Lister Hill National Center for
Biomedical Communications
An Intramural Research Division of the U.S. National Library of Medicine

# Open-i$^{SM}$: imaging, informatics, natural language processing, and multi-modal information retrieval – research and development

## A Report to the Board of Scientific Counselors
## September 2016

Dina Demner-Fushman, MD, PhD
Sameer Antani, PhD
Suchet Chachra, MS
Michael Kushnir, JD
Soumya Gayen, MS

Communications Engineering Branch

**U.S. National Library of Medicine, LHNCBC**
**8600 Rockville Pike, Building 38A**
**Bethesda, MD 20894**

# Contents

## Abstract

The Open-i project combines research in text processing, image analysis and machine learning to create a system (also called Open-i) that enables about 10,000 users a day to retrieve relevant images and expanded citations from the open-access biomedical literature, as well as from clinical and historic image collections. Searching may be done by text as well as image queries. Images include a wide range of clinical imaging modalities, graphs, charts, photographs and other illustrations. The images are indexed by text in captions and mentions in the article, as well as by image features. This report presents the underlying research in natural language processing, biomedical image analysis, and informatics leading to the design, development and practical implementation of this system.

## Background

The importance of quick and easy access to images is well-established. For example, Reiner et al. have shown that improved access to a Picture Archiving and Communication System (PACS) containing radiology images resulted in a twofold increase in the number of radiology images reviewed by medical team members and an average saving of 44 minutes of clinician time [1]. In a survey of information needs of researchers and educators, Sandusky and Tenopir found that tables and figures are often the first parts of a scientific article scanned or read by researchers [2]. In addition, the survey participants indicated that having access to the illustrations[1] prior to obtaining the whole publication would greatly enhance their search experience. In the biomedical domain, Divoli et al. [3] asserted that bioscience literature search systems, such as PubMed, should show figures from articles alongside search results, and that captions should be searched, along with the article title, metadata, and abstract. In our earlier work, Simpson et al. [4] showed that, for the system presented in this report, retrieval of case descriptions similar to a patient's case was significantly improved with the use of image-related text.

The clear need for a multimodal retrieval system on the one hand, and the current state-of-the-art in natural language processing (NLP), information retrieval (IR) and content-based image retrieval (CBIR) techniques on the other, motivated us to implement a multimodal retrieval system (Open-i) for advanced information services. We presented a prototype to the LHNCBC Board of Scientific Counselors (BSC) and the NLM Board of Regents in 2010 and 2011. Both Boards also recommended scaling the system to large collections and to continue research and development of highly accurate image retrieval methods. Both Boards encouraged combining our research efforts to develop publicly-available services supporting real-time access to biomedical images. Implementing these recommendations provided for exciting research venues in the areas of image processing and visual question answering, full text processing and summarization, multimodal retrieval that combines  text and image features in image panel segmentation, modality classification and retrieval, as well as creating a system architecture for big data processing and seamless real-time image retrieval for hundreds of thousands of transactions a day (which includes human searchers as well as automated crawlers), search engine optimization, and responsive design, while ensuring security and addressing other NIH requirements for public-facing servers. Figure 1 shows the growth of Open-i collections since it was presented to BSC in 2011, and this report presents an overview of the latest research.

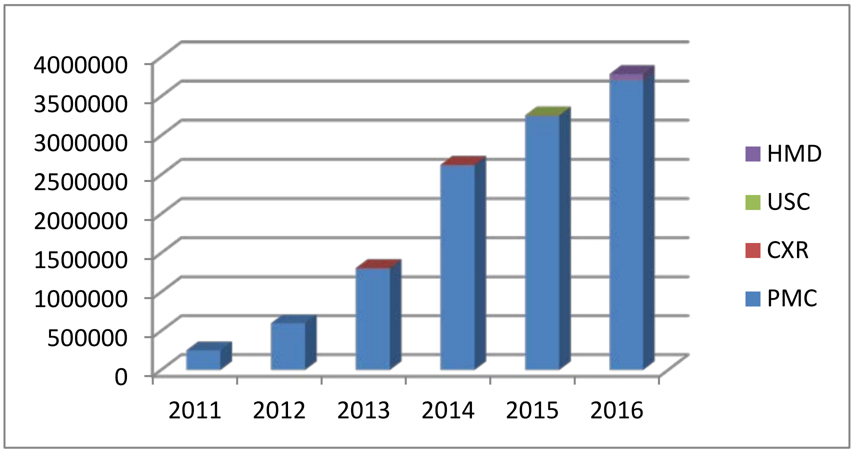


**Figure 1: Number of images in Open-i by year. PMC** – PubMed Central® Open Access subset; **CXR** -- Indiana University Chest X-Ray collection; **USC** -- University of Southern California Orthopedic collection; **HMD** -- Images from the History of Medicine Division of the U.S. National Library of Medicine

[1] We use "images," "illustrations," and "figures" interchangeably when referring to visual material in our set of medical articles and images.

The growth of the Open-i collection is paralleled by the growth in the number of unique users, up to 10,000 a day. Among the intentional uses for quick access to images and occasional requests for image reuse, Open-i has inspired the following example:

> Dear Open-i,
>
> During my search through the internet for medical images, i came across the Open-i website.
>
> It is well designed and very informative.
>
> The reason I'm looking for these website, is the following:
>
> We're going to shoot my second movie in August, which is about a young boy (hypochondriac) that gets to know an old mechanic.
>
> Slowly but surely, the world of measurements, charts and parts gets imprinted in the boys mind.
>
> It would be great if I can use your website in the film it will be just some shorts shots of the boy using the website and scrolling past the images.
>
> Greetings,
>
> Maurice (Netherlands)

## Objectives

The objectives of the NLM Open-i project are two-fold: first, to investigate methods for image and text processing to enhance access to image collections in biomedicine and second, to research system architecture, search engine optimization and graphical user interface design that enable high-quality real-time user experience with rapidly growing amounts of data. The system should enable:

- Searching by textual, visual and hybrid queries
- Retrieving illustrations (medical images, charts, graphs, diagrams)
- Retrieving bibliographic citations, enriched by relevant images
- Retrieving from collections of journal literature, patient records, and independent image databases
- Linking patient records to the literature and image databases, to support visual diagnosis and clinical decision making

Open-i success is measured in contributions to multimodal indexing, NLP, and image processing research while maintaining and improving access to biomedical information.

## Significance

Based on considerable evidence for a strong need to supplement traditional bibliographic citations with relevant visual material, the Open-i project has demonstrated the feasibility of providing high-quality multimodal retrieval services to NLM patrons and the biomedical community. Delivering such services would be essentially unaffordable if the citations enhanced with relevant visual materials were to be manually created, in contrast to the automated methods we have developed.

The automated techniques developed for Open-i are essential for processing large amounts of images and millions of small files, which poses unique challenges. Furthermore, these techniques offer building blocks for the development of advanced information services that enable users to search by textual as well as visual queries, and retrieve citations enriched by relevant images, charts, graphs, diagrams, and other illustrations, not only from the journal literature, but also drawn from patient records and independent image databases. In addition to promoting greater, and more targeted access to the biomedical literature, our techniques promise to enhance visual diagnoses and clinical decision support.

Since the beginning, the Open-i project has provided an excellent training venue for young researchers as well as numerous in-house and worldwide collaborations. Some of the innovative Open-i algorithms, such as multi-panel figure segmentation, translation of image features into visual words to facilitate indexing, and detection of regions of interest in the images were developed by interns and post-doctoral fellows working with in-house researchers. An example of international collaboration is the ImageCLEFmed community-wide evaluation of various aspects of medical image processing and multi-modal retrieval that the Lead Investigators of the Open-i project organized in collaboration with the international community of researchers [5, 6].

## Methods and Procedures

The methods fall into several distinct areas:  1) text and image processing that come together to produce MEDLINE citations enriched with image-related information (henceforth, "enriched citations") as shown in Figure 2; 2) multimodal retrieval; 3) system architecture to enable timely processing of millions of documents and reliable 24/7 access to the Web site; 4) responsive GUI design to enable optimal viewing and navigation of the Web site across devices ranging from phones to desktop computers; and 5) search engine optimization to improve access to information.

