EaaS: Evaluation-as-a-Service and Experiences from the VISCERAL project

Henning Müller and Allan Hanbury

Abstract The Cranfield paradigm has dominated information retrieval evaluation for almost 50 years. It has had a major impact on the entire domain of information retrieval since the 1960s and, compared with systematic evaluation in other domains, is very well developed and has helped very much to advance the field. This chapter summarizes some of the shortcomings in information analysis evaluation and how recent techniques help to leverage these shortcomings. The term Evaluation-as-a-Service (EaaS) was defined at a workshop that combined several approaches that do not distribute the data but use source code submission, APIs or the cloud to run evaluation campaigns. The outcomes of a white paper and the experiences gained in the VISCERAL project on cloud-based evaluation for medical imaging are explained in this paper. In the conclusions, the next steps for research infrastructures are imagined and the impact that EaaS can have in this context to make research in data science more efficient and effective.

1 Introduction

Information retrieval evaluation has largely followed the Cranfield paradigm over the last more than 50 years (Cleverdon et al, 1966; Cleverdon, 1962). This means that an information retrieval test collection consisting of documents is created, then topics are defined on the test collection and ground truthing is performed to determine an optimal response. The ground truth can be the relevance of all documents in the collection but is usually only for part of it, done using pooling techniques. The Cranfield tests helped to identify that automatic indexing of terms had as good or better performance as manually attached keywords of experts by system-

Henning Müller
HES–SO Valais, Sierre, Switzerland, e-mail: henning.mueller@hevs.ch

Allan Hanbury
TU Wien, Austria, e-mail: allan.hanbury@tuwien.ac.at

atic analysis, so their impact on the field of text retrieval was very important. On the basis of this paradigm, information retrieval developed systematic approaches for retrieval system evaluation very early (Jones and van Rijsbergen, 1975; Salton, 1971) and could thus demonstrate steady scientific progress and also develop commercial applications in many fields by showing the performance obtained over the years. Evaluation campaigns such as TREC[1] (Text REtrieval Conference) (Harman, 1992) or CLEF[2] (Cross-Language Evaluation Forum) (Braschler and Peters, 2002) developed yearly evaluation cycles for a variety of domains and application scenarios of textual information retrieval. Visual information retrieval systems have been evaluated in ImageCLEF[3] (Image retrieval tasks of CLEF) (Kalpathy-Cramer et al, 2015; Müller et al, 2010) and also the TRECvid campaigns for many years (Smeaton et al, 2003). The impact of these evaluation campaigns both in terms of commercial impact and scholarly impact have also been analyzed for TREC (Rowe et al, 2010), TrecVid (Video retrieval tasks of TREC) (Thornley et al, 2011) and CLEF/ImageCLEF (Tsikrika et al, 2011, 2013). All of these analyses have shown that the commercial impact is massive and that national agencies save much money by supporting such campaigns by sharing resources and allowing larger and more impactful evaluations. The scholarly impact was considered very important, as publicly available test collections foster data reuse and many of the overview papers of the popular evaluation campaigns are highly cited. Participant papers with good results also often get a high citation count as the techniques are often reimplemented or reused (particularly if the source code is also made available). There are also several criticisms of evaluation campaigns, particularly when the tasks are artificial and not related to user needs (Forsyth, 2002), or that benchmarking favors small changes to existing algorithms over completely new techniques. In general, challenges improve the performance and not doing any evaluation mainly means that performance improvements cannot be measured. ImageNET (Deng et al, 2009) has also shown that large scale evaluation can lead to disruptive changes, in this case on the use of deep learning for computer vision and object recognition (Krizhevsky et al, 2012).

