

画像を主とするデジタルレコードに対する カタログ付与のためのデータセットの構築・公開

鈴木 稔規^{1,a)} 永溝 孝太^{1,b)} 池田 大輔^{1,c)}

概要：デジタルトランスフォーメーション（DX）が推進されている。DX 推進に伴い、多種多様なデジタル化されたレコードの増加が予想される。デジタルレコードを即座にアクセス出来る様に、適切に管理することが求められる。カタログとは、電子図書館におけるデジタルレコードを参照・管理するためのメタデータである。カタログは、レコードの対象分野について予め作成されたシソーラスから選択される。一方で、シソーラスが事前に用意されないレコードへのカタログ付与は喫緊の課題である。本研究では、画像とキーワードからなるレコードに、シソーラスの代わりとしてオンライン百科事典ウィキペディアのエントリを付与するシステムの開発・評価用データセットを構築した。データセットの構築及び利用促進のための公開活動について報告する。

キーワード：画像データ、カタログ、メタデータ、ウィキペディア

Developing and Opening a Dataset for Cataloging Image-based Digital Archives

Abstract: The Digital Transformation (DX) is being promoted by Japanese government as their major governmental policy. As DX is penetrated, various records in digital formats are expected to be multiplied. It is becoming more important to organize such digital records to be ready for access easily and immediately. Cataloging, which assigns metadata (catalog) to records, is an effective way for record management. While a catalog is selected from a predefined thesaurus for a collection, cataloging a record, which do not have any thesauri, is requested to fulfill future demands. In this study, we developed a dataset to develop and evaluate a system which assign Wikipedia entries instead of thesaurus to a record consisting of an image and keywords. We report the development of the dataset and our activity of promoting the use of it.

Keywords: image data, record management, catalogue, meta data, wikipedia

1. 初めに

日本では、デジタルトランスフォーメーション（DX）が推進されており、IT の浸透がより深化していくことが予想される。データの利活用に関して、特にコロナ禍において、厚労省のコロナウィルス感染症に関するオープンデータ^{*1}を始め、多種多様なデータの公開・共有化が図られ、感染症の状況分析に役立てられている。データ公開が進ん

だ社会では、分析者などデータを必要とする人が、所望のデータに素早くアクセスし、活用出来る環境を整備することは、これまで以上に重要になる。

データ活用の基盤として、適切なメタデータ（カタログと呼ぶ）をデジタルレコードに付与することで、参照的にアクセスが可能となり、データの検索性が高まる。特に、データがテキストではない、画像などにおいて非常に重要となる。長年、画像へのタグ付与は、コンペ型ワークショップを中心に研究がされてきている [9]。ただ、画像へのタグ付与は、予め決められた語彙集合からタグを選択するため、検索タグの重複は考慮がされていない。画像対象が広がる応用では、検索性の低下が問題となる。

本研究では、画像付与タグの重複を排除することで、画

¹ 九州大学

Kyushu University, Fukuoka 819-0395, Japan

a) tokinori@inf.kyushu-u.ac.jp

b) nagamizo.koto.749@s.kyushu-u.ac.jp

c) ikeda.daisuke.899@m.kyushu-u.ac.jp

*1 <https://www.mhlw.go.jp/stf/covid-19/open-data.html>

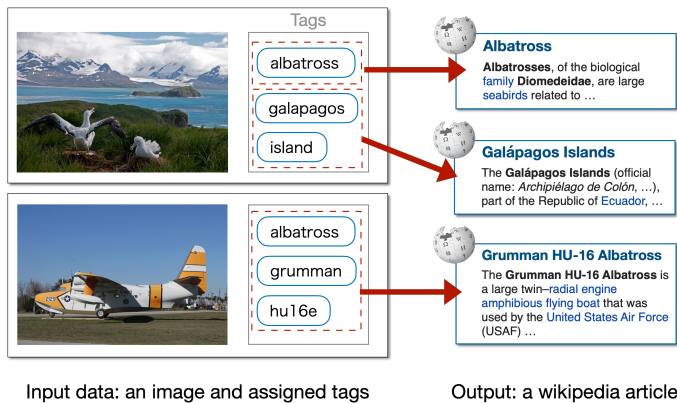


図 1 画像タグ曖昧性解消課題の入力及び出力例。

像データに対するカタログとするシステム構築を目標とする。タグの重複例として、図 1 では、同名タグ ‘albatross’ が、同図上の鳥の写真及び、下の飛行機の写真が付与される。このような重複タグを、対応するシソーラスに対応づけることで曖昧性を排除し、カタログとする。具体的には、画像とタグのペアを入力として、タグをウィキペディアの項目に対応付けをする課題（タグ曖昧性解消）を扱う。図 1 に、タグ曖昧性解消の入力と対応ウィキペディア記事を例示している。二つの異なる画像内容に応じて、同名タグ ‘albatross’ を、アホウドリ（‘Albatross’）や飛行機のモデル（‘Grumman HU-16 Albatross’）に対応付けをし分ける。