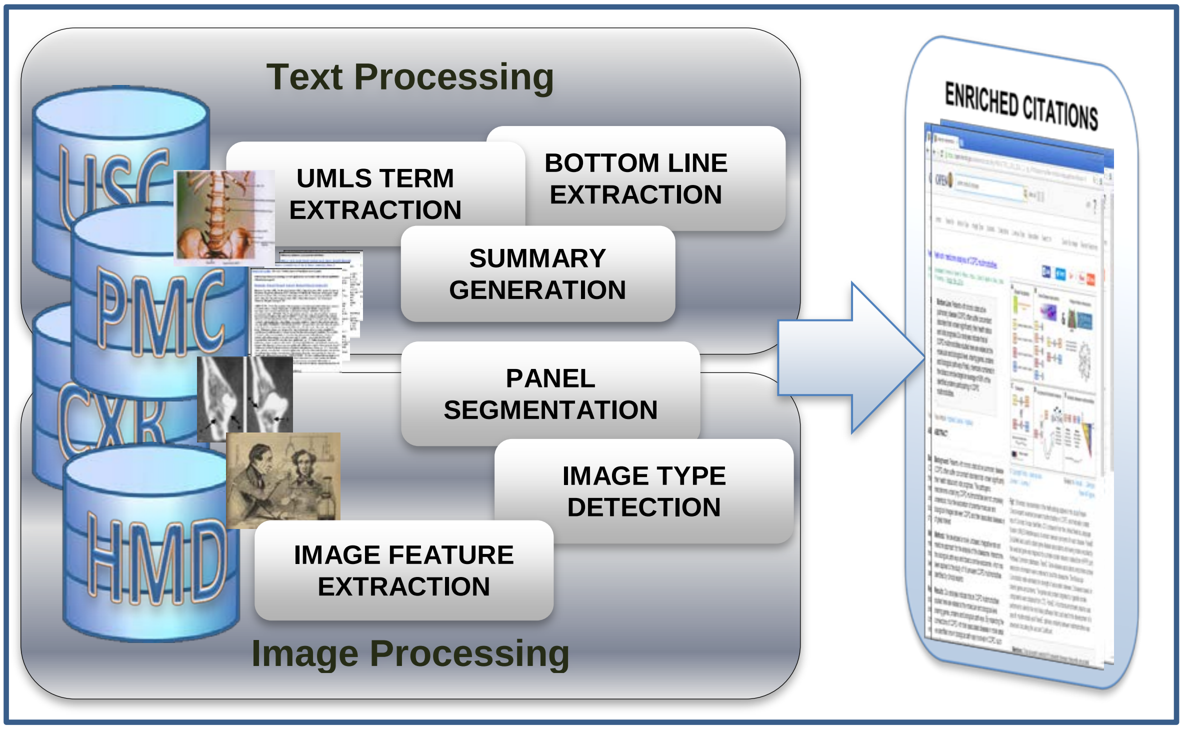


**Figure 2 Open-i document processing pipeline:** The complex text and image processing pipelines consist of many basic modules, such as caption extraction and named entity recognition in text and visual keywords generation and image feature extraction from the images. The modules highlighted in the figure are continually updated or have been developed recently and are discussed in-depth in this report.

In this section we provide a summary of the basic approaches described in-depth in our 2010 report [7] and more detailed descriptions of recent research.

## Data Sources

The sources of the images and text data in Open-i are shown on the left in Figure 2. The data are obtained from the open access subset of PubMed Central® (PMC®), using the PMC file transfer protocol (FTP) services[2]; by acquiring data from the Indiana University hospital network; by linking to the Orthopedic Surgical Anatomy Teaching Collection[3] at the USC Digital Library; and images from the NLM History of Medicine Division[4]. Open-i collection contains over 3.7 million images from about 1.2 million PubMedCentral articles; 7,470 chest x-rays with 3,955 radiology reports from Indiana; 67,517 images from NLM History of Medicine collection; and 2,064 orthopedic illustrations.

## Building Blocks – Text and Image Processing

To prepare documents for indexing and retrieval, we combine our tools and those publicly available, in a pipeline that starts with acquiring data and ends in the generation of enriched citations shown in Figures 2 and 3 respectively. The initially separate text and image processing pathways merge to detect image types, segment multi-panel images and create multimodal indexes, for use with specialized multimodal information retrieval algorithms. The text processing module extracts descriptions of images and image captions from the full text articles to enrich the MEDLINE citation of the article containing the image. The image processing module extracts low-level visual features used in image modality classification and image clustering. The image clusters are labeled with alphanumeric strings (“cluster words”). Subsequently, image features are represented using the cluster words. The cluster words pertaining to an image are added to its enriched citation, along with the image modality label.

The output of the document processing pipeline is a set of enriched MEDLINE/PubMed citations and documents from other collections converted to this standard XML format that is subsequently indexed with a domain-specific search engine Essie [8]. We are currently exploring if the same functionality can be achieved with the open-source search engine Lucene™[5].

The Open-i document processing system is developed in Java, and uses Hadoop MapReduce (Apache Software Foundation, Los Angeles, CA, USA) to parallelize text processing and image feature extraction.

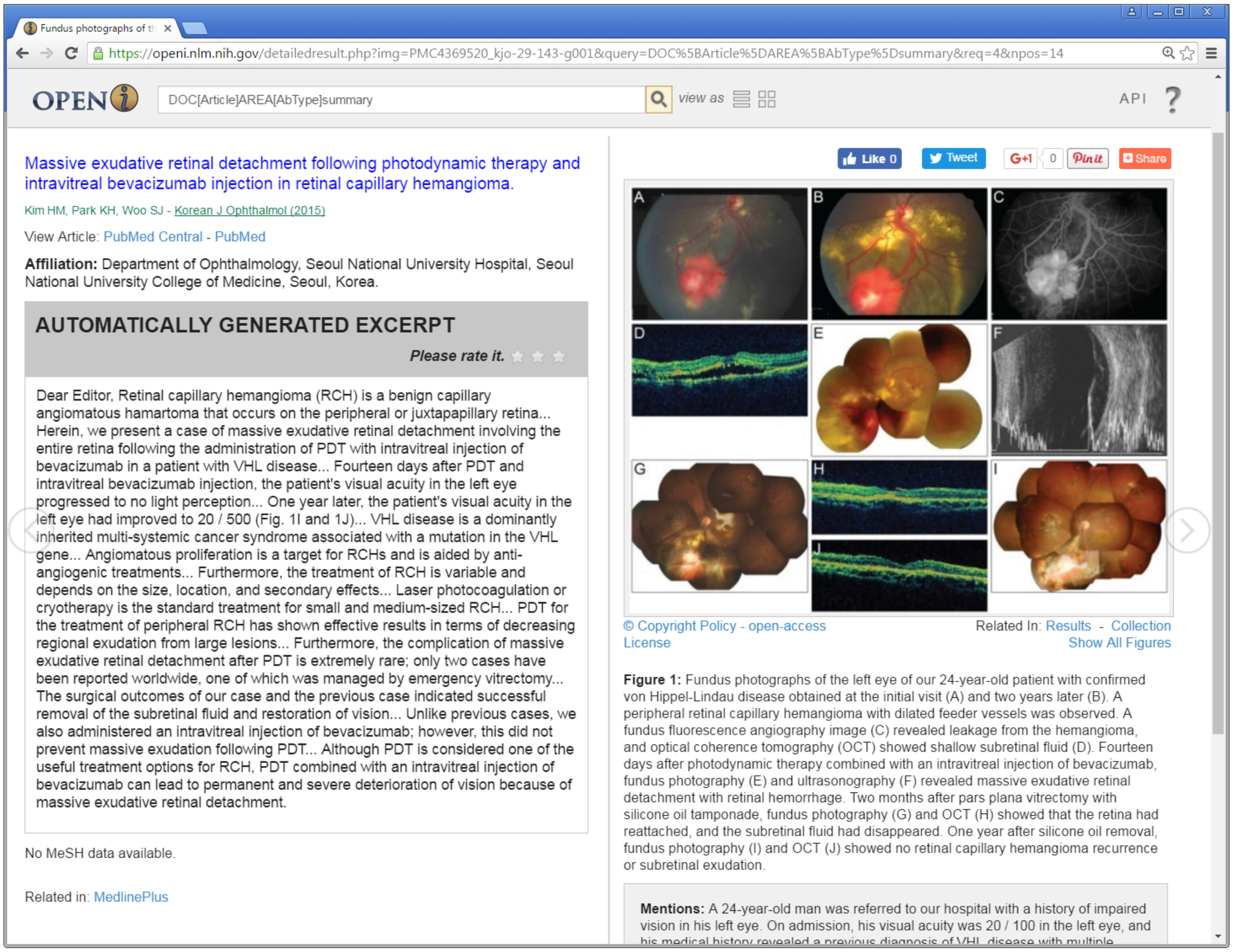
One challenge in image processing arises from several illustrations combined into one figure. These multipanel (or compound) images (see Figure 3) found in many articles reduce the quality of image features, if the features are extracted from the whole image. For feature extraction, therefore, these images need to be first separated into distinct panels, as described below. In addition to the text and image features necessary for retrieval, each enriched citation also contains meta-information derived from the basic features (such as the medical terms found in the captions and mapped to unified medical language system (UMLS) [9] concepts). This meta-information is used to filter and re-rank search results. For example, the results could be restricted to radiology images using the modality classification results, or re-ranked to promote articles focused on genetics (identified as such by genetics-related concepts in the titles, Medical Subject Heading (MeSH) terms, and captions). In addition to the multi-panel image, a detailed view of an enriched citation in Figure 3 shows the results of automated text summarization for one of about 50,000 full-text articles that do not have author-generated abstracts. We present this example of a recently developed text processing approach next.

