Chapter 1
VISCERAL: Evaluation-as-a-Service for Medical Imaging

Allan Hanbury and Henning Müller

Abstract Systematic evaluation has had a strong impact in many data analysis domains, for example TREC and CLEF in information retrieval, ImageCLEF in image retrieval and many challenges in conferences such as MICCAI for medical imaging and ICPR for pattern recognition. With Kaggle, a platform for machine learning challenges has also had a significant success in crowdsourcing solutions. This shows the importance to systematically evaluate algorithms and that the impact is far larger than simply evaluating a single system. Many of these challenges also showed the limits of the commonly-used paradigm to prepare a data collection and tasks, distribute these and then evaluate the participants' submissions. Extremely large datasets are cumbersome to download, while shipping hard disks containing the data becomes impractical. Confidential data can often not be shared, for example medical data, but also data from company repositories. Real-time data will never be available via static data collections as the data changes over time and data preparation often takes much time. The Evaluation-as-a-Service (EaaS) paradigm tries to find solutions for many of these problems and has been applied in the VISCERAL project. In EaaS the data are not moved but remain on a central infrastructure. In the case of VISCERAL, all data were made available in a cloud environment. Participants were provided with virtual machines on which to install their algorithms. Only a small part of the data, the training data, were visible to participants. The major part of the data, the test data, were only accessible to the organizers who ran the algorithms in the participants' virtual machines on the test data to obtain impartial performance measures.

Allan Hanbury
TU Wien, Institute of Software Technology and Interactive Systems, Favoritenstraße 9-11/188, 1040 Vienna, Austria. e-mail: allan.hanbury@tuwien.ac.at

Henning Müller
Information Systems Institute, HES-SO Valais, Rue du Technopole 3, 3960 Sierre, Switzerland. e-mail: henning.mueller@hevs.ch

1.1 Introduction

Scientific progress can usually be measured via clear and systematic experiments (Lord Kelvin: "If you can not measure it, you can not improve it."). In the past, scientific benchmarks, such as TREC (Text Retrieval Conference) and CLEF (Conference and Labs of the Evaluation Forum), have given a platform for such scientific comparisons and have had a significant impact [15, 17, 18]. Commercial platforms such as Kaggle[1] have also shown that there is a market for a comparison of techniques based on real problems that companies can propose.

Much data are available and can potentially be exploited for generating new knowledge based on data, including notably medical imaging, where extremely large amounts have been produced for many years [1]. Still, constraints are often that data need to be manually anonymised or can only be used in restricted settings, which does not work well for very large data sets.

Several of the problems encountered in traditional benchmarking that often relies on the paradigm of creating a dataset and sending it to participants can be summarized in the following points:

- *very large* data sets can only be distributed with very much effort, usually by sending hard disks through the post;
- *confidential* data are extremely hard to distribute and they can usually only be used in a closed environment, in a hospital or inside the company firewalls;
- *quickly changing* data sets cannot be used for benchmarking if it is necessary to package the data and send them around.

To answer these problems and challenges, the VISCERAL project proposed a change in the way that benchmarking has been organized by proposing to keep the data in a central space and move the algorithms to the data [3, 10].

Other benchmarks equally realized these difficulties in running benchmarks and came up with a variety of propositions for running benchmarks without fixed data packages that are distributed. These ideas were discussed in a workshop organized around this topic and named Evaluation-as-a-Service (EaaS) [6]. Based on the discussions at the workshop, a detailed White Paper was written [4], which outlines the roles involved in this process and also the benefits that researchers, funding organizations and companies can gain from such a shift in scientific evaluations.

This chapter highlights the role of VISCERAL in the EaaS area, which benchmarks were organized and how the benchmarks helped advance this field and gain concrete experience with running scientific evaluations in the cloud.

[1] `http://www.kaggle.com`

1.2 VISCERAL Benchmarks

The VISCERAL project organised a series of medical imaging Benchmarks described below:

1.2.1 Anatomy Benchmarks

A set of medical imaging data in which organs are manually annotated is provided to the participants. The data contains segmentations of several different anatomical structures as well as positions of landmarks in different image modalities, e.g. CT and MRI. Participants in the Anatomy Benchmarks have the task of submitting software that automatically segments the organs for which manual segmentations are provided, or detecting the locations of the landmarks. After submission, this software is tested on images which are inaccessible to the participants. Three rounds of the Anatomy Benchmark have been organised, and this Benchmark is continuing beyond the end of the VISCERAL project. These benchmarks are described in more detail in Chapter 7. Chapters 9–12 are reports of some participants in the Anatomy Benchmarks.