Evaluation campaigns have many other shortcomings. Even though the topics are usually created with clear user models in mind, often based on surveys or log file analysis (Markonis et al, 2012, 2013; Müller et al, 2007), they only measure static behavior, so changes in user targets or the impact of a user interface and of interaction cannot be measured. For this reason interactive retrieval evaluation was introduced into TREC and CLEF (Borlund and Ingwersen, 1997; Gonzalo et al, 2005) to make it possible to measure the human factor in text and image retrieval, which is extremely important for building working systems. Such evaluation is much harder than with static collections but it is very complementary and much can be learned. Even with existing test collections and evaluation campaigns, several problems remain and have been reported in the literature. It is often easier to adapt existing data sets to make the challenges easier than to actually improve the techniques (Müller

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

[2] http://www.clef-campaign.org/

[3] http://www.imageclef.org/

et al, 2002). Even when standard data sets exist, the use of them and the increased performance reported often does not add up (Armstrong et al, 2009b). This is linked to the fact that existing baselines are not well defined and often comparison is not done against the best systems in an evaluation campaign but to less well performing own algorithms, or they are compared against older results, even though better results have been published later on (comparison to a low baseline). (Blanco and Zaragoza, 2011) shows another problem: when trying many different approaches and manually optimizing them on the data it is possible to get better results but these results are often meaningless, even though they are statistically significant. It is impossible to reproduce such results and statistically speaking significance rules should in this case be multiple and not single hypotheses testing. Often, only positive results are reported and not exactly how these results were obtained. (Ioannidis, 2005) even goes a step further, stating that most research findings published are false, as small sample sizes are used and often incorrect statistics and strong publication bias towards positive results add to this. Interestingly, the more a domain is competitive the more the results are false, as quick publication and pressure lead to increased incorrectness.

To overcome some of the mentioned problems, several infrastructures have been proposed in the past. In (Müller et al, 2001), an online evaluation with a retrieval communication protocol was implemented but it was never used by a large number of systems, even though such a system could guarantee a high level of reproducibility because the ground truth is never released and manual optimizations are difficult. The EvaluatIR initiative (Armstrong et al, 2009a) also developed a framework for evaluation where components could be shared and separately evaluated. Again, such a system could save much time and effort but it was discontinued after only a short period of time. Actually understanding the interplay of the many components of a retrieval system has also been subject of detailed research in the past (Hanbury and Müller, 2010). Still, many of the approaches never reached a critical mass and non-integrated systems with intermediate results were often inefficient, and thus not taken up by either researchers or commercial partners. Academic systems to manage evaluation campaigns such as DIRECT have also been used and shown their usefulness (Agosti et al, 2012; Silvello et al, 2017). For good reproducibility both data (Silvello, 2018) and code (Niemeyer et al, 2016) need to become citable and reusable easily (Mayernik et al, 2016). Beyond the academic field, machine learning has shown the need for evaluation infrastructures, highlighted by the commercial success of the Kaggle[4] platform and several similar initiatives, for example TopCoder[5] and CrowdAI[6] that addresses a more open type of challenges in machine learning.

[4] http://www.kaggle.com/

[5] http://www.topcoder.com/

[6] http://www.crowdai.org/

2 Evaluation-as-a-Service

The actual term EaaS[7] (Evaluation-as-a-Service) was detailed in a workshop in early 2015 (Hopfgartner et al, 2015). 12 researchers of several evaluation initiatives and institutions met in Sierre, Switzerland, and discussed their approaches for providing Evaluation-as-a-Service in several environments and with differing approaches. EaaS in this case means that no data sets are distributed but the evaluation part of the campaign is provided as a service. Figure 1 details some of the outcomes of the workshop; also made available in a white paper (Hanbury et al, 2015) in a much more detailed form and with many references to the relevant literature. In (Hopfgartner et al, 2018), a shorter version of the main aspects of EaaS was published.

Fig. 1: Overview of several aspects and the set of stakeholders in the EaaS field; the figure shows the large number of possibly implied groups and roles (image taken from (Hanbury et al, 2015)).