本報告では、タグ曖昧性解消システム開発及び評価のために構築したデータセットを紹介する。画像共有サイト Flickr ^{*2}から収集した 420 件の画像、関連タグ 2,464 件について正解ウィキペディア記事を、人手で判定付与した。本データセットは、Mendeley Data で公開 ^{*3}されており、誰でも無料で利用可能である。このデータセットは、タグ曖昧性解消システムの構築・評価に加えて、上述した飛行機モデルなど詳細な対象が含まれているため、より細かい分類クラスを対象にした画像分類システムの評価にも用いることが可能である。

本報告の構成は、次の通りである。第 2 節で関連研究について簡単にまとめる。第 3 節では、データセットの作成及び構築したデータセットの評価を行う。第 4 節では、タグ曖昧性解消手法のベースライン手法と課題の評価方法について説明する。第 5 節では、データセットの内容（データ及びプログラム）の紹介をする。第 6 節で、本報告をまとめる。

2. 関連研究

デジタルレコードに対するカタログ付与は、音楽レコー

ド [7] 及びデジタルライブラリ [3], [6] において、シソーラスの代替としてリンクトデータを活用する研究がされている。これらの研究では、管理するコレクションに対するシソーラス構築にかかるコスト削減を目的としている。画像をコンテンツとするデジタルレコードについては、検索タグ付与に関する研究が、ワークショップのテーマに設定されて研究が行われる [9] など、主流である。近年では、畳み込みニューラルネットワーク中間層から特徴抽出 [4] や画像からテキスト特徴量への変換手法 [10] など深層学習を活用したタグ付与手法が提案されている。検索タグ付与では、重複タグが考慮されていなく、レコード管理への即座の活用は出来ない。本研究では、検索タグを曖昧性を解消したカタログとして活用することを目的とする。

3. データセットの構築

本節では、データセットの構築について、対象データの収集及び正解記事の判定について説明する。

3.1 画像と関連タグの収集

画像データは、Flickr から、動物名のタグが付与される画像を対象として収集した。タグ曖昧性解消課題評価のために、タグ名に曖昧性がある、動物名称のタグを対象とした。例えば、動物名は、商品名やスポーツチームのニックネームなど動物以外の写真に用いれる。動物名タグには、一般的な動物の名称 ^{*4}から、無作為に 14 種を設定した。図 3 の横軸に、動物名タグを示す。その他、画像に付与される全タグを収集した。写真の内容に無関係なタグ、‘nikond5’、‘500mmf4’などは除外した。

3.2 正解記事のアノテーション

画像タグに対する適切なウィキペディア記事を人手により判定付与をした。正解判定は、作業者が画像タグに対する正解ページをウィキペディアから検索をする形式で行った。本節では、正解判定作業とアノテーション結果について説明する。

作業者は図 2 で示す作業用スプレッドシートを用いて正解判定を進めた。作業シートに表示される画像とその関連タグに対して、一単語もしくは複数単語について、画像との関係性を判断し対応する英語版ウィキペディア記事を正解として記入する。

図 2 中のタグ ‘South’、‘America’ を例に説明する。作業者は、これらのタグと表示されている画像（アルパカ）から対応する記事の判定を行う。タグを検索質問にするもしくは、手がかりとしてウィキペディアの対応記事を検索する。検索過程では、作業者は検索質問の修正及び検索を繰り返して候補を探す。タグに対する見つけた候補につい

