al. in this volume describes the activities of LifeCLEF.

3.1.17 News Recommendation Evaluation (2014–2017)

The NewsREEL Lab at CLEF provided the opportunity to evaluate algorithms both based on live data and offline simulated streams. The development of recommender services based on stream data is a challenging task. Systems optimized for handling streams must ensure highly precise recommendations taking into account the continuous changes in the stream as well as changes in the user preferences. In addition the technical complexity of the algorithms must be considered ensuring the seamless integration of recommendations into existing applications as well as ensuring the scalability of the system. Researchers in academia often focus on the development of algorithms only tested using static datasets due to the lack of access to live data. The benchmarking of the algorithms in the NewsREEL Lab considered both the recommendation precision (measured by the ClickThrough-Rate) and technical aspects (measured by reliability and response time) (Lommatzsch et al, 2017).

The chapter by Hopfgartner et al. in this volume includes a description of the activities of NewsReel.

[17] See https://www.imageclef.org/lifeclef2019

3.1.18 Living Labs (2015–2016)

In recent years, a new evaluation paradigm known as *Living Labs* has been proposed. The idea is to perform experiments in situ, with real users doing real tasks using real-world applications. Previously, this type of evaluation had only been available to (large) industrial research labs. The main goal with the Living Labs for IR Evaluation (LL4IR) Lab at CLEF was to provide a benchmarking platform for researchers to evaluate their ranking systems in a live setting with real users in their natural task environments. The Lab acted as a proxy between commercial organizations (live environments) and lab participants (experimental systems), facilitated data exchange, and made comparison between the participating systems. This initiative was a first of its kind for IR. It dealt with evaluation of ranking systems in a live setting with real users in their natural task environments.

The chapter by Hopfgartner et al. in this volume details the activities of Living Labs.

3.1.19 Social Book Search (2015–2016)

The goal of the Social Book Search (SBS) Lab was to evaluate approaches to support users in searching collections of books. The SBS Lab investigated the complex nature of relevance in book search and the role of traditional and user generated book metadata in retrieval. The aims were 1) to develop test collections to evaluate systems in terms of ranking search results and 2) to develop user interfaces and conduct user studies to investigate book search in scenarios with complex information needs and book descriptions that combine heterogeneous information from multiple sources. Techniques were studied to support users in complex book search tasks that involved more than just a query and results list, relying on semi-structured and highly structured data. The Lab included an interactive task which was a result of a merge of the INEX Social Book Search track and the Interactive task of CHiC. User interaction in social book search was gauged by observing user activity with a large collection of rich book descriptions under controlled and simulated conditions, aiming for as much real-life experiences as possible intruding into the experimentation. The aim was to augment the other Social Book Search tracks with a user-focused methodology. This Lab is discussed in the chapter by Kamps et al. in this volume.

3.1.20 Microblog Cultural Contextualization (2016–2018)

The MC2 lab mainly focused on developing processing methods and resources to mine the social media sphere and microblogs surrounding cultural events such as festivals, concerts, books, movies and museums, dealing with languages, dialects and informal expressions. The underlying scientific problems concern both IR and the Humanities.

The Lab began with a pilot activity in 2016. This examined the contextualization of data collected on the Web, and the search of content captured or produced by internet users. Participants were given access to a massive collection of microblogs and related urls to work with. The MC2 Lab at CLEF 2017 dealt with how the cultural context of a microblog affects its social impact at large. This involved microblog search, classification, filtering, language recognition, localization, entity extraction, linking open data, and summarization. Participants had access to the massive multilingual microblog stream of The Festival Galleries project. Microblog search topics were in four languages: Arabic, English, French and Spanish, and results were expected in any language.

In the 2018 Lab, two main tasks were offered: cross-language cultural microblog search; and argumentation mining. The first task was specific to movies. Topics were extracted from the French VodKaster website that allows readers to get personal short comments (microcritics) about movies. The challenge was to find related microblogs in four different languages in a large archive. The second task was about argumentation mining, a new problem in corpus-based text analysis that addresses the challenging task of automatically identifying the justifications provided by opinion holders for their judgment. The idea was to perform a search process on a massive microblog collection that focused on claims about a given festival. More details can be found in the chapter by Kamps et al. in this volume.

3.1.21 Dynamic Search for Complex Tasks (2017–2018)

DynSe, the CLEF Dynamic Tasks Lab, attempted to focus attention towards building a bridge between batch TREC-style evaluation methodology and Interactive Information Retrieval evaluation methodology - so that dynamic search algorithms can be evaluated using reusable test collections.