Figure 1 shows the many implications that EaaS has and the various stakeholders in the evaluation environment. Whereas challenge stakeholders can range from chal-

[7] http://eaas.cc/

lenge organizers and challenge participants (academic and commercial researchers), there are several roles that are often not described in much detail, such as the data providers and human annotators (who might have an important problem that they would like to see solved). In terms of EaaS, infrastructure providers play an important role, for example cloud providers, and also funding agencies that could see important gains in a more efficient global research infrastructure. A similar diversity can be seen in the types of technologies used, the policy aspects of challenges and their infrastructures and also the business parts of it, as all of these can have a strong impact on how data are shared and how well the science advances. EaaS research usually mentions only a few domains of high potential but all these aspects can be taken into account for a full picture. Following the initial workshop, a second workshop was organized in Boston, MA, USA, in November 2015, focusing on the distributed and cloud aspects of the evaluation and with a focus also on medical applications, which is one of the use cases for EaaS with a clearly visible potential (Müller et al, 2016). The various stakeholders from funding agencies, infrastructure and use case providers were invited to this workshop to get a clearer idea of the roles of these partners and their interest.

The example cases that were used as the basis of the white paper include the use of an API for challenges, as in the TREC Microblog task (Ounis et al, 2011) that uses Twitter data that cannot be distributed in another way for copyright reasons. The TIRA[8] system uses code submission and then runs the code on the test data in a sandboxed environment on the university servers (Gollub et al, 2012). The CLEF NewsReel task (Hopfgartner et al, 2014) also relies on an API, this time of a news recommendation web page that adds recommendations provided by participating systems to the real recommendations and measures the number of clicks in these provided items to evaluate the quality of the results. VISCERAL (Hanbury et al, 2012) uses virtual machines in the cloud to run a challenge and only a small data set can be seen by the participants. In C-BIBOP[9] (Cloud-Based Image BiOmarker Platform), Docker containers were used to bring the algorithms to the data in a more lightweight way compared to virtual machines. Finally, the BioAsq project is directly included in the process of assigning MeSH terms to new texts and then compares these with the terms that are manually attached to the texts (Tsatsaronis et al, 2015).

Several other challenges have used these concepts as well, for example the Mammography Dream challenge (Trister et al, 2017) that made an unprecedented amount of data available for research in a cloud environment. The Mammography Dream challenge was run in several submission phases that had as objective to improve the initial results. They fostered particularly in the later collaborative phase of the challenge a strong collaboration among the research groups that obtained the best results in the previous phases.

[8] http://tira.io/

[9] http://cbibop.github.io/

3 The VISCERAL Experience

The ideas of the VISCERAL[10] (VISual Concept Extraction challenge in RAdioLogy) project were developed on the basis of several challenges and problems of previous evaluation initiatives that the project partners had encountered, notably:

- it is difficult to move *very large data sets*, as even in the Terabyte (TB) range it is currently hard to download data. Sending hard disks through the postal service also becomes cumbersome and prone to physical errors;
- *confidential data* can often not be shared easily but only after manual checking, which becomes infeasible for very large data sets, for example limiting medical data sharing in many cases;
- *quickly changing data sets* cannot be evaluated, as the time to prepare test collections and then transmit them to researchers and obtain results is often already too long and new data have become available in the meantime that need to be taken into account for a final evaluation. This would require a system where algorithms can be run again when new data become available, to always work on the latest data and know what works best.

All these challenges support the idea of moving the algorithms towards the data set rather than the current practice to moving the data to the researchers and their algorithms (Hanbury et al, 2012). The initial idea was to use a cloud infrastructure, in our case the Azure system, to store the entire data set of medical imaging data and only make a small part of it available to researchers directly, keeping the remaining data only accessible to the algorithms and not to the researchers (Langs et al, 2012). Algorithms have actually become much more mobile than the increasingly large data sets. Figure 2 shows the first step of the process, where each participant obtains access to a small data set in the cloud via a virtual machine (VM). This data set makes it possible to get used to the data format and the virtual machine makes it possible to install all necessary tools and test them on the small data set. VMs with both Linux and Windows were available for the participants to avoid creating any limitations. Algorithms and scripts can be tested on the available data, so they will then run automatically on the unknown data.