^{*2} <https://flickr.com>

^{*3} <https://data.mendeley.com/datasets/msyc6mzvgh/1>

^{*4} https://en.wikipedia.org/wiki/List_of_animal_names

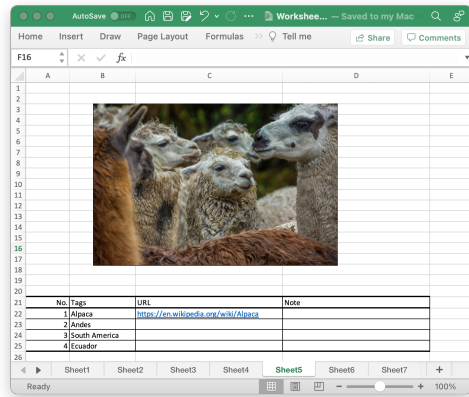


図 2 アノテーション作業のスクリーンショット。

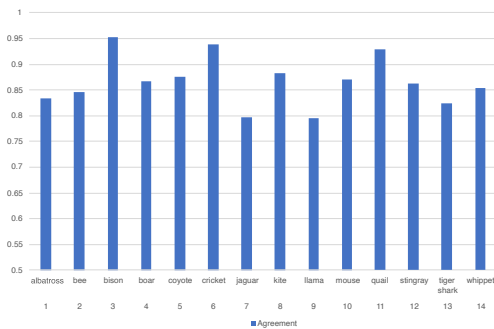


図 3 各タグにおける作業者の一緻率。

て、関連性を判断して、対応記事であれば作業シートに記事の URL を記入する。対応する記事がないタグについては、対応なしと記述をする。上述の正解判定の作業を、アノテーション指示者との数回の練習後に実施した。

正解判定では一つのタグに付き二名の作業者を割り当てた。動物名タグ毎の作業者間一致率を、図 3 に示す。タグ間のマクロ平均は、0.87 であった。また、作業者間の一致についてのカッパ係数は、0.9 を超えた。

4. タグ曖昧性解消のベースライン手法及び評価

4.1 ベースライン手法

タグ曖昧性解消に対してのベースラインとして三手法 (BM25, 手法 1, 手法 2) の実装を本データセットで提供する。各手法のプログラム及び出力はデータセットに含まれる。

BM25 は手法は、タグによるキーワード検索手法である。タグを検索質問として、BM25 のスコア計算 [8] によってウィキペディア記事を検索する。手法の実装には、Apache Solr 検索エンジン *5 を用いた。データセットでは、BM25 のパラメータ設定 $k_1 = 1.2, b = 0.75$ を用いた結果を格納する。

手法 1 及び手法 2 は、対象タグについて BM25 の検索結

果を候補にして、再順位付けをして改善する手法である。手法 1 では、画像に付与される全タグについて、word2vec [5] を用いた文脈情報を類似度計算に用いる。手法 2 では、全タグに加えて、画像分類 [2] による分類ラベルを類似度計算に用いる。手法 1, 2 では、候補ウィキペディア T 、画像に付与されたタグ Q 、画像分類による分類ラベル L から次式により類似度を計算する。

$$\text{sim}(T, Q, L) = \alpha \frac{1}{|T|} \sum_{w_q \in Q} \text{sim}(T, w_q) + (1 - \alpha) \sum_{w_l \in L} \text{sim}(T, w_l),$$

T はウィキペディア記事テキストで、本データセットでは記事の見出しとした。 w_q 及び w_l は、それぞれタグ中の一単語、画像ラベル中の一単語を表す。 α は、タグ及び画像ラベルの重みづけパラメータである。本データセットでは、手法 2 は、 $\alpha = 0.5$ の設定に基づく。手法 1 は、上式の左辺のみから類似度計算がされる ($\alpha = 1$)。単語間の類似度 sim は、次式により計算する。

$$\text{sim}(w_1, w_2) = \cos(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1^T \mathbf{w}_2}{\|\mathbf{w}_1\| \|\mathbf{w}_2\|},$$

ここで、 $\mathbf{w}$ は単語 w のベクトル表現であり、両単語ベクトルのコサインを計算する。

4.2 評価方法

タグ曖昧性解消課題における出力は、タグに対するウィキペディア候補を順位付きリストである。評価尺度として、Mean Reciprocal Rank (MRR), Precision@1 (P@1), Precision@10 を用いる。一つ目に、MRR は次式により計算する。

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q},$$

Q は、画像に付与されるタグ q の集合、 rank_q は、出力された順位付きリスト中の正解記事の順位である。 P@1 及び P@10 は、それぞれ、順位付きリスト、第 1 位及び第 10 位までに正解記事が含まれる課題の割合である。

5. データセットの紹介

本節では、第 1 節で紹介したタグ曖昧性解消課題に対して、構築データセットに含まれる構成データについて説明する。

本データセットは、写真共有サイト Flickr から収集した 420 件の画像及び、合計 2,464 組の関連タグと対応ウィキペディア記事から構成される。データセットは、図 4 に示す通り、直下のディレクトリに、三つのディレクトリ (data, img, results) と一つのプログラムファイルを主とする。各要素について説明する。