1.2.2 Detection Benchmark

A set of medical imaging data that contains various lesions manually annotated in anatomical regions such as the bones, liver, brain, lung, or lymph nodes is distributed to participants. Participants in the Detection Benchmark have the task of submitting software that will automatically detect these lesions. The software is tested on detecting lesions on images that the participants have not seen. The Benchmark data and ground truth continue to be available beyond the end of the VISCERAL project as the Detection2 Benchmark. As this was the most challenging benchmark that was organised, no solutions were submitted. There is therefore no chapter on this benchmark included, although the data and ground truth continue to be available.

1.2.3 Retrieval Benchmark

One of the challenges of medical information retrieval is similar case retrieval in the medical domain based on multimodal data, where cases refer to data about specific patients (used in an anonymised form), such as medical records, radiology images and radiology reports or cases described in the literature or teaching files. The Retrieval Benchmark simulates the following scenario: a medical professional is assessing a query case in a clinical setting, e.g., a CT volume, and is searching for

cases that are relevant in this assessment. The participants in the Benchmark have the task of developing software that finds clinically-relevant (related or useful for differential diagnosis) cases given a query case (imaging data only or imaging and text data), but not necessarily the final diagnosis. The Benchmark data and relevance assessments continue to be available beyond the end of the VISCERAL project as the Retrieval2 Benchmark. This benchmark is described in more detail in Chapter 8. Chapters 13 and 14 give reports of two of the participants in the Retrieval Benchmark.

1.3 Evaluation-as-a-Service in VISCERAL

Evaluation-as-a-Service is an approach to the evaluation of data science algorithms, in which the data remains centrally stored, and participants are given access to this data in some controlled way.

The access to the data can be provided through various mechanisms, including an API to access the data or virtual machines on which to install and run the processing algorithms. Mechanisms to protect sensitive data can also be implemented, such as running the virtual machines in sandboxed mode (all access out of the virtual machine is blocked) while the sensitive data are being processed, and destroying the virtual machine after extracting the results to ensure that no sensitive data remains in a virtual machine [13]. An overview of the use of Evaluation-as-a-Service is given in [4] and [6].

We now give two examples of Evaluation-as-a-Service in use, illustrating the different types of data for which EaaS is useful. In the TREC Microblog task [11], search on Twitter was evaluated. As it is not permitted to redistribute Tweets, an API (Application Programming Interface) was created allowing access to the Tweets stored centrally. In the CLEF NewsREEL task [5], news recommender systems were evaluated. In this case, an online news recommender service sent requests for recommendations in real-time based on actual requests from users, and the results were evaluated based on the clicks of the recommendations by the users of the online recommender service. As this was real-time data from actual users of a system, a platform, the Open Recommendation Platform [2], was developed to facilitate communication between the news recommender portal and the task participants.

In the VISCERAL project, we were dealing with sensitive medical data. Even though the data had been anonymised by removing potentially personal meta-data and blurring the facial regions of the images, it was not possible to guarantee that the anomymisation tools had completely anonymised the images. We were therefore required to keep a large proportion of images, the test set, inaccessible to participants. Training images were available to participants as they had undergone a more thorough control of the anonymisation effectiveness. The EaaS approach allowed this to be done in a straightforward way.

The training and test data are stored on the cloud in two separate storage containers. When each participant registers, he/she is provided with a virtual machine on

the cloud that has access to the training data container, as illustrated in Figure 1.1. During the *Training Phase*, the participant should install the software that carries out the benchmark task on the virtual machine, following the specifications provided, and can train algorithms and experiment using the training data as necessary. Once the participant is satisfied with the performance of the installed software, the virtual machine is submitted to the organisers. Once a virtual machine is submitted, the participant loses access to it, and the *Test Phase* begins. The organisers link the submitted virtual machine to the test data, as shown in Figure 1.2, run the submitted software on the test data, and calculate metrics showing how well the submitted software performs.

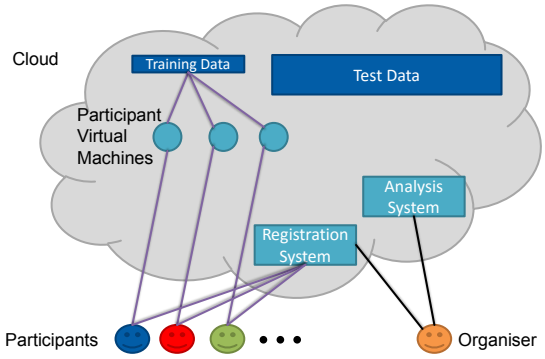

Fig. 1.1 Training Phase. The participants register and each get their own virtual machine in the cloud, linked to a training dataset of the same structure as the test data. Software for carrying out the competition objectives is placed in the virtual machines by the participants. The test data is kept inaccessible to participants.