[2] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[3] http://digitallibrary.usc.edu/cdm/landingpage/collection/p15799coll50

[4] https://www.nlm.nih.gov/hmd/

[5] https://lucene.apache.org/



Fundus photographs of th
https://openi.nlm.nih.gov/detailedresult.php?img=PMC4369520_kjo-29-143-g001&query=DOC%5BArticle%5DAREA%5BAbType%5Dsummary&req=4&npos=14
OPEN i
DOC[Article]AREA[AbType]summary
view as
API
?
Like 0
Tweet
G+1
0
Pin it
Share
A
B
C
D
E
F
G
H
J
I

Massive exudative retinal detachment following photodynamic therapy and intravitreal bevacizumab injection in retinal capillary hemangioma.

Kim HM, Park KH, Woo SJ - Korean J Ophthalmol (2015)

View Article: PubMed Central - PubMed

**Affiliation:** Department of Ophthalmology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Korea.

**AUTOMATICALLY GENERATED EXCERPT**

*Please rate it.* ☆ ☆ ☆

Dear Editor, Retinal capillary hemangioma (RCH) is a benign capillary angiomatous hamartoma that occurs on the peripheral or juxtapapillary retina... Herein, we present a case of massive exudative retinal detachment involving the entire retina following the administration of PDT with intravitreal injection of bevacizumab in a patient with VHL disease... Fourteen days after PDT and intravitreal bevacizumab injection, the patient's visual acuity in the left eye progressed to no light perception... One year later, the patient's visual acuity in the left eye had improved to 20 / 500 (Fig. 1I and 1J)... VHL disease is a dominantly inherited multi-systemic cancer syndrome associated with a mutation in the VHL gene... Angiomatous proliferation is a target for RCHs and is aided by anti-angiogenic treatments... Furthermore, the treatment of RCH is variable and depends on the size, location, and secondary effects... Laser photocoagulation or cryotherapy is the standard treatment for small and medium-sized RCH... PDT for the treatment of peripheral RCH has shown effective results in terms of decreasing regional exudation from large lesions... Furthermore, the complication of massive exudative retinal detachment after PDT is extremely rare; only two cases have been reported worldwide, one of which was managed by emergency vitrectomy... The surgical outcomes of our case and the previous case indicated successful removal of the subretinal fluid and restoration of vision... Unlike previous cases, we also administered an intravitreal injection of bevacizumab; however, this did not prevent massive exudation following PDT... Although PDT is considered one of the useful treatment options for RCH, PDT combined with an intravitreal injection of bevacizumab can lead to permanent and severe deterioration of vision because of massive exudative retinal detachment.

No MeSH data available.

Related in: MedlinePlus

© Copyright Policy - open-access License

Related In: Results - Collection
Show All Figures

**Figure 1:** Fundus photographs of the left eye of our 24-year-old patient with confirmed von Hippel-Lindau disease obtained at the initial visit (A) and two years later (B). A peripheral retinal capillary hemangioma with dilated feeder vessels was observed. A fundus fluorescence angiography image (C) revealed leakage from the hemangioma, and optical coherence tomography (OCT) showed shallow subretinal fluid (D). Fourteen days after photodynamic therapy combined with an intravitreal injection of bevacizumab, fundus photography (E) and ultrasonography (F) revealed massive exudative retinal detachment with retinal hemorrhage. Two months after pars plana vitrectomy with silicone oil tamponade, fundus photography (G) and OCT (H) showed that the retina had reattached, and the subretinal fluid had disappeared. One year after silicone oil removal, fundus photography (I) and OCT (J) showed no retinal capillary hemangioma recurrence or subretinal exudation.

**Mentions:** A 24-year-old man was referred to our hospital with a history of impaired vision in his left eye. On admission, his visual acuity was 20 / 100 in the left eye, and his medical history revealed a previous diagnosis of VHL disease with multiple

**Figure 3: Enriched MEDLINE/PubMed citation:** the detailed view presents bibliographic information and an automatically generated summary on the left, and a multi-panel image with its caption and description extracted from the text (mention) on the right. Additional links allow viewing license information extracted from the text, MedlinePlus, social media and searching for visually similar results within search results or an entire collection.

## Text processing: Extractive Summarization of Missing Abstracts

We combined two methods, an in-house bottom line extractor [10] and key-phrase based classifier, to produce baseline summaries for about 50,000 articles in the open access subset of PMC collection. The bottom-line extractor computes the likelihood of containing a health outcome for each sentence in full text. This technique employs an ensemble of semantic, rule-based, Naive Bayes, n-gram, positional and document-length classifiers. The rule-based classifier analyzes each sentence based on existence of cue phrases indicating bottom line, e.g., *significant differences*. The Naive Bayes classifier generates a likelihood score based on a bag of words representation of the sentence. The n-gram based classifier looks for uni- and bi-grams that provide a high information gain measure and are strong positive predictors of outcomes, e.g. *superior*. Positional and document length classifiers evaluate the position in the supplied text and the length of each sentence to provide probability estimates. The semantic classifier detects biomedical concepts belonging to specific semantic types within the sentence and concept discovery information from previous sentences to generate a likelihood score. The probability scores from each classifier are combined to compute the final score for each sentence.

The key-phrase based classifier follows the ideas introduced by Luhn who extracted the most significant sentences from technical papers to automatically generate abstracts [11]. The significance of a sentence was measured by the significance and position of the words in the sentence, whereas the significance of a word was measured by the frequency of its normalized form in the paper, after excluding what is now known as stop words, i.e., words that could be found in any text, such as articles, pronouns, etc. We compared the

Keyphrase Extraction Algorithm (KEA) [12], which identifies key phrases using term-frequency / inverse document frequency of the phrases and a Naïve Bayes classifier, to Microsoft Text Analytics (MSTA) which we accessed online through Microsoft Azure Machine Learning suite. For both key-phrase extraction methods, we defined most salient sentences as those that contain the higher density of distinct key phrases. We then experimented with selecting most salient sentences using textual entailment[6], similar to the graph-based algorithm proposed by Gupta et al. [13].

To evaluate the performance of the approaches, two annotators judged the summaries generated by the three base methods for 1) coverage of the main points in the article, 2) informativeness, and 3) extraction of bottom-line. The summaries were presented to the judges in random order and the judges were blinded to the system that generated a given summary. The annotators also manually generated extractive summaries for the same 300 articles. The manual judgments correlated sufficiently well with the Rouge scores, a widely used automatic approach to summarization evaluation [14], which allowed us to use the manually generated summaries to judge further developments of the summarization system. The overlap between KEA and MSTA summaries was significant, with KEA getting better scores from both judges. Merging the KEA-based and the bottom-line extraction summaries resulted in the best performance, with different approaches to selecting sentences using textual entailment increasing recall, while keeping precision stable and under-performing the upper bound established by inter-annotator agreement at 43% $F_1$ score only by 8%.

