IMGpedia: a Linked Dataset with Content-based Analysis of Wikimedia Images

Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan

Center for Semantic Web Research
Department of Computer Science, Universidad de Chile
{sferrada,bebustos,ahogan}@dcc.uchile.cl

Abstract. IMGpedia is a large-scale linked dataset that incorporates visual information of the images from the Wikimedia Commons dataset: it brings together descriptors of the visual content of 15 million images, 450 million visual-similarity relations between those images, links to image metadata from DBpedia Commons, and links to the DBpedia resources associated with individual images. In this paper we describe the creation of the IMGpedia dataset, provide an overview of its schema and statistics of its contents, offer example queries that combine semantic and visual information of images, and discuss other envisaged use-cases for the dataset.
Resource type: Dataset
Permanent URL: https://dx.doi.org/10.6084/m9.figshare.4991099.v2

1 Introduction

Many datasets have been published on the Web following Semantic Web standards and Linked Data principles. At the core of the resulting "Web of Data", we can find linked datasets such as DBpedia [6], which contains structured data automatically extracted from Wikipedia; and Wikidata [10], where users can directly add and curate data in a structured format. We can also find various datasets relating to multimedia, such as LinkedMDB describing movies, BBC Music describing music bands and genres, and so forth. More recently, DBpedia Commons [9] was released, publishing metadata extracted from Wikimedia Commons[1]: a rich source of multimedia containing 38 million freely usable media files (image, audio and video).

Related Work Amongst the available datasets describing multimedia, the emphasis has been on capturing the high-level metadata of the multimedia files (e.g., author, date created, file size, width, duration) rather than audio or visual features of the multimedia content itself. However, as mentioned in previous works (e.g., [8,1,4]), merging structured metadata with multimedia content-based descriptors could lead to a variety of applications, such as semantically-enhanced multimedia publishing, retrieval, preservation, etc. While such works have proposed methods to describe the audio or visual content of multimedia files in Semantic Web formats, we are not aware of any public linked dataset incorporating content-based descriptors of multimedia files. For example, DBpedia Commons [9] does not extract any audio/visual features directly from the multimedia files of Wikimedia Commons, but rather only captures metadata from the documents describing the files.

[1] http://commons.wikimedia.org

Contribution Along these lines, we have created IMGpedia: a linked dataset incorporating visual descriptors and visual similarity relations for the images of Wikimedia Commons, linked with both the DBpedia Commons dataset (which provides metadata for the images, such as author, license, etc.) and the DBpedia dataset (which provides metadata about resources associated with the image). The initial use-case we are exploring for IMGpedia is to perform *visuo-semantic* queries over the images, where, for example, using SPARQL federation over IMGpedia and DBpedia, we could request: *given a picture of the Cusco Cathedral, retrieve the top-k most similar cathedrals in Europe*. More generally, as discussed later, we foresee a number of potential use-cases for the dataset as a test-bed for research in the potentially fruitful intersection of the Multimedia and Semantic Web areas.

Outline In this paper, we describe the IMGpedia dataset[2]. We first introduce the image analysis used to extract visual descriptors and similarity relations from the images of Wikimedia Commons. Next we give an overview of the lightweight ontology used to represent the resulting visual information as RDF. We then provide some high-level statistics of the resulting dataset and the best-practices used in its publication. Thereafter, we provide some example visuo-semantic queries and their results. Finally we conclude with discussion of other use-cases we envisage as well as our future plans to improve upon and extend the IMGpedia dataset.

2 Image Analysis

Wikimedia Commons is a dataset of 38 million freely-usable media files contributed and maintained collaboratively by users. Around 16 million of these media files are images, which are hosted on a mirror server accessible via `rsync`[3]. We downloaded the images, with a total size of 21 TB, in order to be able to process them offline. The download took 40 days with a bandwidth of 500 GB/day. In order to facilitate later image processing tasks, we only consider images with (commonly supported) `JPG` or `PNG` encodings, equivalent to 92% of the images.