The second part of the VISCERAL challenge is shown in Figure 3. Once the algorithms are ready, then the participant can submit the virtual machine and the proposed algorithms are run on the protected test data. The participants do not have any further access to the machine and only the challenge organizers can access the system to run the algorithms installed on the larger data set. These data thus never get exposed to the researchers and by running the VMs in a sandboxed environment with all communication closed, no data can be communicated out, even if the researchers installed such code on the VMs.

In the case of the VISCERAL benchmarks, the training sets were relatively small and these were made fully available to the participants and the test sets were the large data set (Jimenez-del-Toro et al, 2016). It would also be possible to have training

[10] `http://www.visceral.eu/`

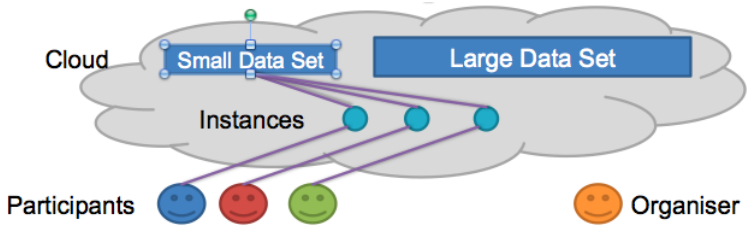

Fig. 2: The participants each have their own computing instance (VM) in the cloud, linked to a small data set of the same structure as the large one. Software for carrying out the competition objectives is placed in the instances by the participants. The large data set is kept separate. (image taken from (Hanbury et al, 2012))

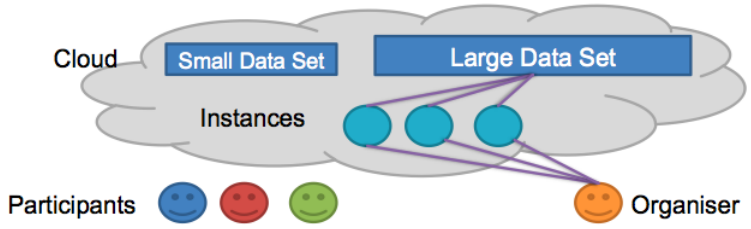

Fig. 3: On the competition deadline, the organiser takes over the instances containing the software written by the participants, links them to the large data set, performs the calculations and evaluates the results. (image taken from (Hanbury et al, 2012))

data in the larger data set and thus run training and testing in the cloud directly autonomously. Many more details on the experiences gained in the VISCERAL project can be found in a published book on the project (Hanbury et al, 2017). This book details all challenges that were run (lesion detection, similar case retrieval, organ segmentation), the experiences gained in data annotation and quality controls. The final system to manage the data and challenges can be seen in Figure 4. The entire data annotation in 3D volumes was also run in the cloud including quality control and several double annotations to measure subjectivity of the tasks (to compare algorithm outcome with the quality of human annotation). A fully automatic evaluation system was developed including the extraction of many performance measures and an automatically generated leaderboard. When a participant submits a new algorithm all steps are executed automatically and the participant is asked to publish or not the results in the continuous leaderboard of the task.

The system developed also allows many additional possibilities for exploiting the data and the algorithms. Having the algorithms of 15 segmentation tools as executables makes it possible to also run them on new data for which no annotations or ground truth exist. Combining results of several automatic algorithms with label fusion leads to much better results than any single algorithm and we called this out-

