# Consolidating the ImageCLEF Medical Task Test Collection: 2005-2007*

William Hersh, MD[1], Henning Müller, PhD[2],
Jayashree Kalpathy-Cramer, PhD[1], Eugene Kim[1]

[1]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University

[2]Section of Medical Informatics, University & Hospitals of Geneva, Geneva, Switzerland

3181 SW Sam Jackson Park Rd., BICC

Portland, OR 97239

+1-503-494-4563

e-mail: hersh@ohsu.edu

**Abstract**

The goal of the ImageCLEF medical image retrieval task (ImageCLEFmed) has been to improve understanding and system capability in search for medical images. This has been done by developing a test collection that allows system-oriented evaluation of medical image retrieval systems. From 2005-2007, test collections were developed and used for ImageCLEFmed. This paper describes our recent work consolidating the test collections into a single unified collection of 66,662 images and their annotations; 85 topics classified by amenability to visual, textual, or mixed retrieval methods; and relevance judgments. This will provide a comprehensive test collection for further testing of systems and algorithms in medical image retrieval..

## 1 Introduction

Images play a variety of uses in health care and biomedical research. Despite their widespread use, however, we know little about how those who use them find and manage them. Two small analyses have found that the image use tends to be related to the "role" of the user, such as clinician, educator, researcher, etc. [1, 2]. As there are growing numbers of image collections and search interfaces proliferating on the World Wide Web as well as closed

* This work was supported by a supplement to National Science Foundation (NSF) grant ITR-0325160. We also acknowledge the European Commission IST projects program in facilitating this work (through the SemanticMining Network of Excellence, grant 507505) and the Swiss National Funds (grant 205321-109304/1). Instructions for obtaining the data described in this paper can be obtained from the ImageCLEFmed Web site (http://ir.ohsu.edu/image/).

networks, we believe it is important to understand users' needs as well as provide systems that meet those needs.

The goal of the ImageCLEF medical image retrieval task (ImageCLEFmed) is to improve understanding and system capability in search for medical images [3]. This has been done by developing a test collection that allows system-oriented evaluation of medical image retrieval systems. As with most collections, we have strived to make the content and search topics for this collection as realistic as possible. For three years running, ImageCLEF has featured a medical retrieval task based around ad hoc retrieval. The collection of images came from four sources initially, with two additional ones added in the third year. Each collection is used "as is," i.e., its annotations are used from the original source. This paper describes the recent effort by the project to consolidate the three years of test collections into a single collection that aims to provide a test bed for evaluating systems and algorithms that perform medical image retrieval.

## 2 Background

ImageCLEF is a part of the Cross-Language Evaluation Forum (CLEF, www.clef-campaign.org), a challenge evaluation for information retrieval from diverse languages [4]. CLEF itself is an outgrowth of the Text Retrieval Conference (TREC, trec.nist.gov), a forum for evaluation of text retrieval systems [5]. TREC and CLEF operate on an annual cycle of test collection development and distribution, followed by a conference where results are presented and analyzed.

The goals of TREC and CLEF are to build realistic test collections that simulate real-world retrieval tasks and enable researchers to assess and compare system performance [6]. The goal of test collection construction is to assemble a large collection of *content* (documents, images, etc.) that resemble collections used in the real world. Builders of test collections also seek a sample of realistic *tasks* to serve as *topics* that can be submitted to systems as *queries* to retrieve content. The final component of test collections is *relevance judgments* that determine which content is relevant to each topic. A major challenge for test collections is to develop a set of realistic topics that can be judged for relevance to the retrieved items. Such benchmarks are needed by any researcher or developer in order to evaluate the effectiveness of new tools.

Test collections usually measure how well systems or algorithms retrieve relevant items. The most commonly used evaluation measures are recall and precision. *Recall* is the proportion of relevant documents retrieved from the database whereas *precision* is the proportion of relevant documents retrieved in the search. Often there is a desire to combine recall and precision into a single aggregate measure. Although many approaches have been used for aggregate measures, the most frequently used one in TREC and CLEF has been the mean average precision (MAP) [7]. In this measure, which can only be used with ranked output from a search engine, precision is calculated at every point at which a relevant document is obtained. The average precision for a topic is then calculated by averaging the precision at each of these points. MAP is then calculated by taking the mean of the average precision

values across all topics in the run. MAP has been found to be a stable measure for combining recall and precision, but suffers from its value arising from being a statistical aggregation and having no real-world meaning [8].