After the acquisition of the images, we proceeded to compute different *visual descriptors*, which are high-dimensional vectors that capture different elements of the content of the images (such as color distribution or shape/texture information); later we will use these descriptors to compute visual similarity between images, where we say that two images are visually similar if the distance between their descriptors is low. The descriptors computed are the following:

- **Gray Histogram Descriptor:** We transform the image from color to grayscale and divide it into a fixed number of blocks. A histogram of 8-bit gray intensities is then calculated for each block. The concatenation of all histograms is used to generate a description vector with 256 dimensions.
- **Histogram of Oriented Gradients Descriptor:** We extract edges of the grayscale image by computing its gradient (using Sobel kernels), applying a threshold, and computing the orientation of the gradient. Finally, a histogram of the orientations is made and used as a description vector with 288 dimensions.

[2] In a previous short paper, we proposed the idea of the project and gave details of initial progress [3]; this paper describes the dataset resulting from that initial work.

[3] `rsync://ftpmirror.your.org/wikimedia-images/`

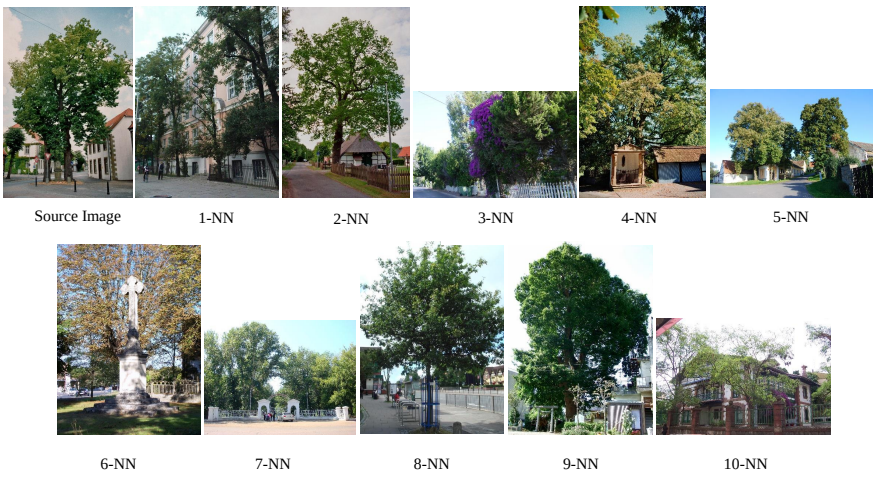


Fig. 1: 10 nearest neighbors of an image of Hopsten Marktplatz using HOG

– **Color Layout Descriptor:** We divide the image into blocks and for each block we compute the mean (YCbCr) color. Afterwards the Discrete Cosine Transform is computed for each color channel. Finally the concatenation of the transforms is used as the descriptor vector, with 192 dimensions.

Computing the descriptors was performed on a machine with Debian 4.1.1, a 2.2 GHz 24-core Intel® Xeon® processor, and 120 GB of RAM. With multi-threading, computing GHD took 43 hours, HOG took 107 hours, while CLD took 127 hours. We have made implementations to compute these visual descriptors available in multiple programming languages under a GNU GPL license [3][4].

The next task is to use these descriptors to compute the visual similarity between pairs of images. Given the scale of the dataset, in order to keep a manageable upper-bound on the resulting data (we selected ∼4 billion triples as a reasonable limit), we decided to compute the 10 nearest neighbors for each image according to each visual descriptor. To avoid $\binom{n}{2}$ brute-force comparisons, we use approximate search methods where we selected the Fast Library for Approximated Nearest Neighbors (FLANN) since it has been proven to scale for large datasets [7][5]. In order to facilitate multi-threading, we divide the images into 16 buckets, where for each image, we initialize 16 threads to search for the 10 nearest neighbors in each bucket. At the end of the execution we have 160 candidates to be the global 10 nearest neighbors so we choose the 10 with the minimum distances among them to obtain the final result. This process took about 13 hours with the machine previously described. In Figure 1 we show an example of the results of the similarity search based on the HOG descriptor, which captures information about edges in the image.