Information Retrieval research has traditionally focused on serving the best results for a single query - so-called ad-hoc retrieval. However, users typically search iteratively, refining and reformulating their queries during a session. IR systems can respond to each query in a session independently of the history of user interactions, or alternatively adopt their model of relevance in the context of these interactions. A key challenge in the study of algorithms and models that dynamically adapt their response to a user's query on the basis of prior interactions is the creation of suitable evaluation resources and the definition of suitable evaluation metrics to assess the effectiveness of such IR algorithms. Over the years, various initiatives have been proposed which have tried to make progress on this long standing challenge. However, while significant effort has been made to render the simulated data as realistic as possible, generating realistic user simulation models remains an open problem (Kanoulas and Azzopardi, 2017).

In its first edition, the Dynamic Search lab ran in the form of a workshop with the goal of addressing one key question: how can we evaluate dynamic search algorithms, commonly used by personalized session search, contextual search, and dialog systems. The workshop provided an opportunity for researchers to discuss the

challenges faced when trying to measure and evaluate the performance of dynamic search algorithms, given the context of available corpora, simulation methods, and current evaluation metrics. To seed the discussion, a pilot task was run with the goal of producing search agents that could simulate the process of a user, interacting with a search system over the course of a search session. The outcomes of the workshop were used to define the tasks of the 2018 Lab.

3.1.22 Early Risk Prediction on the Internet (eRisk, 2017–2019)

This Lab is exploring evaluation methodologies and effectiveness metrics for early risk detection on the Internet (in particular risks related to health and safety). The challenge consists of sequentially processing pieces of evidence from social media and microblogs and detecting, as soon as possible, early traces of diseases, such as depression or anorexia. 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. The main goal is to pioneer a new interdisciplinary research area, 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.

3.1.23 Evaluation of Personalised Information Retrieval (2017–2019)

The objective of the PIR-CLEF Lab is to develop and demonstrate the effectiveness of a methodology for the repeatable evaluation of Personalised Information Retrieval (PIR). PIR systems are aimed at enhancing traditional IR systems to better satisfy the information needs of individual users by providing search results that are not only relevant to the query but also to the specific user who submitted the query. In order to provide a personalised service, a PIR system maintains information about the user and their preferences and interests. These personal preferences and interests are typically inferred through a variety of interactions modes between the user with the system. This information is then represented in a user model, which is used to either improve the user's query or to re-rank a set of retrieved results so that documents that are more relevant to the user are presented in the top positions of the ranked list. Existing work on the evaluation of PIR has generally relied on a user-centered approach, mostly based on user studies; this approach involves real users undertaking search tasks in a supervised environment. While this methodology has the advantage of enabling the detailed study of the activities of real users, it has the significant drawback of not being easily reproducible and does not support the extensive exploration of the design and construction of user models and their exploitation in the search process. These limitations greatly restrict the scope for algorithmic exploration in PIR. This means that it is generally not possible to make

definitive statements about the effectiveness or suitability of individual PIR methods and meaningful comparison between alternative approaches (Pasi et al, 2017).

The PIR-CLEF Lab began with a pilot task in 2017. This was undertaken by 10 users employing a clearly defined and novel methodology. Data was gathered on the activities of each participant during search sessions on a subset of the ClueWeb12 collection[18], including details of relevant documents as marked by the searchers. The intention was to allow research groups working on PIR to gain experience with and provide feedback on the proposed PIR evaluation methodology. The input from the pilot task was used in the definition of the methodology employed in the 2018 and 2019 Labs. The Labs provide a framework for the evaluation of *Personalized Information Retrieval (PIR)*: 1) to facilitate comparative evaluation by offering participating research groups a mechanism for the evaluation of their personalisation algorithms; 2) to give the participating groups the means to formally define and evaluate their own novel user profiling approaches for PIR.

This is the first evaluation benchmark based on the Cranfield paradigm in this research area, with the potential benefits of producing evaluation results that are easily reproducible.

3.1.24 Automatic Identification and Verification of Political Claims (2018–19)

The CheckThat! Lab aims at fostering the development of technology capable of both spotting and verifying check-worthy claims in political debates in English and Arabic. Investigative journalists and volunteers work hard trying to get to the root of a claim in order to present solid evidence in favor or against it. However, manual fact-checking is very time-consuming, and automatic methods have been proposed as a way of speeding-up the process. For instance, there has been work on checking the factuality/credibility of a claim, of a news article, or of an entire news outlet. However, less attention has been paid to other steps of the fact-checking pipeline, e.g., check worthiness estimation has been severely understudied as a problem. By comparing a claim against the retrieved evidence, a system can determine whether the claim is likely true or likely false (or unsure, if no supporting evidence either way could be found). CheckThat! aims to address these understudied aspects. It is fostering the development of technology capable of spotting check-worthy claims in English political debates in addition to providing evidence-supported verification of Arabic claims.