### Image Feature Extraction

Low-level visual features, such as color, texture, edges, and shape, are individually insufficient for capturing image semantics, but they are necessary primary building blocks for describing the visual content in an image. They are very effective if judiciously selected features are incorporated into a suitable machine-learning framework that supports multi-scalar and concept-sensitive visual similarity.

Images in the open access PMC collection are of different sizes. In order to obtain a uniform measure of visual content while maintaining computational efficiency we compute features from images that are normalized to a common size measuring 256 x 256 pixels.

We evaluated feature extraction and similarity computation for larger size images (512 x 512, 768 x 768, and 1024 x 1024) and found no significant improvement in retrieval results, but a much greater computational cost. However, higher resolutions may be more useful in the future for visual question-answering where responses to multimodal image and text queries could be other images significant parts of which are mapped to visual concepts and used in matching. We will revisit this issue in future research.

Color plays an important role in the human visual system and measuring its distribution can provide valuable discriminating data about the image. We evaluated several image descriptors to represent the color in the image [16]. Results from the analysis indicated that descriptors that captured both the color and its layout were most effective. To represent the spatial structure of the color in images, we utilize the Color Layout Descriptor [17] (CLD) specified by MPEG-7[7]. The CLD represents the spatial layout of the color, both the luminance and the chroma information, in images in a compact form.

Texture measures the degree of "smoothness" (or "roughness") in an image. This "tactile" measure is expressed by the perturbations in the intensity of image pixels and edges in a given region. When combined with the color layout it can be a powerful visual descriptor. In our study, the Fuzzy Color Texture Histogram

[6] "Textual entailment is defined as a directional relationship between pairs of text expressions, denoted by T - the entailing "Text", and H - the entailed "Hypothesis". We say that T entails H if the meaning of H can be inferred from the meaning of T, as would typically be interpreted by people" [15].

[7] http://mpeg.chiariglione.org/standards/mpeg-7

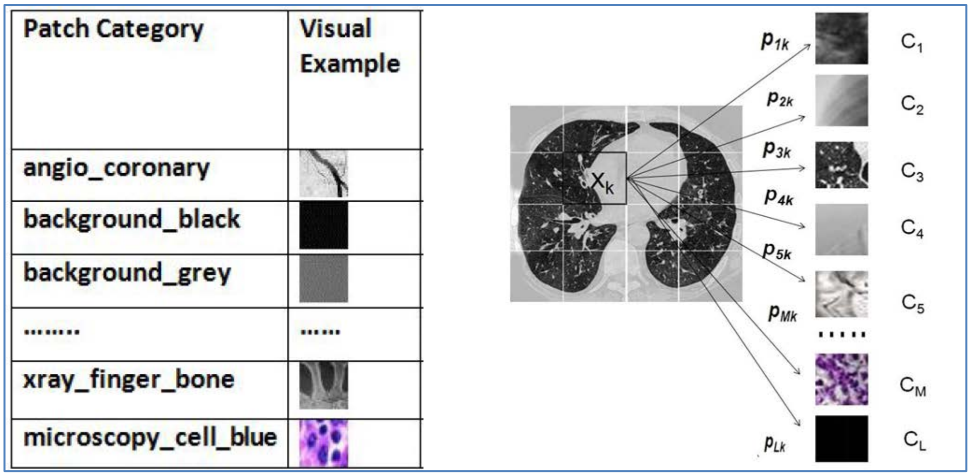
(FCTH) [18] from Apache Lucene Image Retrieval (LIRe[8]) library was found to be very effective in comparison with other texture features. Another feature, the Color Edge Direction Descriptor (CEDD) [19], also from LIRe, incorporates color and texture information into a single histogram and requires low computational power compared to MPEG-7 descriptors. This descriptor is robust with respect to image deformation, noise, and smoothing.

Edges are useful in determining object outlines and capturing overall image content layout. The Edge Histogram Descriptor [17] (EHD), also specified by MPEG-7, captures local edge distributions (strength, and orientation) and quantizes it into a directional histogram.

In addition to the image-appearance based features, we also developed a meta-image feature that combines these low level features with biomedical concepts. The SemanticConcept or SConcept feature [20, 21] shown in Figure 4, extends the recently popular "Bag of Visual Words" (BoVW) method [22]. In the BoVW approach, generally the low-level visual features at local regions, or patches, are vector quantized to generate "visual words". This allows one to treat each image as a "visual document" and explore a variety of region-based image matching techniques that follow from prior work in information retrieval research. Although the method has proved to be effective for visual content representation, the unsupervised clustering to generate the words or dictionary largely neglects the semantic context of the local features. As a result, commonly generated visual words are still not as expressive as keywords in text documents.

Nevertheless, in a heterogeneous collection of medical images from journal articles, such as in Open-i, it is critical that we annotate images with semantic concepts in addition to computing visual descriptors. One approach is to identify specific local patches that are perceptually and/or semantically distinguishable, such as homogeneous texture patterns in gray level radiological images, or differential color and texture structures in microscopic pathology and dermatologic images. For example, if we consider a computed tomography (CT) image of the lung or chest, which appears in many radiographic or medical journals, we observe several image regions with texturally different visual patterns that are semantically distinguishable from each other, such as a slightly bright and hazy appearance that can be mapped to the pattern "Ground Glass" opacity, or hexagonal structures that can be mapped to a "Honey Combing" pattern.  As illustrated in Figure 4, the variation in these local patches can be effectively modeled by using supervised classification techniques, such as multi-class SVM, to perform concept-based image representation (i.e., "Bag of Visual Concepts") that may be used to enable concept-based image region-of-interest retrieval [23].

[8] http://www.lire-project.net



| Patch Category | Visual Example |
|---|---|
| angio_coronary | |
| background_black | |
| background_grey | |
| …….. | …… |
| xray_finger_bone | |
| microscopy_cell_blue | |

**Figure 4: Example image patches (left) and concept (C1-CL) based image annotation process (right)**

## Image Type Detection

Searching and filtering retrieval results by image type is essential to efficient image retrieval. As mentioned before, images in Open-i are of various types (imaging modalities) and can be loosely categorized by appearance, and/or acquisition method. This hierarchy [24] is illustrated in Figure 5. The taxonomy is not intended to be exhaustive, but largely tries to distinguish images based on appearance.

Some of the prominent and visually distinct image types shown in Figure 5 are currently used in Open-i. We plan to introduce other types as our research progresses in classifying visual content understanding of images that may be highly similar, and yet categorically different, for example, endoscopic images that could be mistaken for funduscopic.

For the classification, we use the image features described above and apply supervised machine learning using SVM classifier to train models that distinguish the following types: Graphics, X-ray, CT Scan, MRI, Microscopy, PET, Photographs, and Ultrasound.