[4] https://github.com/scferrada/imgpedia

[5] We configured FLANN with a goal precision of 90% and tested it on a brute-forced gold standard of 20,000 images. FLANN achieved an actual precision of 79% on this dataset. However, while the gold standard took 3.5 days to compute with 16 threads, FLANN finished in 13 minutes with 1 thread. We concluded that FLANN offers a good precision/efficiency trade-off for a large-scale collection of images such as ours.

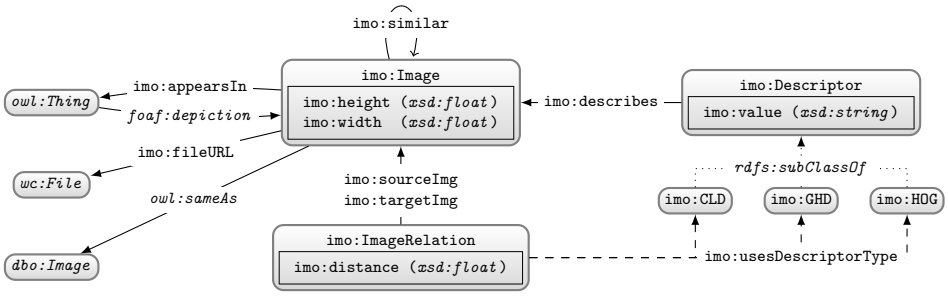


Fig. 2: IMGpedia ontology overview: classes are shown in boxes; solid edges denote relations between instances of both classes, dotted lines are between the classes themselves, while dashed lines are from instances to classes; external terms are italicized; datatype properties are listed inside the class boxes for conciseness.

3 Ontology and Data

The visual descriptors and similarity relations of the images form the core of the IMGpedia dataset. To represent this information as RDF, we create a custom lightweight IMGpedia ontology. All IMGpedia resources are identified under the `http://imgpedia.dcc.uchile.cl/resource/` namespace. The vocabulary is described in RDFS/OWL at `http://imgpedia.dcc.uchile.cl/ontology`; this vocabulary (authoritatively) extends related terms from the DBpedia Ontology, schema.org and the Open Graph Protocol where appropriate, and has been submitted to the Linked Open Vocabularies (LOV) service. In Figure 2, we show the classes, datatype- and object-properties available for representing images, their visual descriptors and the similarity links between them.

An `imo:Image` is an abstract resource representing an image of the Wikimedia Commons dataset, describing the dimensions of the image (height and width), the image URL in Wikimedia Commons, and an `owl:sameAs` link to the complementary resource in DBpedia Commons. In Listing 1 we see an example of the RDF for the `imo:Image` representation of Hopsten Marktplatz.

Listing 1: RDF example of a visual entity

```
@prefix imo: <http://imgpedia.dcc.uchile.cl/ontology#>
@prefix im: <http://imgpedia.dcc.uchile.cl/resource/>
@prefix dbcr: <http://commons.dbpedia.org/resource/File:>
im:Hopsten_Marktplatz_3.jpg a imo:Image ;
  owl:sameAs dbcr:Hopsten_Marktplatz_3.jpg;
  imo:width 400 ; imo:height 300 ;
  imo:fileURL <http://commons.wikimedia.org/wiki/File:Hopsten_Marktplatz_3.jpg>.
```

An `imo:Descriptor` respresents a visual descriptor of an image and is linked to it through the `imo:describes` relation. An `imo:Descriptor` can be of type `imo:GHD`, `imo:HOG`, or `imo:CLD` corresponding to the three types of descriptors previously discussed. In Listing 2 we show an example of a visual descriptor in RDF. To keep the number of output triples manageable, we store the vector of the descriptor as a string; storing individual dimensions as (192–288) individual objects would inflate the output triples to an unmanageable volume; in addition, we do not currently anticipate SPARQL queries over individual values of the descriptor.