3.1.25 Reproducibility (2018–2019)

The goal of CENTRE@CLEF is to run a joint task across CLEF/NTCIR/TREC on reproducibility, a primary concern in many areas of science.

[18] https://lemurproject.org/clueweb12/

Information Retrieval is especially interested in reproducibility since it is a discipline strongly rooted in experimentation, where experimental evaluation represents a main driver of advancement and innovation. In 2015, the ECIR conference began a new track focused on the reproducibility of previously published results. This conference track led to 3-4 reproducibility papers accepted each year but, unfortunately, this valuable effort did not produce a systematic approach to reproducibility: submitting authors adopted different notions of reproducibility, they adopted very diverse experimental protocols, they investigated the most disparate topics, resulting in a very fragmented picture of what was reproducible and what not, and the results of these reproducibility papers are spread over a series of potentially disappearing repositories and Web sites. It is clear that there is a need and urgency for a systematic approach to reproducibility in IR. The joint task at CENTRE@CLEF challenges participants:

- to reproduce the best results of the best/most interesting systems in previous editions of CLEF/NTCIR/TREC by using standard open source IR systems;
- to provide the community with the additional components and resources that were developed to reproduce the results with the hope of improving existing open source systems.

4 IR Tools and Test Collections

CLEF activities over these last two decades have resulted in the creation of a considerable amount of valuable resources, extremely useful for many types of text processing and benchmarking activities in the IR domain. In this section, we provide some pointers with respect to their availability.

Much attention was paid in the first years of CLEF 1.0 to the processing requirements of different languages; these vary considerably depending on levels of morphological and syntactic complexity. This resulted in many comparative studies and the development of a variety of morphological processors (light and more aggressive stemmers), see the discussion in the chapter by Savoy and Braschler in this volume. Jacques Savoy also maintains an important site at the University of Neuchâtel which provides information on and links to many IR multilingual tools[19].

The test collections, created as a result of the diverse experimental evaluation initiatives conducted in CLEF represent the end results of much collaborative work aimed at providing understanding and insights into how system performances can best be improved and how progress can be achieved. As already stated, the CLEF evaluation campaigns have mainly adopted a comparative evaluation approach in which system performances are compared according to the Cranfield methodology (see the chapter by Voorhees for a description of Cranfield). The test collections produced are thus made up of documents, topics and relevance assessments. The topics are created to simulate particular information needs from which the systems derive

[19] http://members.unine.ch/jacques.savoy/clef/

the queries to search the document collections. System performance is evaluated by judging the results retrieved in response to a topic with respect to their relevance, and computing the relevant measures, depending on the methodology adopted by the Track/Lab. The chapter by Agosti et al. in this volume describes the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system which manages and provides access to much of the data used and produced within CLEF.

During the campaigns, participating groups are provided with access to the necessary data sets on signing a data agreement form which specifies the conditions of use. An objective of CLEF is that, at the end of an evaluation, the test collections produced should, whenever possible, be made available to the wider R&D community. Here below we give some examples of collections that are now publicly accessible. If you do not find what you were looking for, our advice is to contact the coordinators of the relevant Track or Lab to see if they can help you. Contact information can be found via the CLEF web site[20] and/or annual working notes[21].

4.1 ELRA Catalogue

A number of official CLEF Test Suites consisting of the data created for the monolingual, bilingual, multilingual and domain-specific text retrieval and question answering tracks in the CLEF 1.0 Campaigns are available, generally for a fee, in the catalogue of the European Language Resources Association (ELRA)[22]. These packages consist of multilingual document collections in many languages; step-by-step documentation on how to perform system evaluation; tools for results computation; multilingual sets of topics; multilingual sets of relevance assessments; guidelines for participants (in English); tables of the results obtained by the participants; publications. The following data collections are included:

- CLEF multilingual corpus of more than 3 million news documents in 14 European languages. This corpus is divided into two comparable collections: 1994-1995 - Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, Swedish; 2000-2002 - Basque, Bulgarian, Czech, English, Hungarian. These collections were used in the Ad-Hoc, and Question Answering packages.
- The GIRT-4 social science database in English and German (over 300,000 documents) and two Russian databases: the Russian Social Science Corpus (approx. 95,000 documents) and the Russian ISISS collection for sociology and economics (approx. 150,000 docs); Cambridge Sociological Abstracts in English (20,000 docs). These collections were used in the domain-specific package.