For the initial VISCERAL benchmarks, the organisers set a deadline by which all virtual machines must be submitted. The values of the performance metrics were then sent to participants by e-mail. This meant that a participant had only a single possibility to get results of their computation on the test data. For the final round of the Anatomy Benchmark (Anatomy3), a continuous evaluation approach was adopted. Participants have the possibility to submit their virtual machine multiple times for assessment of the software on the test set (there is a limit on how often this can be done to avoid "training on the test set"). The evaluation on the test set is carried out automatically, and participants can view the results on their personal results page. Participants can also choose to make results public on the global leaderboard.

Chapter 2 presents a detailed description of the VISCERAL cloud environment.

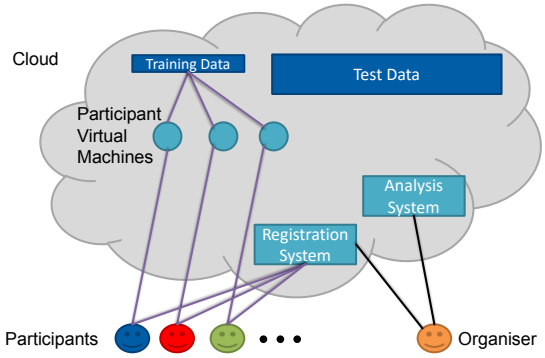

Fig. 1.2 Test Phase. On the Benchmark deadline, the organiser takes over the virtual machines containing the software written by the participants, links them to the test dataset, performs the calculations, and evaluates the results.

1.4 Main Outcomes of VISCERAL

As a result of running the Benchmarks, the VISCERAL project generated data and software that will continue to be useful to the medical imaging community. The first major data outcome are manually annotated MR and CT images, which we refer to as the *Gold Corpus*. The use of the EaaS paradigm also gave the possibility to compute a *Silver Corpus* by fusing the results of the participant submissions. One of the challenges in creating datasets for use in medical imaging benchmarks is obtaining permission to use the image data for this purpose. In order to provide guidelines for researchers intending to obtain such permission, we present an overview of the processes necessary at the three institutes that provided data for the VISCERAL Benchmarks in Chapter 3. All data created during the VISCERAL project are described in detail in Chapter 5. Finally, particular attention was paid to ensuring that the metrics comparing segmentations were correctly calculated, leading to the release of new open source software for efficient metric calculation.

1.4.1 Gold Corpus

The VISCERAL project produced a large corpus of manually annotated radiology images, called the Gold Corpus. An innovative manual annotation coordination system was created, based on the idea of tickets, to ensure that the manual annotation was carried out as efficiently as possible. The Gold Corpus was subjected to an extensive quality control process, and is therefore small but of high quality. Annotation in VISCERAL served as the basis for all three Benchmarks. For each Benchmark,

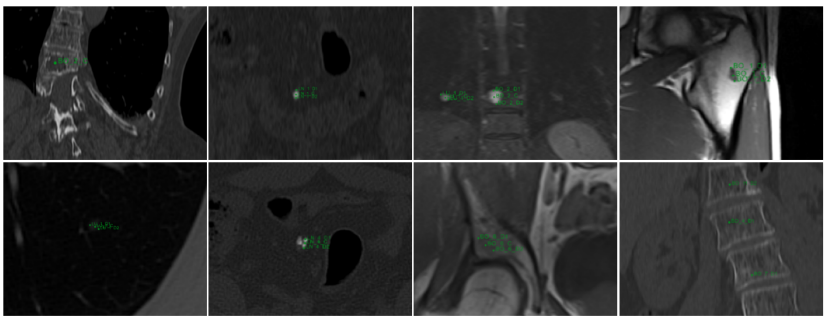

Fig. 1.3 Examples of lesion annotations.

training data was distributed to the participants and testing data was kept for the evaluation.

For the Anatomy Benchmark series [8], volumes from 120 patients were manually segmented by the end of VISCERAL by radiologists, where the radiologists trace out the extent of each organ. The following organs were manually segmented: left/right kidney, spleen, liver, left/right lung, urinary bladder, rectus abdominis muscle, 1st lumbar vertebra, pancreas, left/right psoas major muscle, gallbladder, sternum, aorta, trachea, and left/right adrenal gland. The radiologists also manually marked landmarks in the volumes, where the landmarks include: lateral end of clavicula, crista iliaca, symphysis below, trochanter major, trochanter minor, tip of aortic arch, trachea bifurcation, aortic bifurcation, and crista iliaca.

