DBpedia Commons: Structured Multimedia Metadata from the Wikimedia Commons

Gaurav Vaidya[1], Dimitris Kontokostas[2], Magnus Knuth[3], Jens Lehmann[2], and Sebastian Hellmann[2]

[1] University of Colorado Boulder, United States of America
gaurav.vaidya@colorado.edu
[2] University of Leipzig, Computer Science, AKSW, Germany
{lastname}@informatik.uni-leipzig.de
[3] Hasso Plattner Institute, University of Potsdam, Germany
magnus.knuth@hpi.de

Abstract. The Wikimedia Commons is an online repository of over twenty-five million freely usable audio, video and still image files, including scanned books, historically significant photographs, animal recordings, illustrative figures and maps. Being volunteer-contributed, these media files have different amounts of descriptive metadata with varying degrees of accuracy. The DBpedia Information Extraction Framework is capable of parsing unstructured text into semi-structured data from Wikipedia and transforming it into RDF for general use, but so far it has only been used to extract encyclopedia-like content. In this paper, we describe the creation of the DBpedia Commons (DBc) dataset, which was achieved by an extension of the Extraction Framework to support knowledge extraction from Wikimedia Commons as a media repository. To our knowledge, this is the first complete RDFization of the Wikimedia Commons and the largest media metadata RDF database in the LOD cloud.

Keywords: Wikimedia Commons • DBpedia • Multimedia • RDF

1 Introduction

Wikipedia is the largest and most popular open-source encyclopedia project in the world, serving 20 billion page views to 400 million unique visitors each month[1]. Wikipedia has over 200 language editions, from the English Wikipedia (4.6 million articles) to the Tumbuka Wikipedia (177 articles). Every article contains metadata such as its page title, its list of contributors, the categories it belongs to, and the other articles it links to. Articles may also contain structured data, such as the latitude and longitude of geographically situated articles. Since 2007, the DBpedia project has been extracting this metadata and structured data and making it publicly available as RDF [**?**].

[1] http://reportcard.wmflabs.org/

Until recently, this extracted data had almost no information on the media files used to illustrate articles. While some media files are stored within a particular language edition of Wikipedia, over twenty-five million of them are located in a centralized repository known as the Wikimedia Commons[2]. The Wikimedia Commons acts as a media backend to all of Wikipedia; media files uploaded to it under an open-access license can be easily inserted into articles in any language. Metadata and structured data associated with the files are stored on the Wikimedia Commons' MediaWiki instance, in a format similar to that used by Wikipedia. This will likely be superceded by Wikidata, the Wikimedia Foundation's new structured data store, but this project is still under discussion[3]. We make this large and well maintained media resource accessible for semantic tools by extending the DBpedia Extraction Framework to read data from the Wikimedia Commons in addition to other Wikipedia language editions.

In this paper, we describe the dataset and the extraction process required to provide *DBpedia Commons* (DBc). We report on the extensions to the DBpedia Information Extraction Framework (DIEF) to support File pages, multiple languages on the same page, and proper Wikimedia Commons media URL construction. In addition we describe the ontological changes we made in the DBpedia ontology for annotating media files and the additional external vocabularies we chose for the media representation. To our knowledge, this is the first complete RDFization of the Wikimedia Commons and the largest media metadata RDF database in the LOD cloud.

2 Wikimedia Commons

The Wikimedia Commons follows many of the same conventions as Wikipedia itself: regular pages can contain textual content and embedded media files, pages may be placed in more than one category, and namespaces allow project and policy pages to be separated from content pages. Two main differences distinguish the Wikimedia Commons from Wikipedia: (a) Every Wikipedia edition is written entirely in a single language. The Wikimedia Commons is designed to be used by users of every language: where possible, page content is written in multiple languages so that it can be understood by all these users. (b) Most Wikipedia content is in its page content, i.e. its articles. Most Wikimedia Commons content is associated with individual files in the `File` namespace: thus, rather than describing a subject, as Wikipedia articles do, most Wikimedia Commons content describes a media file.