The ELRA catalog also lists test suites derived from CLEF eHealth activities. These packages contain data used for user-centred health information retrieval tasks

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

[21] In the CEUR Workshop Proceedings - http://http://ceur-ws.org/

[22] Information and conditions of purchase can be found at: http://catalog.elra.info/.

conducted at the CLEF eHealth Labs in 2013 and 2014 and include: a collection of medical-related documents in English; guidelines provided to the participants; queries generated by medical professionals in several languages; a set of manual relevance assessments; the official results obtained by the participants; working notes papers.

4.2 Some Publicly Accessible CLEF Test Suites

Many Labs make evaluation test suites available free-of-charge for research and system training purposes. Here below, we list what is currently available at the time of writing (April 2019).

- QA@CLEF: Question Answering
  In addition to what can be found on the ELRA Catalogue, datasets for advanced tasks are accessible at http://nlp.uned.es/clef-qa/repository/pastCampaigns.php
- PAN: Digital Text Forensics
  Datasets designed for Authorship, Author Profiling, Credibility Analysis, Deception Detection, and Text Reuse Detection tasks. Accessible at https://pan.webis.de/data.html.
- RepLab: Online Reputation Management
  - RepLab 2013: +500,000 reputation expert annotations on Twitter data, covering named entity disambiguation (filtering task), reputational polarity, topic detection and topic reputational priority (alert detection). Accessible at http://nlp.uned.es/replab2013/
  - RepLab 2014: additional annotations on RepLab 2013 tweets covering reputational dimensions of tweets (Products/Services, Innovation, Workplace, Citizenship, Governance, Leadership, and Performance) and author profiling: (i) identification of opinion makers and (ii) classification of author types (journalist, professional, authority, activist, investor, company or celebrity). Accessible at http://nlp.uned.es/replab2014/
- WePS: Web People Search
  WePS 3 included two tasks concerning the Web entity search problem:
  - Task 1 is related to Web People Search and focuses on person name ambiguity and person attribute extraction on Web pages;
  - Task 2 is related to Online Reputation Management (ORM) for organizations and focuses on the problem of ambiguity for organization names and the relevance of Web data for reputation management purposes. Test collections accessible at http://nlp.uned.es/weps/weps-3
  Previous WePS datasets are also accessible at http://nlp.uned.es/weps/weps-1/weps1-data and http://nlp.uned.es/weps/weps-2
- Social Book Search
  2.8 million book records in XML format. Accessible at http://social-book-search.humanities.uva.nl/

- Protest News
  - Annotated data from the publicly available English Reuters news text Corpus RCV1 will be made freely accessible. See Lab website for details.
- The ImageCLEF and LifeCLEF initiatives make a number of existing datasets for system training purposes. Full details and information concerning conditions of use of the following collections can be found at the ImageCLEF website, see https://www.imageclef.org/datasets.
  - ImageCLEF/IAPR TC 12 Photo Collection
  - Segmented IAPR dataset
  - The COLD Database: contains image sequences captured using a regular and omni-directional cameras mounted on different mobile robot platforms together with laser range scans and odometry data. Data recorded at three different indoor laboratory environments located in three different European cities under various weather and illumination conditions.
  - The IDOL2 Database: consists of 24 image sequences accompanied by laser scans and odometry data acquired using two mobile robot platforms, within an indoor laboratory environment consisting of five rooms of different functionality, under various illumination conditions and across a span of 6 months.
  - The INDECS Database: several sets of pictures taken indoors, in five rooms of different functionality under various illumination and weather conditions at different periods of time.
  - ImageCLEF VCDT test collections: test collections of the ImageCLEF Visual Concept Detection and Annotation Task (VCDT) from 2009-2011
  - ImageCLEF Wikipedia Image Retrieval Datasets - The Wikipedia image retrieval task ran as part of ImageCLEF for four years: 2008-2011.
- Other test collections used in ImageCLEF tasks are listed here:
  - 2012 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 253000 images. Accessible at http://doi.org/10.5281/zenodo.1038533
  - 2013 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 253000 images. Accessible at http://doi.org/10.5281/zenodo.257722
  - 2014 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 505122 images. Accessible at http://doi.org/10.5281/zenodo.259758
  - 2015 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 500000 images. Accessible at http://doi.org/10.5281/zenodo.1038547
  - 2016 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 500000 images. Accessible at http://doi.org/10.5281/zenodo.1038554
  - ImageCLEF 2016 Bentham Handwritten Retrieval Dataset: images of scanned pages of a manuscript and queries to retrieve. Language: English. Size: 363 pages train, 433 pages development, 200 pages test. Accessible at http://doi.org/10.5281/zenodo.52994