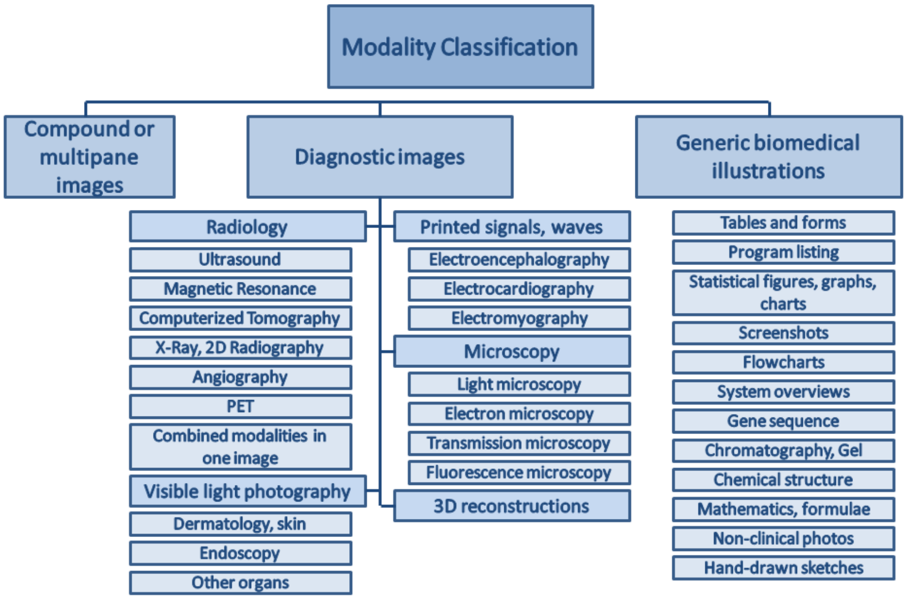


**Figure 5: Image type (modality) classification hierarchy**

Our method uses an SVM to classify images into multiple modality categories. The degree of membership in each category can then be used to compute the image modality. In its basic formulation, the SVM is a binary classification method that constructs a decision surface and maximizes the inter-class boundary between the samples. To extend it to multi-class classification, we combine all pair-wise comparisons of binary SVM classifiers, known as one-against-one or pair-wise coupling (PWC). The PWC method constructs binary SVMs for all possible pairs of classes. Hence, for M classes this method uses M * (M-1) / 2 binary classifiers, each of which provides a partial decision for classifying an image. Each SVM is trained for a single image feature. The class with the greatest probability for each feature accumulates one vote. The class with the greatest number of votes after classifying for all features is deemed to be the winning class, and the modality category of the class is assigned to the image. This category is used when a user requests a specific image type: a hard constraint on exact match on the assigned image modality category of the enriched citation is imposed.

## Multimodal processing: Multi-panel figure segmentation

The panel splitting module that splits multi-panel images (Figure 3) combines the output of the caption processing module that uses patterns to predict the number of panels and panel labels and two image-feature based modules detecting panel boundaries and panel labels.

The caption processing module determines if the caption belongs to a multi-panel figure. This rule-based system is looking for sequences of alphanumeric characters that are included within repeating tags, or followed by a repeating punctuation sign (for example, A. B. C.)

The image-feature based panel segmentation module determines if an image contains homogeneous regions that cross the entire image. If no homogeneous regions are found, the image is classified as single-panel. If the homogeneous regions are found, the panel segmentation module iteratively determines if each panel contains homogenous regions, and finally outputs coordinates of each panel.

The second image-feature based module attempts to detect labels in the panels [25]. The module first binarizes the image into black and white pixels, then searches the image for connected components that could represent panel labels, and then applies optical character recognition (OCR) methods to the connected components. Finally, the most probable label sequences and locations are selected from all candidate labels using Markov Random Field modeling.

The splitting module takes the outputs of the caption splitting, panel segmentation and label detection modules and splits the original figure if the following conditions are met: all three modules agree on the number of panels and the caption splitting and label detecting modules agree on the labels (this happens for approximately 30% of the multi-panel figures). If the panel segmentation or the label detection modules fail completely, the image cannot be split. However, if the modules partially agree on labels and position of some of the labels at the corners of the corresponding panels, heuristics help to compensate for the partial errors of individual modules and the combined information helps correctly split another 40% of the multi-panel images [26]. For example, if the OCR module outputs labels A, C, and D; the panel splitting module identifies four panels; and the caption splitting predicts A, B, C, D, then label B is inferred for the panel with the missing label.

### Cluster Words Generation

Content-based image retrieval systems commonly define visual similarity as a distance between extracted visual descriptors. Computing this distance exactly can be an expensive operation that increases the response time of a retrieval system. In order to avoid the computational complexity of computing these distances, we create a textual representation of the descriptors that we integrate with our existing textual features following our global feature mapping approach [27]. For each descriptor, the method clusters the feature vectors extracted from all the images in the collection using a hierarchical version of the k-means++ algorithm [28]. The method then assigns each cluster of features a unique alphanumeric code word and represents each image as a bag of these "visual words." The method produces, for each image, a textual signature of our visual features that has the advantage of being indexed and searched using traditional text-based information retrieval systems.

### Generation and Indexing of Enriched Citations

The final step in processing of PMC articles is assembling the following parts into one XML document: the original PubMed/MEDLINE citation; "visual words" that represent image features; bottom-line advice extracted from the abstract or an automatically generated summary for the articles without abstracts; image modality, caption, mention, licensing information and links to the original sources of the data. Other collections are processed similarly, after their native document structures are mapped to Open-i schema. Finally, indexing with Essie or Lucene produces a collection ready for searching and retrieval.

## System Architecture

To make possible the exceptionally high performance of Open-i, its production, development, and research environments are supported by the world-class infrastructure of the Lister Hill Center's Communications Engineering Branch (Figure 6).

The production environment is a fully-redundant architecture that has provided 99.999% uptime year to date with a total of 4 minutes of downtime year to date. The intensive mission of real-time text and image search is carried out by four application servers each equipped with 32 processors, 256GB of RAM, and 4 terabytes of solid state storage. High-availability and fault-tolerance is provided by a failover cluster of two intelligent load balancers that are able to not only detect whether the application servers are up, but also perform sanity checks before allowing a given application server to accept user traffic. An additional application server at the NCCS facility in Virginia serves as a failover site should the entire NLM site fail.

The development and application-support environment is a combination of bare-metal and virtual resources. One pure-hardware development server is augmented by 6 virtual machines supported by our state-of-the-art private virtualization cloud and NetApp mass storage.

The research environment for Open-i is supported entirely by our private cloud. Twenty four virtualization hosts (offering a total of 512 processors and 4 terabytes of RAM), two NetApp units  providing 160 terabytes of fully-redundant storage, and 30 terabytes of state-of-the-art distributed and replicated storage support Open-i research and development initiatives. The private cloud enables the project to quickly allocate and re-allocate massive computing resources as necessary and to employ distributed computing platforms such as Apache Hadoop, Apache Spark, and Apache Strom to overcome various Big Data and Big Compute challenges particularly inherent in the processing of images.

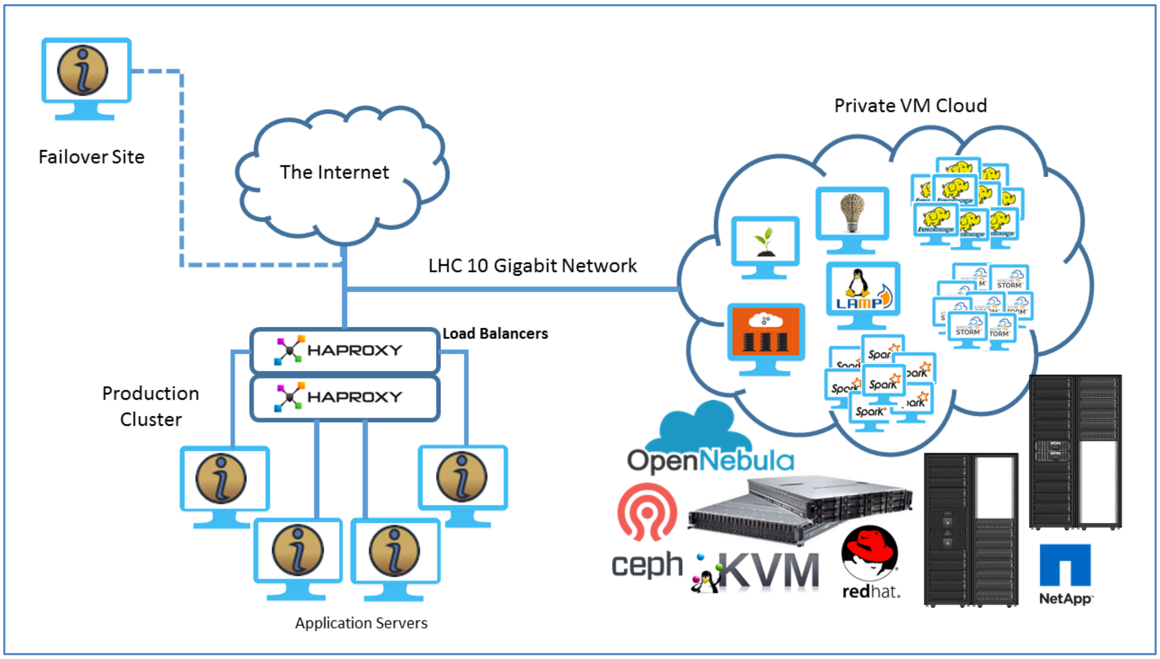


**Figure 6: Open-i IT Infrastructure**

With the exception of the two NetApp units, the vast majority of the project’s infrastructure is based on commodity hardware as well as open-source and internally-developed software with no dependencies on proprietary technologies or risk of vendor lock-in or skyrocketing costs. The majority of the hardware supporting our private cloud (including the NetApp units and 80% of the virtualization hosts) were obtained from government surplus pools at no cost to the NLM.