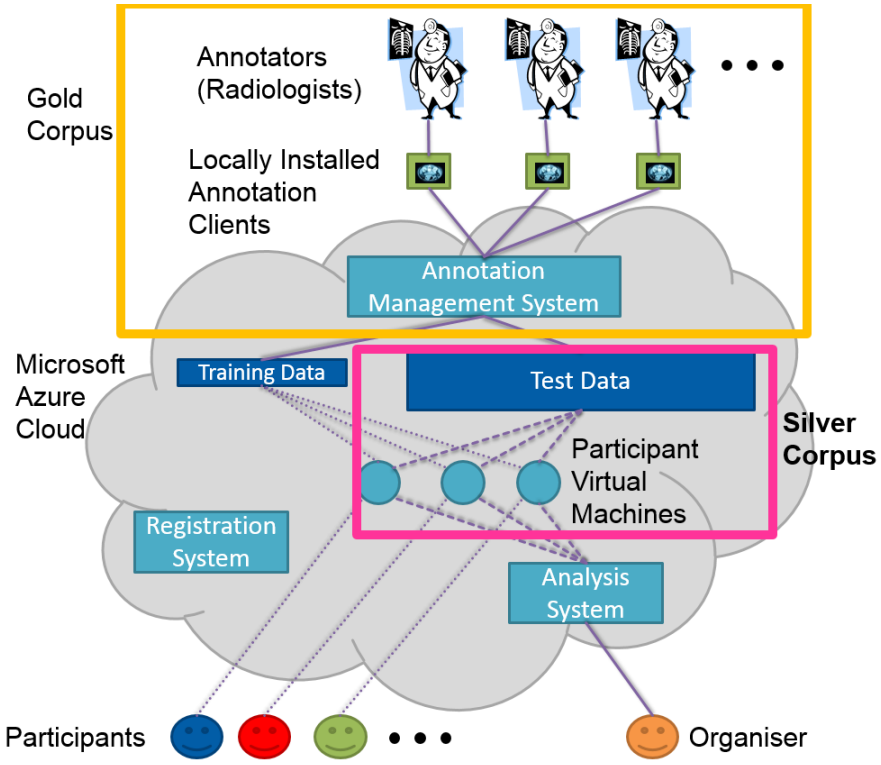

Fig. 4: Final overview of the VISCERAL infrastructure including a system for data annotation and quality control in the cloud and in addition to the cloud-based evaluation (image taken from (Hanbury et al, 2015))

come a Silver Corpus (Krenn et al, 2016) and showed that it has a very good quality for most of the organs that were segmented in the challenge. This Silver Corpus is now made available to other researchers and can be used to train algorithms such as deep learning algorithms that require large amounts of training data. Thus, it possibly improves results of future systems that can use the additional training data. This also makes it possible to manually annotate only those cases where the algorithms have the highest disagreement, as this can limit the annotation costs with a maximum information gain. Such active learning can massively reduce the efforts of manual annotation and still generate very large and meaningful training data sets.

In general the risk of medical data being exposed to research is not the data itself but rather the risk of matching several data sources or databases, which can possibly allow a re-identification of patients. In the case of VISCERAL no human sees the larger data set, only the algorithms, thus limiting ethical risks to an absolute minimum. Developing such models inside hospitals can even limit the risks further (if security mechanisms for running the code in a protected environment exist). With

Docker a light-weight container technology is available that allows to move code and all its dependencies. Hospitals can thus make data of their clinical challenges available for researchers and take advantage of knowledge gained by using the best algorithms. In any case, hospitals will require large computing and storage infrastructures with the advent of digital medicine. Genomics but also the analysis of imaging data and other data sources for decision support will require strong computation in the future, and part of this can also be used for EaaS by sharing data and tasks with the academic world. Even multi-center studies can be envisioned where aggregated data from each client site are combined in a central location. This can be particularly interesting for rare diseases where each institution only encounters few cases over the study period.

4 Conclusions

This chapter summarizes the concepts of EaaS and the implications that this can have for scientific challenges in data science, far beyond the initial targets of information retrieval, as in CLEF. Many problems in the academic field such as an exploding number of publications (that are impossible to follow even in a very narrow field (Fraser and Dunstan, 2010)) and also the impression that many current publications are not fully correct (Ioannidis, 2005) motivate the feeling that new infrastructures for academic research and particularly in data science are necessary. Full reproducibility needs to be created for publications in this field and with executable code containers and digital information this is possible, even in the long term. Storing and linking parameters of all experiments is very important (Di Nunzio and Ferro, 2005) to be able to learn a maximum amount from existing experiments. With EaaS, the executable code is available and can be kept in a light-weight manner, via Docker containers, for example. Data are available including ground truth, topics and scripts for running and evaluating tools on the data. If new data become available or if errors in the data are found the code can simply be rerun and all results can be updated.