5 The CLEF Association

The CLEF Association[23] is an independent non-profit legal entity, established in October 2013 as a result of the activity of the PROMISE Network of Excellence[24], which sponsored CLEF from 2010 to 2013.

The Association has scientific, cultural and educational objectives and operates in the field of information access systems and their evaluation. Its mission is:

- to promote access to information and use evaluation;
- to foster critical thinking about advancing information access and use from a technical, economic and societal perspective.

Within these two areas of interest, the CLEF Association aims at a better understanding of the use and access to information and how to improve this. The two areas of interest translate into the following objectives:

- providing a forum for stakeholders with multidisciplinary competences and different needs, including academia, industry, education and other societal institutions;
- facilitating medium / long-term research in information access and use and its evaluation; increasing, transferring and applying expertise.

The CLEF Association currently plays a key role in CLEF by ensuring the continuity, self-sustainability and overall coordination. CLEF 2014 was the first edition of CLEF not supported in any way by a main European project but run on a totally volunteer basis with the support of the CLEF association membership fees paid by its multidisciplinary research community.

6 Impact

Shared evaluation campaigns have always played a central role in IR research. They have produced huge improvements in the state-of-the-art and helped solidify a common systematic methodology, achieving not only scholarly impact (Tsikrika et al, 2013, 2011; Thornley et al, 2011; Angelini et al, 2014) but also economic results (Rowe et al, 2010), estimated in a return-on-investment about 3-5 times the funding provided. The twenty years of CLEF campaigns have had a significant scientific impact on European and global research. This is documented in the chapter by Birger Larsen in the final part of this volume.

During their life-span, these large-scale campaigns also produce a huge amount of extremely valuable experimental data. This data provides the foundations for subsequent scientific production and system development and constitutes an essential reference for the literature in the field. Papers by Agosti et al., Müller and Hanbury,

[23] http://www.clef-initiative.eu/association

[24] http://www.promise-noe.eu/

and Potthast et al. in this volume explore the infrastructures developed in CLEF over the years to run the experiments and to manage the resulting experimental data. Section 4 provides information on the availability of many of the IR resources and test collections created as a result of CLEF experiments.

Up until the end of the 20th century, IR research was predominantly conducted on test collections in English. Thus, when we launched CLEF 1.0, one of our declared objectives was to stimulate research in our domain on collections in many different languages - not only English - and across language boundaries. As a European initiative our primary focus was on European languages. This is the topic of the chapter by Savoy and Braschler. This goal was so well achieved that in CLEF 2.0 we could almost state that *multilinguality* in European IR research activities is taken for granted; even if the main theme is *multimodality*, all of the CLEF 2.0 Labs handle data in more than one European language.

Another of our goals has been to impact not only academia but also industrial research. IR research can never be considered only at the theoretical level, clearly the over-riding factors are the requirements of society at large. An important step in this direction, which began in CLEF 1.0 with ImageCLEF medical retrieval experiments (see the chapter by Müller et al. in this volume) but has certainly been increasingly reinforced in CLEF 2.0, is the involvement of real world user communities. Thus, just to cite a few examples, we have seen collaborations with the intellectual property and patent search domain in CLEF-IP (see the chapter by Piroi and Hanbury), with health specialists in E-Health (Suominen et al. this volume), and with news portals in the NewsREEL project (see Hopfgartner et al.). The chapter by Jussi Karlgren in the final part of this volume discusses the challenges involved in applying evaluation benchmarks in operational settings. And this year, CLEF 2019 will host for the first time an Industry Day, jointly organized with the Swiss Alliance for Data-Intensive Services. The goal is to further open CLEF to a wider, industrial community through demo sessions, panels and special keynotes where the very best and most pertinent work of CLEF participants will be made more publicly accessible.