Test collections have been used extensively to evaluate IR systems in biomedicine. A number of test collections have been developed for document retrieval in the clinical domain [9, 10]. More recently, focus has shifted to the biomedical research domain in the TREC Genomics Track [11]. Test collections are also used increasingly for image retrieval outside of medicine [12].

In this paper, we describe our efforts to create a single consolidated test collection. In the remaining sections, we describe the content, topics, relevance judgments, and future plans for the merged collection.

# 3 Content

The conceptual structure of the content of the ImageCLEFmed test collection is as follows. The entire *library* consists of multiple collections. Each *collection* is organized into cases that represent a group of related images and annotations. Each *case* consists of a group of images and an optional annotation. Each *image* is part of a case and has optional associated annotations, which consist of metadata (e.g., HEAL tagging), and/or a textual annotation. All of the images and annotations are stored in separate files. An XML file contains the connections between the collections, cases, images, and annotations. Figure 1 shows a graphical depiction of the library, while Figure 2 shows the XML metadata format.

The image library for ImageCLEFmed 2005 and 2006 consisted of the first four collections listed in Tables 1 and 2 (Casimage, MIR, PEIR, and PathoPIC). In 2007, we added the latter two collections listed in those tables (myPACS and CORI). Table 1 describes the image collections, their image and annotation types, and their origins, while Table 2 lists the numbers of images and annotations (including amounts in each language) as well as the archived file size. Figure 3 shows an example case from the Casimage collection, demonstrating how multiple different images and image types can be part of a case. However, note that the largest collection, PEIR, is not organized into cases per se (or, using our framework, has one image per case). The image library for the consolidated test collection will be the entire library, which is the same as that used for ImageCLEFmed 2007.

# 4 Topics

A total of 85 topics have been developed over 2005-2007 for ImageCLEFmed. Each topic has been provided with an information statement in English, French, and German, as well as an index image of a relevant image for use by visual retrieval systems. Because we discovered early on that results on different tasks varied by whether the topic was amenable to visual or textual retrieval, we classified each topic as visual, textual, or mixed.