Besides the credibility of results, it also seems important to make data science more efficient by building fully on the results of other researchers and not reinventing things with minor variations over and over. Good infrastructures can also help with this by facilitating the sharing of code and favoring challenges that allow collaboration between researchers. Working on the same infrastructures and in a similar framework can make sharing code much easier.

Wide availability of all data sets and sharing of recent results on the same data in a simple way can also make it mandatory to compare algorithms with strong baselines, and with the best results obtained on the same data and task at any given time. This would make the quality and possible impact of publications presenting improved results much stronger. It also encourages moving away from focusing purely on quantitative outcomes and rather on interpretation of the potential of techniques and

possibly also publishing negative findings, which would help to limit publication bias.

In the current data science environment, most often the groups with the biggest hardware have a strong advantage, as they can optimize parameters of more complex models on more training data and often obtain better results. By running challenges in a central infrastructure all participants have access to the same computational power, so this disadvantage would vanish and it would even make it possible to compare algorithm effectiveness and efficiency, as usually a tradeoff between the two is the main objective. With many research centers now using virtual machines and virtualized environments, it does not seem to matter too much anymore where the physical servers are, as long as they remain accessible and if quick access to the data is possible. In such an environment, EaaS is a very natural choice. Using Docker containers instead of VMs is in our experience a big advantage, as the installation overhead is very low and portability is much higher than with VMs.

In the future we expect evaluation campaigns such as CLEF to have access to their own research infrastructures and to make them available for participating researchers. The question of who bears the costs for data storage and computation need to be answered. Overall, the costs will be lower, so funding bodies should have a strong interest in having such a framework installed for scientific research. On a European level, the EOSC (European Open Science Cloud) is a candidate to supply storage and computation for such an approach in data science challenges. Likely there are still barriers to this approach but it is a question of time before models similar to EaaS will become the most common form of performing data science.

Acknowledgements The work leading to the chapter was partly funded by the EU FP7 program in the VISCERAL project and the ESF via the ELIAS project. We also thank all the participants of the related workshops for their input and the rich discussions.

References

Agosti M, Di Buccio E, Ferro N, Masiero I, Peruzzo S, Silvello G (2012) Directions: Design and specification of an ir evaluation infrastructure. In: Springer (ed) Multilingual and Multimodal Information Access Evaluation - Third International Conference of the Cross-Language Evaluation Forum, LNCS, vol 7488, pp 88–99

Armstrong TG, Moffat A, Webber W, Zobel J (2009a) Evaluatir: an online tool for evaluating and comparing ir systems. In: SIGIR'09: Proceedings of the 32nd international ACM SIGIR conference, ACM, p 833, DOI http://doi.acm.org/10.1145/1571941.1572153

Armstrong TG, Moffat A, Webber W, Zobel J (2009b) Improvements that don't add up: ad-hoc retrieval results since 1998. In: CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management, ACM, pp 601–610, DOI http://doi.acm.org/10.1145/1645953.1646031

Blanco R, Zaragoza H (2011) Beware of relatively large but meaningless improvements. Tech. rep., Yahoo Research

Borlund P, Ingwersen P (1997) The development of a method for the evaluation of interactive information retrieval systems. JD 53:225–250

Braschler M, Peters C (2002) The CLEF campaigns: Evaluation of cross–language information retrieval systems. CEPIS UPGRADE III 3:78–81

Cleverdon C, Mills J, Keen M (1966) Factors determining the performance of indexing systems. Tech. rep., ASLIB Cranfield Research Project, Cranfield

Cleverdon CW (1962) Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Tech. rep., Aslib Cranfield Research Project, Cranfield, USA

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A large–scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp 248–255

Di Nunzio GM, Ferro N (2005) Direct: a system for evaluating information access components of digital libraries. In: International Conference on Theory and Practice of Digital Libraries, Springer, pp 483–484