Our strategy for extracting data from Wikimedia Commons content therefore focused on extracting as much information as possible for each page from the `File` namespace. Since the DBpedia Extraction Framework can already extract content from MediaWiki archival dumps, we decided to modify it to support extracting content from archival dumps of the Wikimedia Commons[4]. Note that

[2] `http://commons.wikimedia.org/`

[3] See `https://commons.wikimedia.org/wiki/Commons:Wikidata` and `https://www.mediawiki.org/wiki/Multimedia/Structured_Data`.

[4] Such dumps are created monthly at `http://dumps.wikimedia.org/commonswiki/`.

this means the extraction framework never examines the media files directly; instead, it uses MediaWiki's dump format to infer statements about them.

3 Wikimedia Commons Extraction

We identified three kinds of data that we were interested in: (1) File Metadata, (2) Page Metadata, and (3) Content Metadata. File metadata describes the file that has been uploaded to the Wikimedia Commons, such as its encoding format, image dimensions and file size; these are stored in the backend database used by the MediaWiki software that runs the Wikipedia websites. Page metadata is stored for each MediaWiki page, including those that describe files. This includes the page title, the list of contributors and a history of changes. Finally, the content metadata is stored on the MediaWiki page itself: this includes a list of outgoing external and internal links, the list of categories the page belongs to as well as standard templates that allowed descriptions, sources, authority information and latitude and longitude of the subject of the page to be stored. This is often stored in a MediaWiki template, such as `{{Information}}`. After investigating the available file metadata[5], we decided to focus on Page and Content Metadata, as File metadata would required parsing the database dumps separately, necessitating much new software development. Unfortunately, this means that we cannot currently provide the dimensions or size of Wikimedia Commons files.

The DBpedia Information Extraction Framework (DIEF) has support for reading MediaWiki XML exports. DIEF was modified to read monthly backups of the Wikimedia Commons. Many of the extractors used to extract page metadata from Wikipedia [**?**] functioned flawlessly on the Wikimedia Commons dump, extracting titles, categories, authors and other page and content metadata and transforming them into RDF with only minor changes. Four new File Extractors targeting Wikimedia Commons-specific information were developed (Section 3.1). The DBpedia mapping-based extractor was adapted to work on Wikimedia Commons media and creator pages (Section 3.2). We used this extractor to obtain licensing information through the mapping-based extraction.

IRI Scheme . By using the `http://commons.dbpedia.org` domain and following the existing naming strategy of DBpedia, the DBc resources are published under the `http://commons.dbpedia.org/resource/` namespace. For example, `http://commons.dbpedia.org/resource/File:DBpediaLogo.svg`.

3.1 Media Extractors

FileTypeExtractor. The FileTypeExtractor guesses the media MIME type by examining its file extension, and uses a preconfigured index to assign both the direct type and the transitive closure of the direct type using `rdf:type` (cf. Figure 1 and Section 4). 354,873 files could not be identified by their file extension; an expansion of the preconfigured index will be necessary to include them. The direct type is also linked with `dct:type`. `dct:format` captures the MIME type according

[5] `https://www.mediawiki.org/wiki/Manual:Image_table`.

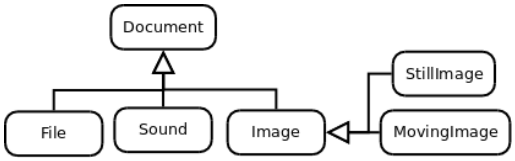

Fig. 1. Hierarchy of main Document classes

to RFC 6838[6]. The file extension is directly queryable with `dbo:fileExtension`. In addition, we provide `dbo:fileURL` for access to the final media URL and `dbo:thumbnail` and `foaf:depiction` for still images. This extractor also provides links to the image itself by using the special page `Special:FilePath`, which provides redirects to the image file. A sample output of this extractor is:

```
@prefix db-com: <http://commons.dbpedia.org/resource/File:>.
@prefix commons-path: <http://commons.wikimedia.org/wiki/Special:FilePath/>.
db-com:DBpediaLogo.svg a dbo:StillImage, dbo:Image, foaf:Image, dbo:File,
          dbo:Document, foaf:Document, schema:CreativeWork, dbo:Work ;
 dct:type dbo:StillImage
 dct:format "image/svg+xml";
 dbo:fileExtension "svg" ;
 dbo:fileURL commons-path:DBpediaLogo.svg;
 dbo:thumbnail commons-path:DBpediaLogo.svg?width=300;
 foaf:depiction commons-path:DBpediaLogo.svg.
```