An aspect of CLEF of which we are particularly proud is the consolidation of a strong community of European researchers in the multidisciplinary context of IR. This year, for the first time, the *European Conference for Information Retrieval (ECIR)* and CLEF have joined forces: ECIR 2019 hosting a session dedicated to CLEF Labs where lab organizers present the major outcomes of their Labs and plans for ongoing activities, followed by a poster session in order to favour discussion during the conference. This is reflected in the ECIR 2019 proceedings, where CLEF Lab activities and results are reported as short papers. The goal is not only to engage the ECIR community in CLEF activities, but also to disseminate the research results achieved during CLEF evaluation cycles at ECIR. This collaboration will of course strengthen European IR research even more. However, this European community should not be seen in isolation. CLEF is part of a global community; we have always maintained close links with our peer initiatives in the Americas and Asia. There is a strong bond connecting TREC, NTCIR, CLEF and FIRE, and a continual, mutually beneficial exchange of ideas, experiences and results.

Despite the acknowledged success of CLEF and other evaluation campaigns over the years, we cannot rest on our laurels. It is fundamental to keep asking what new challenges need to be addressed in the future and how to continue to contribute to progress in the IR field. The chapters in the concluding part of this volume thus explore future perspectives: reproducibility of experiments by Norbert Fuhr, industrial involvement by Jussi Karlgren, and exploitation of Visual Analytics for IR evaluation by Ferro and Santucci.

Acknowledgements CLEF 2000 and 2001 were supported by the European Commission under the Information Society Technologies programme and within the framework of the DELOS Network of Excellence for Digital Libraries (contract no. IST-1999-12262).

CLEF 2002 and 2003 were funded as an independent project (contract no. IST-2000-31002) under the 5th Framework Programme of the European Commission.

CLEF 2004 to 2007 were sponsored by the DELOS Network of Excellence for Digital Libraries (contract no. G038-507618) under the 6th Framework Programme of the European Commission.

Under the 7th Framework Programme of the European Commission, CLEF 2008 and 2009 were supported by TrebleCLEF Coordination Action (contract no. 215231) and CLEF 2010 to 2013 were funded by the PROMISE Network of Excellence (contract no. 258191).

CLEF 2011 to 2015 also received support from the ELIAS network (contract no. 09-RNP-085) of the European Science Foundation (ESF) for ensuring student travel grants and invites speakers.

CLEF 2015, 2017, and 2018 received ACM SIGIR support for student travel grants through the SIGIR Friends program.

Over the years CLEF has also attracted industrial sponsorship: from 2010 onwards, CLEF has received the support of Google, Microsoft, Yandex, Xerox, Celi as well as publishers in the field such as Springer and Now Publishers.

In addition to the support gratefully acknowledged above, CLEF tracks and labs have frequently received the assistance of other projects and organisations; unfortunately, it is impossible to list them all here.

It must be noted that, above all, CLEF would not be possible without the volunteer efforts, enthusiasm, and passion of its community: lab organizers, lab participants, and attendees are the core and the real success of CLEF.

References

Amigó E, Corujo A, Gonzalo J, Meij E, de Rijke M (2012) Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1178/

Amigó E, Gonzalo J, Verdejo MF (2013) A General Evaluation Measure for Document Organization Tasks. In: Jones GJF, Sheridan P, Kelly D, de Rijke M, Sakai T (eds) Proc. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013), ACM Press, New York, USA, pp 643–652

Angelini M, Ferro N, Larsen B, Müller H, Santucci G, Silvello G, Tsikrika T (2014) Measuring and Analyzing the Scholarly Impact of Experimental Evaluation Initiatives. In: Agosti M, Catarci T, Esposito F (eds) Proc. 10th Italian Research Conference on Digital Libraries (IRCDL 2014), Procedia Computer Science, Vol. 38, vol 38, pp 133–137

Angelini M, Fazzini V, Ferro N, Santucci G, Silvello G (2018) CLAIRE: A combinatorial visual analytics system for information retrieval evaluation. Information Processing & Management 54(6):1077–1100

Bollmann P (1984) Two Axioms for Evaluation Measures in Information Retrieval. In: van Rijsbergen CJ (ed) Proc. of the Third Joint BCS and ACM Symposium on Research and Development in Information Retrieval, Cambridge University Press, UK, pp 233–245

Braschler M (2004) Combination Approaches for Multilingual Text Retrieval. Information Retrieval 7(1/2):183–204

Buckley C, Voorhees EM (2004) Retrieval Evaluation with Incomplete Information. In: Sanderson M, Järvelin K, Allan J, Bruza P (eds) Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), ACM Press, New York, USA, pp 25–32

Buckley C, Voorhees EM (2005) Retrieval System Evaluation. In: (Harman and Voorhees, 2005), pp 53–78