Listing 2: RDF example of a descriptor

```
im:Hopsten_Marktplatz_3.jpg.HOG a imo:HOG ;
   imo:describes im: Hopsten_Marktplatz_3.jpg ;
   imo:value "[0.34418711, 0.10582313, 0.05867421, ...]".
```

An `imo:ImageRelation` is a resource that contains the similarity links between two images; it also contains the type of descriptor that was used and the Manhattan distance between the descriptors of both images. Although Manhattan distance is symmetric, these relations are materialized based on a k-nearest-neighbors (k-nn) search, where image a being in the k-nn of b does not imply the inverse relation; hence the image relation captures a source and target image where the target is in the k-nn of the source. We also add a `imo:similar` relation from the source image to the target k-nn image. Listing 3 shows an example of a k-nn relation in RDF.

Listing 3: RDF example of a visual similarity relation

```
im:176147ac95660a47d5d58c57d5260572cdce11f98ad4.HOG a imo:ImageRelation;
   imo:sourceImage im:Hopsten_Marktplatz_3.jpg ;
   imo:targetImage im:Boze_Cialo-glowny.JPG ;
   imo:distance 1.219660e+01 ; imo:usesDescriptor imo:HOG .

im:Hopsten_Marktplatz_3.jpg imo:similar im:Boze_Cialo-glowny.JPG .
```

Finally, aside from the links to DBpedia Commons, we also provide links to DBpedia, which provides a context for the images. To create these links, we use an SQL dump of English Wikipedia and perform a join between the table of all images and the table of all articles, so we can have pairs `(image_name,article_name)` if the image appears in the article. In Listing 4 we give some example links for DBpedia. Such links are not provided by DBpedia Commons.

Listing 4: RDF example of DBpedia links

```
im:Chamomile_original_size.jpg imo:appearsIn dbr:Nephelium_hypoleucum .
im:Rose_Amber_Flush_20070601.jpg imo:appearsIn dbr:Nephelium_hypoleucum .
im:Rose_Amber_Flush_20070601.jpg imo:appearsIn dbr:Acer_shirasawanum .
im:HondaS2000-004.png imo:appearsIn dbr:Alfa_Romeo_Scighera .
```

4 Dataset

The dataset of IMGpedia contains information about 14.7 million images of Wikimedia Commons, the description of their content, links to their most similar images and to the DBpedia resources that form part of their context. A general overview of the size and data of IMGpedia can be seen in Table 2. There we can see that for each visual entity we computed three different descriptors and for each descriptor we computed 10 similarity links using the 10 nearest neighbors, defining a similarity graph with 14.7 million vertices and 442 million edges.

Accessibility and Best Practices IMGpedia is available as a Linked Dataset (with dereferenceable IRIs), as a SPARQL endpoint (using Virtuoso), and as a dump. Locations are provided in Table 1. As aforementioned, we provide a lightweight RDF-S/OWL Ontology that extends well-known vocabularies as appropriate. We also provide a VoID description of the dataset, which includes metadata from DC-terms as well as brief provenance statement using the PROV ontology and licensing information. With respect to the license, the most restrictive licensing clauses allowed for

Table 1: Locations of IMGpedia resources

Resource	Location
LD IRI (example)	http://imgpedia.dcc.uchile.cl/resource/Rose_Amber_Flush_20070601.jpg
SPARQL endpoint	http://imgpedia.dcc.uchile.cl/sparql
Dump	http://imgpedia.dcc.uchile.cl/dumps/20170506/
VoID	http://imgpedia.dcc.uchile.cl/dumps/20170506/void.nt
Ontology	http://imgpedia.dcc.uchile.cl/ontology#
Issue Tracker	http://github.com/scferrada/imgpedia/issues
Datahub	http://datahub.io/dataset/imgpedia