GalleryExtractor. The Wikimedia Commons and Wikipedia both support galleries that make it easy to display a series of related images in a compact format[7]. On Wikimedia Commons, this may be used to display a representative set of images about a single topic, such as the page for Colorado[8]. The GalleryExtractor identifies galleries embedded in pages, extracts the list of individual media items, and links them to the page resource with `dbo:galleryItem`.

```
db-com:Colorado dbo:galleryItem
  db-com:2006_CO_Proof.png, db-com:Anasazi_Heritage_Center.jpg,
  db-com:Bearlakeinspring2.jpg, db-com:Beol_court25.jpg .
```

ImageAnnotationExtraction. The Wikimedia Commons may contain additional annotations for parts of a still image using `{{ImageNote}}` and related templates. The annotations mark a rectangular region within the image and provide a description text in MediaWiki syntax, which may in turn contain hyperlinks to other resources. We extract this information using the W3C *Media Fragments* [**?**] vocabulary. The annotated box is identified by a separate IRI that is linked to the original resource through `dbo:hasAnnotation`. As seen in the example below, the new IRI is based on the original resource by suffixing the the coordinates of the part of the image being annotated as well as its total width and height, extracted from the `{{ImageNote}}` template, which is necessary in case the original image needs to be scaled.

[6] `http://tools.ietf.org/html/rfc6838`.
[7] `http://en.wikipedia.org/wiki/Help:Gallery_tag`
[8] `http://commons.wikimedia.org/wiki/Colorado`

```
@prefix ann: <http://commons.wikimedia.org/wiki/Special:FilePath/Yes_concert.jpg?width
     =1514&height=1024#xywh=pixel:539,380,110,108>.
db-com:Yes_concert.jpg dbo:hasAnnotation ann: .
ann: "Jon Anderson"@en .
```

CommonsKMLExtractor. Keyhole Markup Language or KML[9] is an XML format used to describe map overlays, allowing images of maps to be precisely georeferenced to a location on the planet. The CommonsKMLExtractor extracts the KML data from Wikimedia Commons and stores them as an `rdfs:XMLLiteral` value of the `dbo:hasKMLData` property.

```
db-com:Yellowstone_1871b.jpg dbo:hasKMLData """
  <?xml version=1.0 encoding=UTF-8?>
  <kml xmlns="http://earth.google.com/kml/2.2">
   <GroundOverlay> <!-- KML data --> </GroundOverlay></kml>"""^^rdfs:XMLLiteral .
```

3.2 Infobox to Ontology Mappings using the Mapping Extractor

The DBpedia Information Extraction Framework (DIEF) has a sophisticated system for extracting infoboxes from Wikipedia articles. An 'infobox' is a special template that stores semi-structured data about the subject of an article. For example, `{{Infobox person}}` may record the birth date and location of the person, while `{{Infobox book}}` might record the ISBN and OCLC number of the book. The DBpedia Mapping Extractor allows contributors to the DBpedia Mappings Wiki[10] to describe how template properties map to properties on the DBpedia ontology [**?**, Sec. 2.4].

A similar set of templates provides information on the Wikimedia Commons; for example, the `{{Location}}` template stores the location that is the subject of a media file, such a building being photographed or a city being mapped. A new DBpedia mapping namespace for the Wikimedia Commons was created[11] and DIEF was refactored to extract templates from media file and creator pages and use DBpedia mappings to convert them to RDF statements.