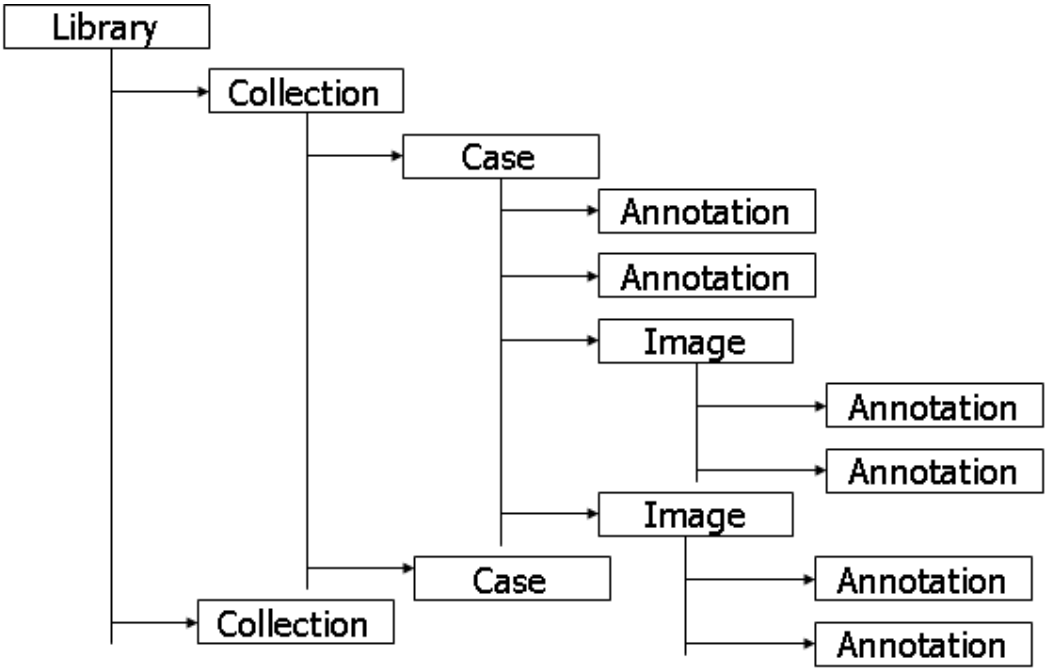


Figure 1 - Structure of ImageCLEF medical image retrieval task (ImageCLEFmed) test collection content.

```
<library>
    <collection>
        <name>name-text</name>
        <cases>
            <case>
                <id>identifier-text</id>
                <images>
                    <image>
                        <id>identifier-text</id>
                        <imagefile>file-name-text</imagefile>
                        <annotation lang=" ">file-name-text</annotation>
                        <annotation lang=" ">file-name-text</annotation>
                    </image>
                <images>
                <annotation lang=" ">file-name-text</annotation>
                <annotation lang=" ">file-name-text</annotation>
            </case>
        </cases>
    </collection>
</library>
```

Figure 2 - Structure of ImageCLEF medical image retrieval task (ImageCLEFmed) XML metadata format for the content.

There were 25 topics in 2005 and 30 each in 2006 and 2007. In each year, each topic was numbered from 1, i.e., 1-25 in 2005 and 1-30 in 2006 and 2007. In the consolidated test collection, the topics from 2005 are numbered 1-25, those from 2006 are numbered 26-55, and those from 2007 are numbered 56-85. A sample topic from the consolidated collection is shown in Figure 4.

Table 1 - ImageCLEF medical image retrieval task (ImageCLEFmed) image collections, image and annotation types, and their origins.

| **Collection Name** | **Image Type(s)** | **Annotation Type(s)** | **Original URL** |
|---|---|---|---|
| Casimage | Radiology and pathology | Clinical case descriptions | http://www.casimage.com/ |
| Mallinckrodt Institute of Radiology (MIR) | Nuclear medicine | Clinical case descriptions | http://gamma.wustl.edu/home.html |
| Pathology Education Instructional Resource (PEIR) | Pathology and radiology | Metadata records from HEAL database | http://peir.path.uab.edu/ |
| PathoPIC | Pathology | Image description - long in German, short in English | http://alf3.urz.unibas.ch/pathopic/e/intro.htm |
| MyPACS | Radiology | Clinical case descriptions | http://www.mypacs.net/ |
| Clinical Outcomes Research Initiative (CORI) Endoscopic Images | Endoscopy | Clinical case descriptions | http://www.cori.org/ |

Table 2 - ImageCLEF medical image retrieval task (ImageCLEFmed) numbers of images and annotations (including amounts in each language) as well as the archived file size

| Collection Name | Cases | Images | Annotations | Annotations by Language | File Size (tar archive) |
|---|---|---|---|---|---|
| Casimage | 2076 | 8725 | 2076 | French - 1899<br>English - 177 | 1.28 GB |
| MIR | 407 | 1177 | 407 | English - 407 | 63.2 MB |
| PEIR | 32319 | 32319 | 32319 | English - 32319 | 2.50 GB |
| PathoPIC | 7805 | 7805 | 15610 | German - 7805<br>English - 7805 | 879 MB |
| myPACS | 3577 | 15140 | 3577 | English - 3577 | 390 MB |
| Endoscopic | 1496 | 1496 | 1496 | English - 1496 | 34 MB |
| Total | 47680 | 66662 | 55485 | French - 1899<br>English - 45781<br>German - 7805 | 5.15 GB |