Table 2: High-level statistics for IMGpedia

Name	Count	Description
Visual Entities	14,765,300	Entities about the images
Links to DBpedia Commons	14,765,300	Links to additional image metadata
Descriptors	44,295,900	Visual descriptors of the images
Similarity Links	442,959,000	Nearest neighbor relations between images
IRIs	502,020,200	Unique resource names
Links to DBpedia	12,683,423	Links to the resource about the article of the image
Triples	3,119,207,705	Number of triples present in the graph

images on Wikipedia Commons are attribution and share-alike[6]; non-derivative or non-commercial clauses are not permitted. Hence we release IMGpedia under an Open Database License (ODC-ODbL) license[7], which is an attribution/share-alike license specifically intended for databases. According to the 5-star model for Linked Open Data [2], IMGpedia is a 5-star dataset since it is an RDF graph that uses IRIs to identify its resources and provides links to other data sources (DBpedia and DBpedia Commons) to provide context. IMGpedia also has an issue tracker on GitHub, so users and collaborators can request features for future versions and report any problems they may find. The dataset is also registered at DataHub so researchers and other public can easily find and use it.

With respect to sustainability, given the large sizes of the dumps, we have yet to find a mirror host to replicate the data. However, internally, data are replicated on NAS storage and the source code is provided to replicate the dataset from the source Wikimedia Commons images. The first author has also secured funding to pursue a PhD on the topic, which will start this year; hence the dataset will be in active maintenance and development. With respect to updating the dataset, while building the original dataset was costly, we are planning to implement an incremental update where `rsync` is used to fetch new images; the descriptors for these images can then be computed, while only the k-nn similarity relations involving new images (potentially pruning old relations) need to be computed.

5 Use-Cases

We first provide some examples of queries that IMGpedia can answer.

First, we can query the visual similarity relations to find images that are similar by color, edges and/or intensity according to the nearest neighbor computation. In Listing 5 we show such a query, requesting the nearest neighbors of the image of Hopsten Marktplatz using the HOG descriptor (capturing visual similarity of edges). The results of this query are the images shown previously in Figure 1.

[6] https://commons.wikimedia.org/wiki/Commons:Licensing
[7] http://www.opendatacommons.org/licenses/odbl/

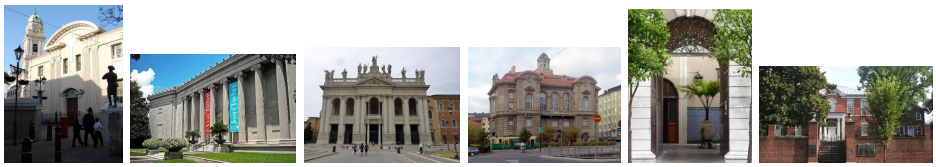
Listing 5: SPARQL Query for similar images to Hopsten Marktplatz

```
SELECT DISTINCT ?Target ?Distance WHERE {
        ?rel imo:sourceImage im:Hopsten_Marktplatz_3.jpg ;
        imo:usesDescriptorType imo:HOG ;
        imo:targetImage ?Target ;
        imo:distance ?Distance . }
ORDER BY ?Distance
```

Second, we can use federated SPARQL queries to perform visuo-semantic retrieval of images, combining visual similarity of images with semantic meta-data through links to DBpedia. In Listing 6, we show an example federated SPARQL query using the DBpedia SPARQL endpoint that takes the images from articles categorized as "*Roman Catholic cathedrals in Europe*" and looks for similar images from articles categorized as "*Museum*". In Figure 3, we show the retrieved images. To obtain more accurate results, SPARQL property paths can be used in order to include hierarchical categorizations, e.g. `dcterms:subject/skos:broader*` can be used in the first `SERVICE` clause to obtain all cathedrals that are labeled as a subcategory of European cathedral, such as French cathedral.