License Extraction. Licenses are encoded in the Wikimedia Commons as templates, e.g. the template `{{cc-by-sa}}` present on a `File` page indicates that the media file has been licensed under the Creative Commons BY-SA license. We used the Mapping Extractor described above to map each template to a URL describing the license, such as https://creativecommons.org/licenses/by-sa/2.0/. However, it is a common practice on the Wikimedia Commons to nest and embed multiple licenses together: for example, the template instruction `{{self|cc-by-sa-4.0}}` indicates that this file was created by the uploader ('self') who has licensed it under a Creative Commons CC-BY 4.0 license. Since nested or wrapped templates are not currently supported in DIEF, we added a preprocessing extraction step to unwrap license templates specifically to make all license mappings identifiable to the Mapping Extractor.

[9] `https://developers.google.com/kml/documentation/`
[10] `http://mappings.dbpedia.org/`
[11] `http://mappings.dbpedia.org/index.php/Mapping_commons`

Table 1. Description of the DBc datasets

Title	Triples	Description
Labels	29,203,989	Labels for resources
Provenance	272,079,712	Provenance information (pageIDs, revisionIDs)
SKOS	94,701,942	SKOS hierarchy based on the category hierarchy
Geo data	18,400,376	Geo coordinates for the media files
File Information	414,118,159	File metadata
Annotations	721,609	Image annotations
Galleries	2,750,063	Image galleries
Types	111,718,049	Resource types
KML	151	KML data
Mappings	95,733,427	Mapped infobox data
Infobox	87,846,935	Unmapped Infobox data
Interlanguage links	4,032,943	Links to other DBpedia editions
Internal links	116,807,248	Internal links to other Wikimedia Commons pages
External links	17,319,980	Links to external resources
Metrics	58,407,978	Article metadata (in/out degree, page size)
Templates	77,220,130	Template metadata and usage

```
1 db-com:DBpediaLogo.svg dbo:license <http://creativecommons.org/publicdomain/mark/1.0/>
```

4 Dataset

A general overview of the datasets provided by DBc is provided in Table 1, where each row provides a summary of one or more similar datasets. A total of 1.4 billion RDF triples were inferred from the Wikimedia Commons dump prepared in January 2015, describing almost 30 million unique IRIs. A diagram for the new classes we introduced for Wikimedia Commons media files is depicted in Figure 1: `dbo:Document` has the subclasses `dbo:File`, `dbo:Sound`, and `dbo:Image`. A `dbo:Image` can be a `dbo:StillImage` (e.g. picture) or a `dbo:MovingImage` (e.g. video). DBc mostly consists of still images (Table 2) with JPEG as the most popular format (Table 4). Table 3 provides the most frequent properties in DBc while Table 5 lists the most common media licenses. One of the largest datasets are the *mappings* (95.7M triples) which is based on the infobox to ontology mappings (Section 3.2), and so include the license information. The authors, with contributions from the DBpedia community, invested significant effort to ensure that 90% of all occurrences of infobox templates and 78% of all template parameters on the Wikimedia Commons have either been mapped to an RDF entity in the DBpedia ontology or have been determined to have no structured data.[12]

Access and Sustainability. DBpedia Commons is part of the official DBpedia knowledge infrastructure and is published through the regular releases of DBpedia along with the rest of the DBpedia language editions. The first DBpedia release that included this dataset is *DBpedia 2014*[13]. DBpedia is a pioneer in adopting and creating best practices for Linked Data and RDF publishing. Thus, being incorporated into the DBpedia publishing workflow guarantees: (a) long-term availability through the DBpedia Association and the Leipzig Computer

[12] `http://mappings.dbpedia.org/server/statistics/commons/`, as of April 25, 2015
[13] `http://downloads.dbpedia.org/2014/commons`

Table 2. Top classes

Count	Class
25,061,835	dbo:StillImage
611,288	dbo:Artwork
90,011	dbo:Agent
49,821	dbo:MovingImage
19,126	dbo:Person

Table 3. Top properties

Count	Property
73,438,813	dct:subject
43,209,414	dbo:license
29,201,812	dce:language
24,496,724	dbo:fileURL
24,496,706	dbo:fileExtension

Table 4. Top MIME types.