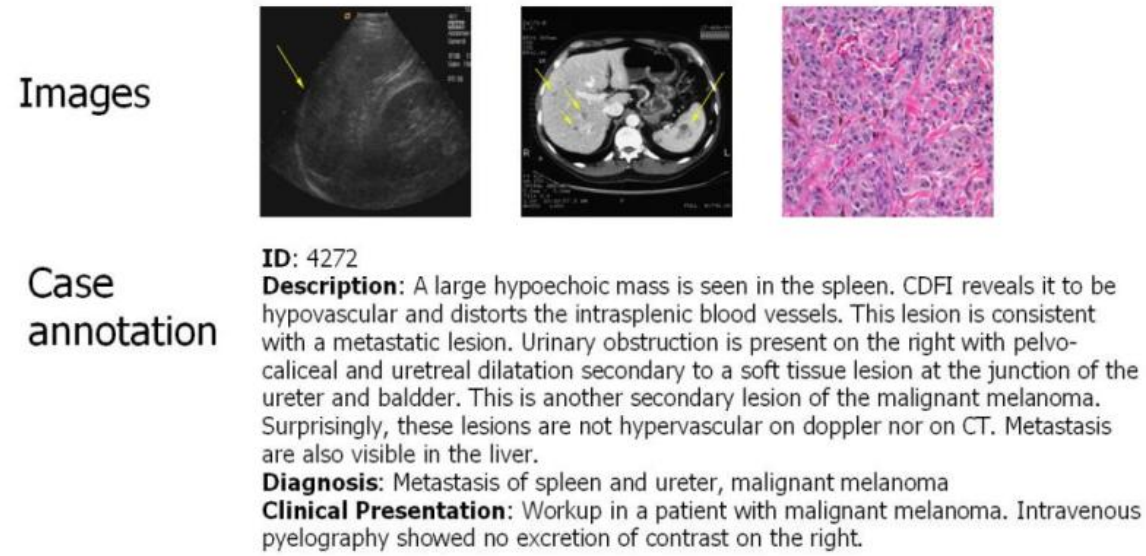


Figure 3 - An example ImageCLEF medical image retrieval task (ImageCLEFmed) case from the Casimage collection.

```
<topic>
  <number>55</number>
  <EN-description>Show me images of findings with Alzheimer's Disease.
      </EN-description>
  <DE-description>Zeige mir Bilder von Fällen mit einer Alzheimer Diagnose.
      </DE-description>
  <FR-description>Montre-moi des images d'observations avec la maladie
      d'Alzeimer.</FR-description>
  <year>2006</year>
  <query-images>
    <image>images2006/3-10a.jpg</image>
    <image>images2006/3-10b.jpg</image>
  </query-images>
  <query-type>semantic</query-type>
</topic>
```

Figure 4 - Topic 55 from the consolidated ImageCLEF medical image retrieval task (ImageCLEFmed) test collection.

# 5 Relevance Judgments

Relevance judgments in ImageCLEFmed have been performed by physicians who are also students in the OHSU biomedical informatics graduate program. They have been paid an hourly rate for their work. The pools for relevance judging have been created by selecting the top ranking images from all submitted runs. The actual number selected from each run varied by year, but was usually about 30-40, with the goal of having pools of about 800-1200 images in size for judging. Judges have been instructed to rate images in the pools are definitely relevant (DR), partially relevant (PR), or not relevant (NR). In ImageCLEFmed 2005 we used only DR images for the gold standard, but in 2006 and 2007 we used DR and PR images.

For the consolidated test collection, we need to perform relevance judgments for the new 2007 images applied to the 2005 and 2006 topics. This process is currently underway. We are also judging some images whose names were erroneous in the 2007 pools due to their being incorrect in the submitted runs. Relevance judging will take place in August-September, 2007.

Although the reliability of judging has been slightly better than that obtained from relevance judgments of textual documents in clinical [9] and genomics [11, 13] tasks, we have found instances of incorrectly judged images, especially with regards to the modality of the image, which is vitally important in image retrieval. To that end, we plan to nominate images for rejudgment, and all future judges will be asked to adhere to the following instructions:

1. Note that a topic can refer to one or more of the following: (a) an imaging modality, (b) an anatomical location, (c) a view and/or (d) a disease or finding. An image should only be considered relevant if it meets all the terms mentioned explicitly in the topic (i.e., should be an AND, not an OR). For instance, in the topic “CT liver abscess,” only CT scans showing a liver abscess should be considered relevant. Pathology or

MRI images of liver abscesses should not be considered relevant. Images of other abscesses should not be considered relevant. An x-ray image associated with an annotation that refers to a need for a CT scan in the future should not be considered relevant.