Listing 6: Query for images of museums similar to European Catholic cathedrals

```
SELECT DISTINCT ?urls ?urlt WHERE{
    SERVICE <http://dbpedia.org/sparql>{
        ?sres dcterms:subject dbc:Roman_Catholic_cathedrals_in_Europe . }
    ?source imo:appearsIn ?sres ;
            imo:similar ?target ;
            imo:fileURL ?urls .
    ?target imo:appearsIn ?tres ;
            imo:fileURL ?urlt .
    SERVICE <http://dbpedia.org/sparql>{
        ?tres dcterms:subject ?sub
        FILTER(CONTAINS(STR(?sub), "Museum"))}}
```



Cathedral of St. Mary and Museum of Fine Arts | Basilica of St. John L. and Nat. Hist. Museum of Helsinki | Cathedral of St. Mary and Dumbarton House Museum

Fig. 3: Results of Listing 6 query

With regards to *usage*, we released IMGpedia to the public on May 6th, 2017 and we keep a log of the SPARQL queries asked through the query endpoint, which at the time of writing (11 weeks later) contains 588 queries. However, we emphasize that IMGpedia was recently published. Our current plan is to further explore the potential of semantically-enhanced image retrieval that IMGpedia offers. The dataset also opens up a number of other use-cases. For example, one could consider combining the semantic information from DBpedia and the visual similarity information of IMGpedia to create a labeled dataset along the lines of ImageNet[8], but with variable levels of granularity (e.g., Catholic cathedral, cathedral, religious building, etc.). Another use-case would be to develop a clustering technique for

[8] http://www.image-net.org/

images based both on visual similarity and semantic context. We also believe that IMGpedia can compliment existing research works in the intersection of the Semantic Web and Multimedia, where it could provide a test-bed for works on media fragments [8,4], or on combining SPARQL with multimedia retrieval [5], etc.

6 Conclusions and Future Work

In this paper we have presented IMGpedia: a linked dataset that offers visual descriptors and similarity relations for the images of Wikimedia Commons; this dataset is also linked with DBpedia and DBpedia Commons to provide semantic context and further metadata. We described the construction of the dataset, the structure and provenance of the data, statistics of the dataset, and the supporting resources made available. Finally, we showed some examples of *visuo-semantic* queries enabled by the dataset and discussed potential use-cases.

There are many things that can be improved and added to IMGpedia. We will develop a web application to make IMGpedia more user-friendly, where users can ask queries intuitively (without needing SPARQL) and browse through results where images are displayed. We also plan to explore more modern visual descriptors that can help us to improve the current similarity relations between images, as well as defining similarity relations that combine descriptors.

Acknowledgments This work was supported by the Millennium Nucleus Center for Semantic Web Research, Grant № NC120004 and Fondecyt, Grant № 11140900. We would also like to thank Camila Faúndez for her assistance.

References

1. Addis, M., Allasia, W., Bailer, W., Boch, L., Gallo, F., Wright, R.: 100 million hours of audiovisual content: digital preservation and access in the PrestoPRIME project. In: INTL- DPIF. ACM (2010)
2. Berners-Lee, T.: Linked Data. W3C Design Issues, July 2006 (2010)
3. Ferrada, S., Bustos, B., Hogan, A.: IMGpedia: Enriching the web of data with image content analysis. In: AMW. CEUR (2016)
4. Kurz, T., Kosch, H.: Lifting media fragment uris to the next level. In: Linked Media Workshop (LIME–SemDev) at ESWC (2016)
5. Kurz, T., Schlegel, K., Kosch, H.: Enabling access to linked media with SPARQL-MM. In: World Wide Web (WWW). pp. 721–726 (2015)
6. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal (2014)
7. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISSAPP. pp. 331–340. INSTICC Press (2009)
8. Troncy, R., Mannens, E., Pfeiffer, S., Deursen, D.V.: Media Fragments URI 1.0. W3C Recommendation (2012)
9. Vaidya, G., Kontokostas, D., Knuth, M., Lehmann, J., Hellmann, S.: DBpedia Commons: Structured multimedia metadata from the Wikimedia Commons. In: The Semantic Web-ISWC 2015, pp. 281–289. Springer (2015)
10. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Comm. ACM 57, 78–85 (2014)