Busin L, Mizzaro S (2013) Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In: Kurland O, Metzler D, Lioma C, Larsen B, Ingwersen P (eds) Proc. 4th International Conference on the Theory of Information Retrieval (ICTIR 2013), ACM Press, New York, USA, pp 22–29

Büttcher S, Clarke CLA, Yeung PCK, Soboroff I (2007) Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements. In: Kraaij W, de Vries AP, Clarke CLA, Fuhr N, Kando N (eds) Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), ACM Press, New York, USA, pp 63–70

Carterette BA (2012) Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. ACM Transactions on Information Systems (TOIS) 30(1):4:1–4:34

Chapelle O, Metzler D, Zhang Y, Grinspan P (2009) Expected Reciprocal Rank for Graded Relevance. In: Cheung DWL, Song IY, Chu WW, Hu X, Lin JJ (eds) Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009), ACM Press, New York, USA, pp 621–630

Cleverdon CW (1967) The Cranfield Tests on Index Languages Devices. Aslib Proceedings 19(6):173–194

Di Nunzio GM, Leveling J, Mandl T (2011) LogCLEF 2011 Multilingual Log File Analysis: Language Identification, Query Classification, and Success of a Query. In: Petras V, Forner P, Clough P, Ferro N (eds) CLEF 2011 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1177/

Ferrante M, Ferro N, Maistro M (2015) Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In: Allan J, Croft WB, de Vries AP, Zhai C, Fuhr N, Zhang Y (eds) Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015), ACM Press, New York, USA, pp 21–30

Ferrante M, Ferro N, Pontarollo S (2017) Are IR Evaluation Measures on an Interval Scale? In: Kamps J, Kanoulas E, de Rijke M, Fang H, Yilmaz E (eds) Proc. 3rd ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2017), ACM Press, New York, USA, pp 67–74

Ferrante M, Ferro N, Pontarollo S (2019) A General Theory of IR Evaluation Measures. IEEE Transactions on Knowledge and Data Engineering (TKDE) 31(3):409–422

Ferro N, Harman D (2010) CLEF 2009: Grid@CLEF Pilot Track Overview. In: Peters C, Di Nunzio GM, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G (eds) Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany, pp 552–565

Ferro N, Silvello G (2016) A General Linear Mixed Models Approach to Study System Component Effects. In: Perego R, Sebastiani F, Aslam J, Ruthven I, Zobel J (eds) Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), ACM Press, New York, USA, pp 25–34

Ferro N, Silvello G (2017) Towards an Anatomy of IR System Component Performances. Journal of the American Society for Information Science and Technology (JASIST)

Ferro N, Maistro M, Sakai T, Soboroff I (2018) Overview of CENTRE@CLEF 2018: a First Tale in the Systematic Reproducibility Realm. In: Bellot P, Trabelsi C, Mothe J, Murtagh F, Nie JY, Soulier L, SanJuan E, Cappellato L, Ferro N (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Nineth International Conference of the CLEF Association (CLEF 2018), Lecture Notes in Computer Science (LNCS) 11018, Springer, Heidelberg, Germany, pp 239–246

Fuhr N (2012) Salton Award Lecture: Information Retrieval As Engineering Science. SIGIR Forum 46(2):19–28

Hanbury A, Müller H (2010) Automated Component-Level Evaluation: Present and Future. In: Agosti M, Ferro N, Peters C, de Rijke M, Smeaton A (eds) Multilingual and Multimodal Information Access Evaluation. Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF 2010), Lecture Notes in Computer Science (LNCS) 6360, Springer, Heidelberg, Germany, pp 124–135

Harman DK (2011) Information Retrieval Evaluation. Morgan & Claypool Publishers, USA

Harman DK, Voorhees EM (eds) (2005) TREC. Experiment and Evaluation in Information Retrieval, MIT Press, Cambridge (MA), USA

Harman DK, Braschler M, Hess M, Kluck M, Peters C, Schaüble P, Sheridan P (2001) CLIR Evaluation at TREC. In: (Peters, 2001), pp 7–23

Hopfgartner F, Hanbury A, Müller H, Eggel I, Balog K, Brodt T, Cormack GV, Lin J, Kalpathy-Cramer J, Kando N, Kato MP, Krithara A, Gollub T, Potthast M, Viegas E, Mercer S (2018) Evaluation-as-a-Service for the Computational Sciences: Overview and Outlook. ACM Journal of Data and Information Quality (JDIQ) 10(4):15:1–15:32