2. When a photograph is the desired imaging modality, i.e., it says "image of" or picture of," only photographic images should be considered relevant. Although, technically, microscopic images of histology/pathology may be considered to be photographs, in this context, they should not be considered relevant.
3. Pathology in the query refers to pathological images (microscopic/gross pathology), not the state of being abnormal.
4. Refer to the sample images provided with each topic for a better understanding of desired imaging modalities.
5. Synonyms of terms should be considered relevant in the topic. For instance, any MeSH synonyms of the search terms should be considered relevant. As an example, cholangiocarcinoma is a synonym of bile duct cancer. But on the other hand, the liver/biliary system/pancreas should not be considered synonymous with the entire gastrointestinal system.

## 6 Future Work

When the work described in sections 2-4 is complete, we will have a medical image retrieval collection with 66,662 images and their annotations; 85 topics categorized by amenability to visual, textual, or mixed retrieval; and about 800-1200 relevance judgments per topic. Our goal is to contact the dozen or so major participants in ImageCLEFmed from 2005-2007 submit baseline runs so that baseline levels of performance can be ascertained. Our hope is that additional researchers will use the collection to evaluate new approaches to image retrieval in the future.

We do plan to continue ImageCLEFmed in 2008 but hope to look at new types of tasks beyond the ad hoc retrieval used in 2005-2007. We aim to expand our previous work in user assessment to develop use cases for new tasks.

## References


1. Hersh WR, et al. *A qualitative task analysis of biomedical image use and retrieval*. *MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*. 2005. Vienna, Austria. http://muscle.prip.tuwien.ac.at/workshop2005_proceedings/hersh.pdf.
2. Müller H, et al. *Health care professionals' image use and search behaviour*. *Proceedings of Medical Informatics Europe 2006*. 2006. Maastricht, Netherlands. 24-32. http://www.sim.hcuge.ch/medgift/publications/MIE2006_Mueller.pdf.

3. Hersh WR, et al., *Advancing biomedical image retrieval: development and analysis of a test collection.* Journal of the American Medical Informatics Association, 2006. 13: 488-496.
4. Braschler M and Peters C, *Cross-language evaluation forum: objectives, results, achievements.* Information Retrieval, 2004. 7: 7-31.
5. Voorhees EM and Harman DK, eds. *TREC: Experiment and Evaluation in Information Retrieval*. 2005, MIT Press: Cambridge, MA.
6. Sparck-Jones K, *Reflections on TREC.* Information Processing and Management, 1995. 31: 291-314.
7. Buckley C and Voorhees EM, *Retrieval System Evaluation*, in *TREC: Experiment and Evaluation in Information Retrieval*, Voorhees EM and Harman DK, Editors. 2005, MIT Press: Cambridge, MA. 53-75.
8. Buckley C and Voorhees E. *Evaluating evaluation measure stability*. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000. Athens, Greece: ACM Press. 33-40.
9. Hersh WR, et al. *OHSUMED: an interactive retrieval evaluation and new large test collection for research*. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1994. Dublin, Ireland: Springer-Verlag. 192-201.
10. Hersh WR, *Interactivity at the Text Retrieval Conference (TREC).* Information Processing and Management, 2001. 37: 365-366.
11. Hersh WR, et al., *Enhancing access to the bibliome: the TREC 2004 Genomics Track.* Journal of Biomedical Discovery and Collaboration, 2006. 1: 3. http://www.j-biomed-discovery.com/content/1/1/3.
12. Clough P, et al. *Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks*. *Evaluation of Multilingual and Multi-modal Information Retrieval - Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*. 2006. Alicante, Spain: Springer Lecture Notes in Computer Science. in press. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/cloughOCLEF2006.pdf.
13. Hersh W, et al. *TREC 2005 Genomics Track overview*. *The Fourteenth Text Retrieval Conference - TREC 2005*. 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/GEO.OVERVIEW.pdf.