For the Detection Benchmark, overall 1,609 lesions were manually annotated in 100 volumes of two different modalities, in five different anatomical regions selected by radiologists: brain, lung, liver, bones, and lymph nodes. Examples of the manual annotation of lesions are shown in Figure 1.3.

For the Retrieval Benchmark [7], more than 10,000 medical image volumes were collected, from which about 2,000 were selected for the Benchmark. In addition, terms describing pathologies and anatomical regions were extracted from the corresponding radiology reports.

Detailed descriptions of the methods used in creating the Gold Corpus are described in Chapter 4.

1.4.2 Silver Corpus

In addition to the Gold Corpus of expert annotated imaging data described in the previous section, the use of the EaaS approach offered the possibility to generate a far larger Silver Corpus, which is annotated by the collective ensemble of participant algorithms. In other words, the Silver Corpus is created by fusing the outputs of all participant algorithms for each image (inspired by e.g. [14]). Even though

this Silver Corpus annotation is less accurate than expert annotations, the fusion of participant algorithm results is more accurate than individual algorithms and offers a basis for large-scale learning. It was shown by experiments that the accuracy of a Silver Corpus annotation obtained by label fusion of participant algorithms is higher than the accuracy of individual participant annotations. Furthermore, this accuracy can be improved by injecting multi-atlas label fusion estimates of annotations based on the Gold Corpus annotated dataset.

In effect, the Silver Corpus is large and diverse, but not of the same annotation quality as the Gold Corpus. The final Silver Corpus of VISCERAL Anatomy Benchmarks contains 264 volumes of four modalities (CT, CTce, MRT1, MRT1cefs), containing 4193 organ segmentations and 9516 landmark annotations. Techniques for the creation of the Silver Corpus are described in [9].

1.4.3 Evaluation Metric Calculation Software

In order to evaluate the segmentations generated by the participants, it is necessary to compare them objectively to the manually created ground truth. There are many ways in which the similarity between two segmentations can be measured, and at least 22 metrics have each been used in more than one paper in the medical segmentation literature. We implemented these 22 metrics in the EvaluateSegmentation software[16], which is available as open source on GitHub,[2] and can read all image formats (2D and 3D) supported by the ITK Toolkit. The software is specifically optimised to be efficient and scalable, and hence can be used to compare segmentations on full body volumes. Chapter 6 goes beyond [16] by discussing the extension to fuzzy metrics and how well rankings based on similarity to the ground truth of organ segmentations by various metrics correlate with rankings of these segmentations by human experts.

1.5 Experience with EaaS in VISCERAL

Based on the examples given there are several experiences to be gained from EaaS in general and VISCERAL more particularly. Some of the experiences, particularly in the medical domain are also discussed in [12].

Initially, the idea to run an evaluation in the cloud was seen by the medical imaging community with some skepticism. Several persons mentioned that they would not participate if they can not see the data and there definitely was a feeling of control loss. It is definitely work to install a virtual machine from scratch in the cloud. Furthermore, VISCERAL provided only a limited set of operating systems under Linux and Windows. There were also concrete questions regarding hardware such

[2] `https://github.com/Visceral-Project/EvaluateSegmentation`

as GPU (Graphical Processing Units) that are widely used for deep learning but that were not available in Azure at the time and prevented a potential participant from participating. These techniques are now easily available, so such problems are often removed quickly with the fast pace in the development of cloud infrastructures. Several participants who did not participate mentioned that they did so because it was additional work to set up the software in the cloud.

Other challenges were regarding the feedback if the algorithm completely failed for a specific image or when the script crashed. We had a few such cases and provided assistance to participants to remove the errors, but this is obviously only possible if the number of participants is relatively small.

In this respect the system also created more work for the organizers than simply making data available for download and receiving calculated results from participants. Once infrastructures that are easier to use and a skeleton for evaluations are available this will also reduce the additional work. The CodaLab[3] software is one such system that makes running a challenge in the cloud much easier and a deeper integration between cloud and executed algorithms could help even further.