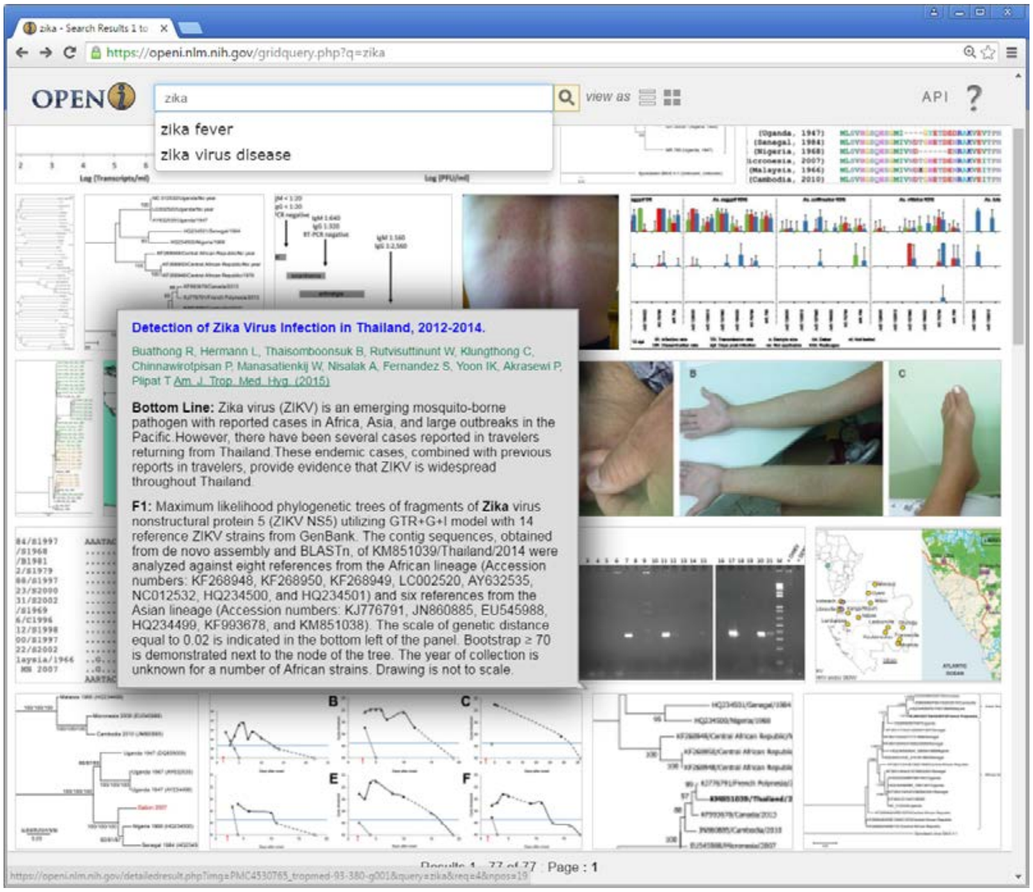
## Responsive Graphical User Interface

The Open-i system supports image retrieval for textual, visual and hybrid queries. The images submitted as queries are represented using cluster words as described above in the cluster word generation section. After this processing step, the cluster words are treated as any other search terms.

Based on the principles developed by Hearst et al. [29], the search results are displayed either on a grid that allows a view of most relevant 100 retrieved images (as shown in Figure 7) or as a traditional list. In either layout, scrolling over the image brings out a pop-up window that, along with the traditional elements of search results display, such as titles and author names, provides captions of the retrieved images and short summaries of the articles.

Once the search results are displayed, the users can drill down to the full enriched citation, view all other images in a given paper, as well as find similar images in the results and in the entire collection.

The search results can also be filtered using the following facets: 1) image type; 2) subsets; 3) clinical specialties; 4) enriched citation fields. These facets are described below.


zika - Search Results 1 to
https://openi.nlm.nih.gov/gridquery.php?q=zika
OPEN i
zika
zika fever
zika virus disease
view as
API
?
Results 1 - 77 of 77 Page : 1
https://openi.nlm.nih.gov/detailedresult.php?img=PMC4530765_tropmed-93-380-g001&query=zika&req=4&npos=19

Detection of Zika Virus Infection in Thailand, 2012-2014.

Buathong R, Hermann L, Thaisomboonsuk B, Rutvisuttinunt W, Klungthong C, Chinnawirotpisan P, Manasatienkij W, Nisalak A, Fernandez S, Yoon IK, Akrasewi P, Plipat T Am. J. Trop. Med. Hyg. (2015)

**Bottom Line:** Zika virus (ZIKV) is an emerging mosquito-borne pathogen with reported cases in Africa, Asia, and large outbreaks in the Pacific.However, there have been several cases reported in travelers returning from Thailand.These endemic cases, combined with previous reports in travelers, provide evidence that ZIKV is widespread throughout Thailand.

**F1:** Maximum likelihood phylogenetic trees of fragments of **Zika** virus nonstructural protein 5 (ZIKV NS5) utilizing GTR+G+I model with 14 reference ZIKV strains from GenBank. The contig sequences, obtained from de novo assembly and BLASTn, of KM851039/Thailand/2014 were analyzed against eight references from the African lineage (Accession numbers: KF268948, KF268950, KF268949, LC002520, AY632535, NC012532, HQ234500, and HQ234501) and six references from the Asian lineage (Accession numbers: KJ776791, JN860885, EU545988, HQ234499, KF993678, and KM851038). The scale of genetic distance equal to 0.02 is indicated in the bottom left of the panel. Bootstrap ≥ 70 is demonstrated next to the node of the tree. The year of collection is unknown for a number of African strains. Drawing is not to scale.

**Figure 7: Grid view of Open-i search results**

Image type: The image type filter is based on our classification of images into medical images modalities, such as MRI, x-ray, CT, ultrasound and others (Figure 5). In addition, the image type filter provides access to Videos downloaded with the PMC collection.

Subsets: Due to the nature of the collection, not all Medline/PubMed subsets, such as the core clinical journals subset are available in Open-i. We used the subject field of the NLM's List of Journals Indexed in MEDLINE to categorize the journals into clinical specialties and subsets. Where available, we used the subset field of the original MEDLINE citation.

Enriched citation fields: The users can search the text in any combinations of the following: titles, abstracts, captions, mentions, MeSH terms, and author names.

Finally, the search results can also be re-ranked according to the users' interests along the following dimensions: 1) by the date of publication (most recent or oldest first; 2) by the clinical task that is discussed in the paper (diagnosis, cause of the problem, prevention, prognosis, treatment, etc.).

## Project Status

The Open-i project has provided access to a constantly growing collection of biomedical images for six years. Open-i services appear to be very useful, judging by the growing numbers of distinct users a day and various innovative uses for its content and API. Open-i architecture is robust and capable of processing millions of images and hundreds of thousands of requests, e.g., when one of the Open-i images goes viral on Pinterest or Reddit the system seamlessly handles bursts of 200,000 views.

The main value of the project, however, is its contributions to research in image processing, NLP, multimodal retrieval, evaluation approaches and processing of big data. We are continually looking at ways to improve both the foundational approaches to text and image processing and the more complex tasks, such as segmentation of multi-panel images, region of interest detection and labeling, and visual question answering. The success of our approaches is attested in part by over 50 papers published in prestigious peer-reviewed journals and presented at international conferences.