Count	MIME type
20,880,240	image/jpeg
1,457,652	image/png
878,073	image/svg+xml
455,947	image/tiff
246,149	application/pdf

Table 5. Top licenses

Count	License
7,433,235	CC-by-sa v3.0
4,096,951	CC-pd v1.0
3,704,043	GNU-fdl v1.2
3,681,840	GNU-fdl
2,116,411	CC-by-sa v2.0

Center long-term hosting platform and (b) a shared codebase with the DBpedia Information Extraction Framework. Besides the stable dump availability we created `http://commons.dbpedia.org` for the provision of a Linked Data interface [**?**], a SPARQL Endpoint and more frequent dataset updates . The dataset is registered in DataHub[14]and provides machine readable metadata as void[15] and DataID[16] [**?**]. Since the project is now part of the official DBpedia Information Extraction Framework, our dataset reuses the existing user and developer support infrastructure, e.g. the general discussion and developer list as well as the DBpedia issue tracker for submitting bugs.

5 Use cases

In the following, we provide several existing or possible use cases of the DBc dataset.

Galleries, Libraries, Archives and Museums. Collectively known as GLAMs, such institutions hold large repositories of documents, photographs, recordings and artifacts. Several have made large contributions of media to the Wikimedia Commons, such as the 128,000 images donated by the National Archives and Records Administration of the United States, containing rich metadata stored using the {{`NARA-image-full`}} template. By mapping parameters in this template to properties in the DBpedia Ontology, we were able to quickly obtain descriptions, authors, notes and local identifiers in RDF for this media[17]. DBc provides a community-edited source of structured data in RDF that can exist in parallel with any structured data being published directly by a GLAM.

Image Recognition Algorithms. Over 96,000 images on the Wikimedia Commons have embedded annotations[18]. By making the coordinates of these annotations

[14] `http://datahub.io/dataset/dbpedia-commons`
[15] `http://commons.dbpedia.org/void.ttl`
[16] `http://dbpedia.s16a.org/commons.dbpedia.org/20150110/dataid.ttl`
[17] See `http://commons.dbpedia.org/resource/File:Douglas_MacArthur_lands_Leyte1.jpg` for an example
[18] `https://commons.wikimedia.org/wiki/Category:Images_with_annotations`

available through our Image annotations dataset, we provide a training dataset that can be used to teach machine-learning algorithms to identify annotations that may be of interest to Wikimedia Commons editors.

License Extraction. Media uploaded to the Wikimedia Commons must either be in the public domain or licensed under an open-access licenses, but individual language Wikipedias may allow users to upload unlicensed images as long as they have a fair-use rationale for them. This means that not all media files embedded in Wikipedia articles may be freely reused. Furthermore, different open-access licenses have different requirements for re-use: some allow any re-use as long as the original creator is cited, while others require any derivative works to carry the same license as the original. Since licenses on the Wikimedia Commons are encoded by licensing template, we were able to use the Mapping Extractor (Section 3.2) to provide license URLs for several million media files. This allows licensing conditions for many Wikimedia Commons media files to be determined automatically. In particular, the German National Library contacted some of the authors specifically for this metadata: they included Wikimedia Commons images in their Linked Data interface and were interested in displaying the license information directly there using our license dataset. This integration is not yet deployed.

6 Conclusions and future work

We introduced DBpedia Commons, to our knowledge the first large-scale knowledge extraction from Wikimedia Commons. We present the adaptations and additions made to the DBpedia Information Extraction Framework to facilitate the correct extraction of media files and their metadata, including license information. The dataset contains 1.4 billion RDF triples that provide file metadata, provenance, descriptions, and license information.

Acknowledgements. This work was funded by the Google Summer of Code 2014 program and by grants from the EU's 7th & H2020 Programmes for projects ALIGNED (GA 644055) and GeoKnow (GA 318159).