On the positive side are several important aspects. First, the three problems mentioned above regarding very large datasets, confidential data and quickly changing data are solved with the given approach. It is also important that all participants take part under the same conditions, so that there is no advantage with a fast Internet connection where data download takes minutes and not days. All participants also had the same environment, hence the same computing power, and there was no difference between computing resources available to participants, also removing a bias. The fact that all participating groups were compared based on the same infrastructure also allowed to compare run time and thus efficiency of algorithms, which is impossible to compare otherwise. In terms of reproducibility the system is extremely good as no one can optimize techniques based on the test data.

The fact that the executables of all participants were available also allowed the creation of the Silver Corpus on new, non-annotated data, done by running all submitted algorithms on the new data and then performing a label fusion. This has shown to deliver much better results than even the best submitted algorithm. Availability of executables can also be used to run the code on new data that has become available or modified data when errors were detected, something that did happen in VISCERAL.

The cloud-based evaluation workshop [12] also showed that there are several ongoing developments that will make the creation of such challenges and use of code much easier. Docker is for example much lighter than virtual machines and submitting Docker containers can be both faster and reduce the amount of work necessary to create the container for participants. Code sharing among participants might also be supported in a more straightforward way, so participants can combine components of other research groups with their own components to optimize results systematically.

[3] `https://github.com/codalab/`

1.6 Conclusion

The VISCERAL project made a number of useful contributions to the medical imaging field, but also to the organisation of data science evaluations in general through advancing the Evaluation-as-a-Service approach. The techniques developed and lessons learned will be useful for evaluation in machine learning, information retrieval, data mining and related areas, allowing the evaluation tasks to be done on huge, non-distributable, private or real-time data. This should not only allow the evaluation tasks to become more realistic and closer to practice, but should also increase the level of reproducibility of the experimental results.

In the area of medical imaging, the VISCERAL project contributed large datasets of annotated CT and MR images. The annotations have been done by qualified radiologists in the creation of the Gold Corpus, but a form of crowdsourcing based on participant submissions allowed the much larger Silver Corpus to be built. Furthermore, a thorough analysis of metrics used in the evaluation of image segmentation was contributed, along with an efficient and scalable implementation of the calculation of these metrics.

1.7 Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 318068 (VISCERAL).

References

[1] (2010) Riding the wave: How Europe can gain from the rising tide of scientific data. Submission to the European Commission, available online at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf, URL http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[2] Brodt T, Hopfgartner F (2014) Shedding Light on a Living Lab: The CLEF NEWSREEL Open Recommendation Platform. In: IIiX'14: Proceedings of Information Interaction in Context Conference, ACM, pp 223–226, URL http://dx.doi.org/10.1145/2637002.2637028

[3] 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

[4] 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. CoRR abs/1512.07454, URL http://arxiv.org/abs/1512.07454

[5] Hopfgartner F, Kille B, Lommatzsch A, Brodt T, Heintz T (2014) Benchmarking News Recommendations in a Living Lab. In: CLEF'14: Proceedings of the 5th International Conference of the CLEF Initiative, Springer Verlag, pp 250–267

[6] 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. SIGIR Forum 49(1):57–65

[7] Jimenez-del-Toro O, Hanbury A, Langs G, Foncubierta-Rodríguez A, Müller H (2015) Overview of the VISCERAL Retrieval Benchmark 2015. In: Multimodal Retrieval in the Medical Domain (MRMD) 2015, Springer, Lecture Notes in Computer Science, vol 9059

[8] 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. Medical Imaging, IEEE Transactions on

[9] Krenn M, Dorfer M, Jiménez-del Toro OA, Müller H, Menze B, Weber MA, Hanbury A, Langs G (2014) Creating a Large-Scale Silver Corpus from Multiple Algorithmic Segmentations. In: MICCAI Workshop on Medical Computer Vision: Algorithms for Big Data, MCV 2015, Springer, vol 8848, pp 163–170

[10] 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

[11] Lin J, Efron M (2013) Overview of the TREC-2013 Microblog Track. In: TREC'13: Proceedings of the 22nd Text REtrieval Conference, Gaithersburg, Maryland

[12] Müller, 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

[13] Potthast M, Gollub T, Rangel F, Rosso P, Stamatatos E, Stein B (2014) Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: CLEF'14: Proceedings of the 5th Int. Conference of the CLEF Initiative, Springer Verlag, pp 268–299

[14] Rebholz-Schuhmann D, Jimeno Yepes AJ, Van Mulligen EM, Kang N, Kors J, Milward D, Corbett P, Buyko E, Beisswanger E, Hahn U (2010) CALBC silver standard corpus. Journal of Bioinformatics and Computational Biology 8(1):163–179

[15] 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
[16] Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging 15(1):1–28
[17] 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
[18] 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