We continuously encourage research interest in multi-modal retrieval and image processing by participating in and organizing community-wide evaluations and through LHC training program. Over the years, dozens of summer interns, Library Associates, postdoctoral fellows and visiting scientists were trained and contributed to Open-i research.

## Evaluation

Open-i evaluation is ongoing along three axes: 1) the algorithms and methods are evaluated in community-wide challenges and in-house, using the collections produced in these challenges; 2) the user interface is evaluated in in-house usability studies and through feedback submitted by the users; and 3) the services are evaluated using Google Analytics.

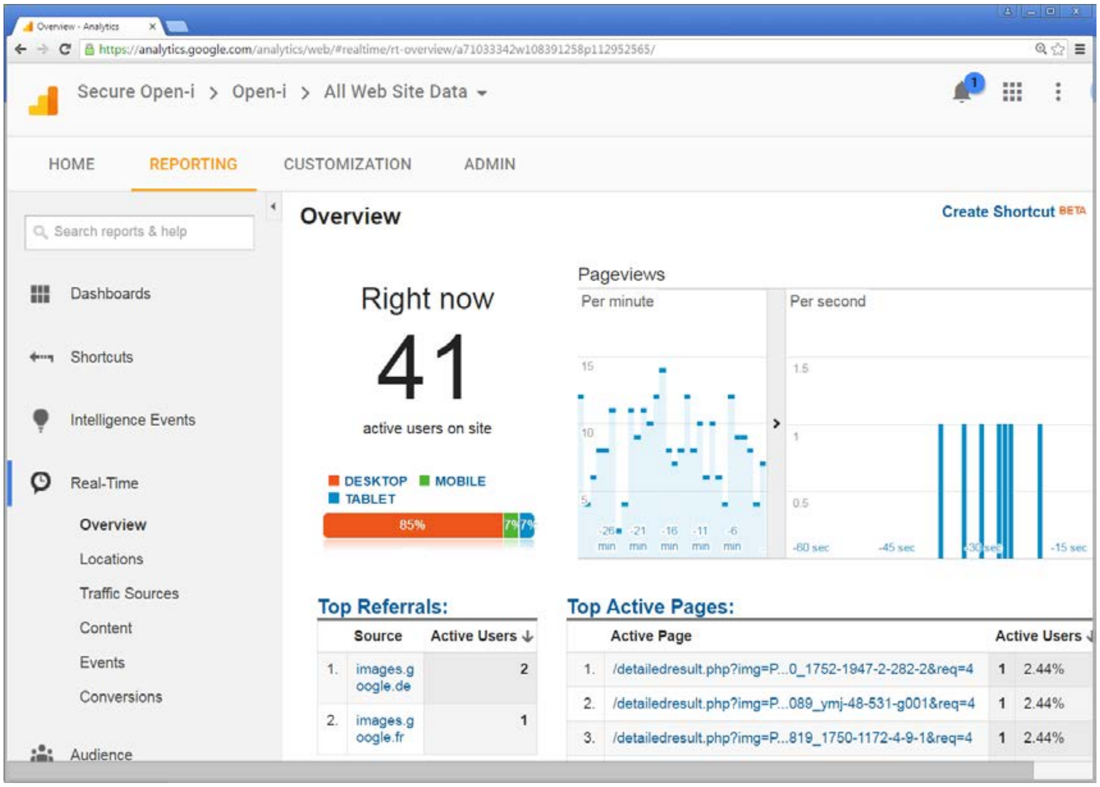
Our methods allowed us to be one of the top five groups for all tasks in all ImageCLEFmed evaluations over the years. In 2013, the last year that provided tasks relevant to Open-i research, our methods were in the top five for image modality classification, second best in compound figure segmentation and the best in retrieval, using an approach that effectively combines image and text features.

### Google Analytics – Worldwide Usage

Users from 226 countries access Open-i 24 hours a day, 7 days a week, 365 days a year. The service currently averages about 10,000 unique visitors per day, with a minimum of 120 unique users visiting the site at any given hour, and ~1,000,000 daily hits, of which 120,000 are directly to images, due to our search engine optimization strategies, and 60GB data transferred daily on average (1.8TB per month).

A large portion of the users comes from the United States; however, 65% of the traffic is international. India is the second largest consumer with 5.76% of overall users followed closely by the United Kingdom at 5.72% and Germany at 4.99%. The international user base is evenly distributed between developed and developing nations as well as countries where English is a primary language and those where English is rarely used. Open-i is widely used in Asia (India, China, Japan, Taiwan, Thailand, Singapore, Pakistan) and the Middle East (Turkey, Iran, Saudi Arabia, Israel). The service also provides information to users in some areas of the world that are most impoverished and in need for access to medical information and resources (i.e. Kenya, Palestine, Cameroon.)

Within the United States, visitors come from many US government entities, as well as the world's most prestigious educational and research institutions, medical schools, and hospitals. Over the last year, the service has been accessed at 929 educational institutions, 742 hospitals and at 97 U.S. based government organizations at the federal, state, and local levels. Users from educational institutions and medical schools such as Harvard, Stanford, Johns Hopkins, Yale, NYU, Columbia, Vanderbilt, Cornel, and MIT use the service daily. Medical institutions range from Kaiser Foundation, to the Nanjing Gulou Hospital in China, Oslo University Hospital in Norway, the University of Michigan Medical Center, and the St. Jude's Children's Hospital. Of the US government agencies, Open-i services are most frequently used by the Department of Veterans Affairs, followed by the National Institutes of Health (including NLM, NCI, and NIEHS), USDA, FDA, and the CDC. Other government consumers include the National Science Foundation, DoJ, DoE, EPA, DHS, CIA, FBI, USPTO, US Courts, 17 identifiable state governments, and 19 identifiable local and regional government entities.



Top Referrals:

| | Source | Active Users ↓ |
|---|---|---|
| 1. | images.google.de | 2 |
| 2. | images.google.fr | 1 |

Top Active Pages:

| | Active Page | Active Users ↓ | |
|---|---|---|---|
| 1. | /detailedresult.php?img=P...0_1752-1947-2-282-2&req=4 | 1 | 2.44% |
| 2. | /detailedresult.php?img=P...089_ymj-48-531-g001&req=4 | 1 | 2.44% |
| 3. | /detailedresult.php?img=P...819_1750-1172-4-9-1&req=4 | 1 | 2.44% |

**Figure 8: Real time monitoring of the site using Google Analytics**

## Summary and Future Work

Based on research showing that enriching citations with relevant images might significantly improve literature retrieval for scientific research and clinical decision making [2,3] the Open-i project was initiated to address issues in multimodal retrieval, which is enabled by bringing together innovative research in image and text processing. Our methods use text and image features extracted from relevant components in a document, database, or case description and create structured representations (the enriched citations). These enriched citations (that contain images and bottom-line advice) are presented to the user as search results. To evaluate and demonstrate our techniques, we have developed the image search engine that has firmly established itself as a leading biomedical image retrieval service, at the same time, providing a means for developing methods in image modality classification, multi-panel image segmentation, detection and labeling of regions of interest, and single document summarization. Our approaches have been shown to be among the best from over a dozen teams from around the world participating in the ImageCLEFmed contests. These competing teams represent academia, industry and other government organizations. Open-i has quickly grown into a service with national and global recognition. The usage statistics are a strong indicator of the global need for the unique and distinctive features the service offers.

We support research in image processing, multimodal retrieval and decision support providing access to Open-i data through an Application Program Interface (API) that has inspired such applications as EDDA[9] Lens that supports generation of systematic reviews and provides Open-i images to illustrate studies clustered by themes for, e.g., studies of potential prognostic biomarkers of oral squamous cell carcinoma. Open-i also provides the Indiana University chest x-ray collection to the research community: the dataset has been downloaded by many research groups.

As next steps, we will continue exploring methods to improve the accuracy of retrieval of literature and images suitable for clinical decision support. These steps include:

1) Regions of Interest / Visual question answering.
A visual-ontology based approach to improving relevance in multimodal information retrieval systems, such as Open-i, is to correlate image regions with related terms, anatomical, disease, procedural, etc., in the caption and/or mentions and the Methods section of the full text of the article. Our research has previously explored methods that identify author placed annotations on the images, such as arrows, symbols, or textual annotations that identify important regions of interest in images [30, 31]. Other approaches that we have explored include correlating image tiles [32], and visually salient key points on the image [33] with a visual concept dictionary. We propose to continue our efforts in this direction exploring use of deep learning methods for automatically understanding visual content. Early work in this direction [34] bears promise for future advanced exploration in visual question-answering techniques.