Hull DA (1993) Using Statistical Testing in the Evaluation of Retrieval Experiments. In: Korfhage R, Rasmussen E, Willett P (eds) Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), ACM Press, New York, USA, pp 329–338

Järvelin K, Kekäläinen J (2002) Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) 20(4):422–446

Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, Cappellato L, Ferro N (eds) (2017) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eighth International Conference of the CLEF Association (CLEF 2017), Lecture Notes in Computer Science (LNCS) 10456, Springer, Heidelberg, Germany

Kanoulas E, Azzopardi L (2017) CLEF 2017 Dynamic Search Evaluation Lab Overview. In: (Jones et al, 2017), pp 361–366

Kekäläinen J, Järvelin K (2002) Using Graded Relevance Assessments in IR Evaluation. Journal of the American Society for Information Science and Technology (JASIST) 53(13):1120—1129

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

Lommatzsch A, Kille B, Hopfgartner F, Larson M, Brodt T, Seiler J, Özgöbek Ö (2017) CLEF 2017 NewsREEL Overview: A Stream-Based Recommender Task for Evaluation and Education. In: (Jones et al, 2017), pp 239–254

McNamee P, Mayfield J (2004) Character N-Gram Tokenization for European Language Text Retrieval. Information Retrieval 7(1-2):73–97

Mizzaro S (1997) Relevance: The Whole History. Journal of the American Society for Information Science and Technology (JASIST) 48(9):810–832

Moffat A, Zobel J (2008) Rank-biased Precision for Measurement of Retrieval Effectiveness. ACM Transactions on Information Systems (TOIS) 27(1):2:1–2:27

Nardi A, Peters C, Vicedo JL, Ferro N (eds) (2006) CLEF 2006 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1172/

Nardi A, Peters C, Ferro N (eds) (2007) CLEF 2007 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1173/

Pasi G, Jones GJF, Marrara S, Sanvitto C, Ganguly D, Sen P (2017) Overview of the CLEF 2017 Personalised Information Retrieval Pilot Lab (PIR-CLEF 2017). In: (Jones et al, 2017), pp 338–345

Peters C (ed) (2001) Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000), Lecture Notes in Computer Science (LNCS) 2069, Springer, Heidelberg, Germany

van Rijsbergen CJ (1974) Foundations of Evaluation. Journal of Documentation 30(4):365–373

Rowe BR, Wood DW, Link AL, Simoni DA (2010) Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. RTI Project Number 0211875, RTI International, USA. http://trec.nist.gov/pubs/2010.economic.impact.pdf

Sakai T (2006) Evaluating Evaluation Metrics based on the Bootstrap. In: Efthimiadis EN, Dumais S, Hawking D, Järvelin K (eds) Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006), ACM Press, New York, USA, pp 525–532

Sakai T (2012) Evaluation with Informational and Navigational Intents. In: Mille A, Gandon FL, Misselis J, Rabinovich M, Staab S (eds) Proc. 21st International Conference on World Wide Web (WWW 2012), ACM Press, New York, USA, pp 499–508

Sakai T (2014a) Metrics, Statistics, Tests. In: Ferro N (ed) Bridging Between Information Retrieval and Databases - PROMISE Winter School 2013, Revised Tutorial Lectures, Lecture Notes in Computer Science (LNCS) 8173, Springer, Heidelberg, Germany, pp 116–163

Sakai T (2014b) Statistical Reform in Information Retrieval? SIGIR Forum 48(1):3–12

Sanderson M (2010) Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval (FnTIR) 4(4):247–375

Saracevic T (1975) RELEVANCE: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science and Technology (JASIST) 26(6):321–343

Savoy J (1997) Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management 33(44):495–512

Spärck Jones K (ed) (1981) Information Retrieval Experiment, Butterworths, London, United Kingdom

Thornley CV, Johnson AC, Smeaton AF, Lee H (2011) The Scholarly Impact of TRECVid (2003–2009). Journal of the American Society for Information Science and Technology (JASIST) 62(4):613–627

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

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

Voorhees EM, Harman DK (1998) Overview of the Seventh Text REtrieval Conference (TREC-7). In: Voorhees EM, Harman DK (eds) The Seventh Text REtrieval Conference (TREC-7), National Institute of Standards and Technology (NIST), Special Publication 500-242, Washington, USA, pp 1–24

Yilmaz E, Aslam JA (2006) Estimating Average Precision With Incomplete and Imperfect Judgments. In: Yu PS, Tsotras V, Fox EA, Liu CB (eds) Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006), ACM Press, New York, USA, pp 102–111

Index