Forsyth DA (2002) Benchmarks for storage and retrieval in multimedia databases. In: Storage and Retrieval for Media Databases, San Jose, California, USA, SPIEProc, vol 4676, pp 240–247, (SPIE Photonics West Conference)

Fraser AG, Dunstan FD (2010) On the impossibility of being expert. BMJ 341

Gollub T, Stein B, Burrows S (2012) Ousting ivory tower research: towards a web framework for providing experiments as a service. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 1125–1126

Gonzalo J, Clough P, Vallin A (2005) Overview of the CLEF 2005 interactive track. In: Working Notes of the 2005 CLEF Workshop, Vienna, Austria

Hanbury A, Müller H (2010) Automated component–level evaluation: Present and future. In: International Conference of the Cross–Language Evaluation Forum (CLEF), Springer, Lecture Notes in Computer Science (LNCS), vol 6360, pp 124–135

Hanbury A, Müller H, Langs G, Weber MA, Menze BH, Fernandez TS (2012) Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In: CLEF conference, Springer Lecture Notes in Computer Science

Hanbury A, Müller H, Balog K, Brodt T, Cormack GV, Eggel I, Gollub T, Hopfgartner F, Kalpathy-Cramer J, Kando N, Krithara A, Lin J, Mercer S, Potthast M (2015) Evaluation–as–a–service: Overview and outlook. ArXiv 1512.07454

Hanbury A, Müller H, Langs G (eds) (2017) Cloud–Based Benchmarking of Medical Image Analysis, vol 6. Springer International Publishing

Harman D (1992) Overview of the first Text REtrieval Conference (TREC–1). In: Proceedings of the first Text REtrieval Conference (TREC–1), Washington DC, USA, pp 1–20

Hopfgartner F, Kille B, Lommatzsch A, Plumbaum T, Brodt T, Heintz T (2014) Benchmarking news recommendations in a living lab. In: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, pp 250–267

Hopfgartner F, Hanbury A, Müller H, Kando N, Mercer S, Kalpathy-Cramer J, Potthast M, Gollub T, Krithara A, Lin J, Balog K, Eggel I (2015) Report on the evaluation–as–a–service (eaas) expert workshop. ACM SIGIR Forum 49(1):57–65

Hopfgartner F, Hanbury A, M´uller H, Balog K, Brodt T, Cormack GV, Eggel I, Gollub T, Kalpathy-Cramer J, Kando N, Krithara A, Lin J, Mercer S, Potthast M (2018) Evaluation-as-a-service in the computational sciences: Overview and outlook. Journal of Data and Information Quality

Ioannidis JP (2005) Why most published research findings are false. PLoS medicine 2(8):e124

Jimenez-del-Toro O, Müller H, Krenn M, Gruenberg K, Taha AA, Winterstein M, Eggel I, Foncubierta-Rodríguez A, Goksel O, Jakab A, Kontokotsios G, Langs G, Menze B, Salas Fernandez T, Schaer R, Walleyo A, Weber MA, Dicente Cid Y, Gass T, Heinrich M, Jia F, Kahl F, Kechichian R, Mai D, Spanier AB, Vincent G, Wang C, Wyeth D, Hanbury A (2016) Cloud–based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL Anatomy Benchmarks. IEEE Transactions on Medical Imaging 35(11):2459–2475

Jones KS, van Rijsbergen C (1975) Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge

Kalpathy-Cramer J, García Seco de Herrera A, Demner-Fushman D, Antani S, Bedrick S, Müller H (2015) Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics 39(0):55 – 61

Krenn M, Dorfer M, Jimenez-del-Toro O, Müller H, Menze B, Weber MA, Hanbury A, Langs G (2016) Creating a Large–Scale Silver Corpus from Multiple Algorithmic Segmentations. In: Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, Cai W, Metaxas D (eds) Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 9, 2015, Revised Selected Papers, Springer International Publishing, pp 103–115