2) Exploring alternative newly developed search engines, such as Elastic Search and Essie 4.

3) Improving extraction of the salient points from patient cases (for example, distinguishing between the findings present in the case description as part of routine examination or the chief complaints)

4) Improving scalability of the feature extraction and document pre-processing methods.

## Acknowledgments

In addition to the authors of this report, Open-i is supported by this team:

Open-i project is supervised by Dr. Clem McDonald and Dr. George Thoma.

Image processing is supported by Dr. Zhiyun Xue.

Evaluation is provided by Dr. Laritza Rodriguez and Sonya Shooshan. MLS

Technical operations are supported by Joseph Chow.

Usability testing and ongoing quality control are provided by Tehseen Sabir and Lan Le.

The following trainees have significantly contributed to Open-i research: Dr. Matthew Simpson, Dr. Md Rahman, Dr. Daekeun You, and Dr. Emilia Apostolova.

[9] Evidence in Documents, Discovery, and Analysis

## References

1. Reiner B, Siegel EL, Hooper F, Protopapas Z. Impact of Filmless Imaging on the Frequency of Clinician Review of Radiology Images. J Digit Imaging. 1998;11:149 - 150.
2. Sandusky RJ, Tenopir C. Finding and using journal article components: impacts of disaggregation on teaching and research practice. J Am Soc Inf Sci Technol. 2008;59(6):970 - 982.
3. Divoli A, Wooldridge MA, Hearst MA. Full text and figure display improves bioscience literature search. PLoS One2010; 5(4):e9619.
4. Simpson MS, Demner-Fushman D, Thoma GR. Evaluating the Importance of Image-related Text for Ad-hoc and Case-based Biomedical Article Retrieval. AMIA Annu Symp Proc. 2010 Nov 13;2010:752-6.
5. Demner-Fushman D, Antani S, Kalpathy-Cramer J, Müller H. A decade of community-wide efforts in advancing medical image understanding and retrieval. Comput Med Imaging Graph. 2015 Jan;39:1-2
6. Kalpathy-Cramer J, de Herrera AG, Demner-Fushman D, Antani S, Bedrick S, Müller H. Evaluating performance of biomedical image retrieval systems-An overview of the medical image retrieval task at ImageCLEF 2004-2013. Comput Med Imaging Graph. 2015 Jan;39:55-61.
7. Demner-Fushman D, Antani S, Simpson MS, Rahman MM.  Combining Text and Visual Features for Biomedical Information Retrieval and Ontologies. Available online at https://lhncbc.nlm.nih.gov/files/archive/tr2010002.pdf
8. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc. 2007 May-Jun;14(3):253-63.
9. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Meth. Inform. Med. 1993; 32: 281-291
10. Demner-Fushman D, Lin J. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. Computational Linguistics. 2007;33(1):63-103.
11. Luhn HP. The automatic creation of literature abstracts. IBM J. Res. Dev. 1958; 2 (2);159-165.
12. Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG. KEA: practical automatic keyphrase extraction. In Proceedings of the Fourth ACM conference on Digital Libraries, August 11-14, 1999, Berkeley, CA, USA.
13. Gupta A, Kaur M, Mirkin S, Singh A, Goyal A. Text summarization through entailment-based minimum vertex cover. In Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014), 2014, Dublin, Ireland.
14. Lin CY. Rouge: A package for automatic evaluation of summaries. InText summarization branches out: Proceedings of the ACL-04 workshop 2004 Jul 25 (Vol. 8).
15. Sammons M, Vydiswaran V, Roth D. Recognizing textual entailment. Multilingual Natural Language Applications: From Theory to Practice. Prentice Hall, Jun. 2011.
16. Simpson MS, You D, Rahman MM, Xue Z, Demner-Fushman D, Antani S, Thoma G. Literature-based biomedical image classification and retrieval. Comp Med Imag Graph. 2015 Jan;39:3-13.
17. Chang, SF, Sikora, T, Puri A. Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology 11(6), 688{695 (2001)
18. Chatzichristos SA, Boutalis YS. FCTH: Fuzzy color and texture histogram: A low level feature for accurate image retrieval. In: Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services. pp. 191{196 (2008)

19. Chatzichristos SA, Boutalis YS. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) Proceedings of the 6th International Conference on Computer Vision Systems. Lecture Notes in Computer Science, vol. 5008, pp. 312{322. Springer (2008)
20. Rahman MM, Antani S, Thoma G.: A medical image retrieval framework in correlation enhanced visual concept feature space. In: Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems (2009)
21. Xue Z, Rahman MM, Antani S, Long LR, Demner-Fushman D, Thoma GR. Modality Classification for Searching Figures in Biomedical Literature. In: Proceedings of the 29th IEEE International Symposium on Computer-Based Medical Systems (2016)
22. Tsai C. Bag-of-Words Representation in Image Annotation: A Review, ISRN Artificial Intelligence, Vol. 2012, Article ID 376804, 19 pages, http://dx.doi.org/10.5402/2012/376804
23. Rahman M, Antani S, Demner-Fushman D, Thoma G. A Biomedical Image Representation Approach Using Visualness and Spatial Information in a Concept Feature Space for Interactive Region-of-Interest (ROI)-Based Retrieval. J. Med. Imag. 2015; 2(4):046502.
24. Müller H, Kalpathy-Cramer J, Demner-Fushman D, Antani S. Creating a classification of image types in the medical literature for visual categorization. Proc. SPIE 8319, Medical Imaging 2012: Advanced PACS-based Imaging Informatics and Therapeutic Applications, 83190P (February 23, 2012); doi:10.1117/12.911186.
25. You D, Antani S, Demner-Fushman D, Govindaraju V, Thoma GR Detecting figure-panel labels in medical journal articles using MRF. Proceedings of 11th International Conference on Document Analysis and Recognition (ICDAR 2011) (pp. 967-971). Beijing, China.
26. Apostolova E, You D, Xue Z, Antani S, Demner-Fushman D, Thoma G. Image Retrieval From Scientific Publications: Text and Image Content Processing to Separate Multipanel Figures. Journal of the American Society for Information Science and Technology (JASIST) Volume 64, Issue 5, pages 893–908, May 2013
27. Simpson MS, Demner-Fushman D, Antani SK, Thoma GR. Multimodal biomedical image indexing and retrieval using descriptive text and global feature mapping. Information Retrieval. 2014; 17(3): 229—264
28. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, SODA'07. January 7-9, 2007. New Orleans, Louisiana.
29. Hearst MA, Divoli A, Buturu H, Ksikes A, Nakov P, Wooldridge MA, Ye J. Biotext search engine: beyond abstract search. Bioinformatics 2007;23(16):2196–7.
30. You D, Simpson M, Antani SK, Demner-Fushman D, Thoma GR. A robust pointer segmentation in biomedical images toward building a visual ontology for biomedical article retrieval. Proc. SPIE 8658, Document Recognition and Retrieval XX, 86580Q (February 4, 2013);
31. KC S, Alam N, Roy PP, Wendling L, Antani S, Thoma G. A simple and efficient arrowhead detection in biomedical images. International J Pattern Recognition and Artificial Intelligence (IJPRAI), 30(5), 1657002, 2016.
32. Rahman MM, Antani SK, Demner-Fushman D, Thoma GR. Biomedical image representation approach using visualness and spatial information in a concept feature space for interactive region-of-interest-based retrieval. J Med Imaging (Bellingham). 2015 Oct;2(4):046502.

33. Pedrosa G, Rahman MM, Antani SK, Demner-Fushman D, Long LR, Traina A. Integrating visual words as bunch of n-grams for effective biomedical image classification. Proceedings of IEEE WACV 2014: IEEE Winter Conference on Applications of Computer Vision. Steamboat Springs, CO. March 24-26, 2014.
34. Shin HC, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016