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp 1097–1105

Langs G, Hanbury A, Menze B, Müller H (2012) VISCERAL: Towards large data in medical imaging — challenges and directions. In: Greenspan H, Müller H, Syeda-Mahmood T (eds) Medical Content-Based Retrieval for Clinical Decision Support, Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, pp 92–98

Markonis D, Holzer M, Dungs S, Vargas A, Langs G, Kriewel S, Müller H (2012) A survey on visual information search behavior and requirements of radiologists. Methods of Information in Medicine 51(6):539–548

Markonis D, Baroz F, Ruiz de Castaneda RL, Boyer C, Müller H (2013) User tests for assessing a medical image retrieval system: A pilot study. In: MEDINFO 2013

Mayernik MS, Hart DL, Mauli DL, Weber NM (2016) Assessing and tracing the outcomes and impact of research infrastructures. Journal of the American Society for Information Science and Technology

Müller H, Müller W, Marchand-Maillet S, Squire DM, Pun T (2001) Automated benchmarking in content–based image retrieval. In: Proceedings of the second International Conference on Multimedia and Exposition (ICME'2001), IEEE Computer Society, IEEE Computer Society, Tokyo, Japan, pp 321–324

Müller H, Marchand-Maillet S, Pun T (2002) The truth about Corel – Evaluation in image retrieval. In: Lew MS, Sebe N, Eakins JP (eds) Proceedings of the International Conference on the Challenge of Image and Video Retrieval (CIVR 2002), Springer–Verlag, London, England, Lecture Notes in Computer Science (LNCS), vol 2383, pp 38–49

Müller H, Boyer C, Gaudinat A, Hersh W, Geissbuhler A (2007) Analyzing web log files of the Health On the Net HONmedia search engine to define typical image search tasks for image retrieval evaluation. In: MedInfo 2007, Brisbane, Australia, IOS press, Studies in Health Technology and Informatics, vol 12, pp 1319–1323

Müller H, Clough P, Deselaers T, Caputo B (eds) (2010) ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol 32. Springer, Berlin Heidelberg

Müller H, Kalpathy-Cramer J, Hanbury A, Farahani K, Sergeev R, Paik JH, Klein A, Criminisi A, Trister A, Norman T, Kennedy D, Srinivasa G, Mamonov A, Preuss N (2016) Report on the cloud–based evaluation approaches workshop 2015. ACM SIGIR Forum 51(1):35–41

Niemeyer KE, Smith AM, Katz DS (2016) The challenge and promise of software citation for credit, identification, discovery, and reuse. Journal Data and Information Quality 7(6):161–165

Ounis I, Macdonald C, Lin J, Soboroff I (2011) Overview of the trec-2011 microblog track. In: Proceeddings of the 20th Text REtrieval Conference (TREC 2011), vol 32

Rowe BR, Wood DW, Link AN, Simoni DA (2010) Economic impact assessment of NIST text retrieval conference (TREC) program. Technical report project number 0211875, National Institute of Standards and Technology

Salton G (1971) The SMART Retrieval System, Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs, New Jersey, USA

Silvello G (2018) Theory and practice of data citation. Journal of the Association for Information Science and Technology 69:6–20

Silvello G, Bordea G, Ferro N, Buitelaar P, Bogers T (2017) Semantic representation and enrichment of information retrieval experimental data. International Journal on Digital Libraries 18(2):145–172

Smeaton AF, Kraaij W, Over P (2003) TRECVID 2003: An overview. In: Proceedings of the TRECVID 2003 conference

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 62(4):613–627

Trister AD, Buist DS, Lee CI (2017) Will machine learning tip the balance in breast cancer screening? JAMA oncology 3(11):1463–1464

Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, et al (2015) An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics 16(1):138

Tsikrika T, García Seco de Herrera A, Müller H (2011) Assessing the scholarly impact of ImageCLEF. In: CLEF 2011, Springer Lecture Notes in Computer Science (LNCS), pp 95–106

Tsikrika T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The scholarly impact of CLEF (2000–2009). In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer, pp 1–12