*5 <https://solr.apache.org/>

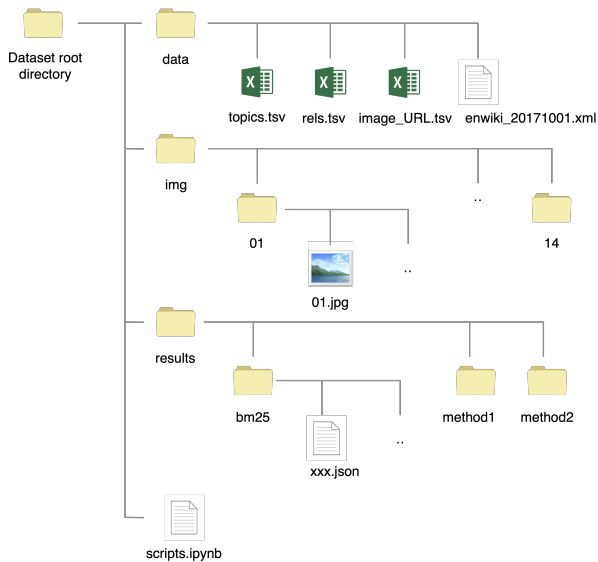


図 4 データセットのディレクトリ構造.

5.1 データディレクトリ

データディレクトリ (data) は, 三つの表ファイルと XML ファイルを含む. 表 (topics.tsv, rels.tsv, image_URL.tsv) は, タブ区切り形式のファイルである. topics.tsv は, 表 1 に示す列から構成され, 曖昧性解消の対象タグ及び画像と関連タグを保持するファイルである. この表の一行を一課題として扱う. rels.tsv は, 課題に対する正解ウィキペディアの情報をまとめた正解判定である. 表 2 に示す列, 課題番号, ウィキペディアの記事名と URL から構成される. image_URL.tsv ファイルは, 画像ファイルと画像がアップロードされる Flickr の URL を保持し, 表 3 に示す列からなる表である.

XML ファイル (enwiki_20171001.xml) は, 英語版ウィキペディアダンプ記事, 1,210,453 件の見出し, 本文を保持する. 2017 年 10 月時点のウィキペディアダンプデータ^{*6}から, Python ライブラリ, WikiExtractor^{*7}を用いて取得した.

5.2 画像ディレクトリ

画像ディレクトリ (img) は, 画像用のディレクトリである. スクリプトファイル, scripts.ipynb (第 5.4 節参照) のダウンロードプログラムによって, 420 件の画像が格納されるプレースホルダである. img ディレクトリ構造は, 'xx/yy.jpg' となり, カテゴリを表すサブディレクトリ (xx) の下に, 該当画像 (yy) を保持する.

本データセットは, 大規模画像データベース ImageNet [1] と同様に, 著作権の観点から画像ファイル自体の提供ではなく, URL 及びダウンロードプログラムを用意する.

表 1 topic.tsv ファイル 表の説明

列名	データ形式	説明
Topic	整数値	課題の識別番号
Query_label	文字列	曖昧性解消対象のタグ
All_labels	文字列のリスト	課題に対応する画像の全タグ
Image_file	文字列	課題に対応する画像ファイル名

表 2 rels.tsv ファイル 表の説明

列名	データ形式	説明
Topic	整数値	課題の識別番号
Title	文字列	正解ウィキペディア記事のタイトル
URL	文字列	正解ウィキペディア記事の URL

表 3 image_URL.tsv ファイル 表の説明

列名	説明	データ形式
Image_file	文字列	画像ファイル名
URL	文字列	画像ファイルの URL

5.3 実験結果ディレクトリ

実験結果ディレクトリ (results) は, 実験結果の JSON ファイルを格納する. 図 4 に示す通り, 性能の比較評価及び評価プログラムのデモ用にベースライン手法 (bm25, method1, method2) の結果が格納されている. 各 JSON ファイルは, 一課題に対応する結果になる. 例えば, 図中の結果ファイル 'xxx.json' は, 表 1 の Topic (課題番号) 'xxx' に対応する.

5.4 プログラムファイル

タグ曖昧性解消課題を再実験するためのプログラム群を, Jupyter Notebook^{*8} ファイル (scripts.ipynb) にまとめる. Jupyter Notebook は統合開発環境であり, 環境下に scripts.ipynb を配置することで, プログラムの改変及び実行の作業を容易に行うことが出来る. 図 5 は, scripts.ipynb のスクリーンショットであり, 各プログラムが実行可能なセルに記述される. このファイルには, 画像ファイルのダウンロード, ベースライン手法の実行, 実験結果の評価に対応する Python プログラムが用意される.

6. 終わりに

本論文では, タグ曖昧性解消課題のシステム構築及び評価用データセットを構築した. また正解判定作業者の一致率などの品質評価及び, ベースライン手法について説明し, データセットのファイル等のコンテンツを紹介した. 容易に再実験が出来るように, Jupyter Notebook 形式の実験プログラムを提供する. 利用促進のために, 本データセッ

^{*6} <https://archive.org/download/enwiki-20171001>

^{*7} <https://github.com/attardi/wikiextractor>

^{*8} <https://jupyter.org>

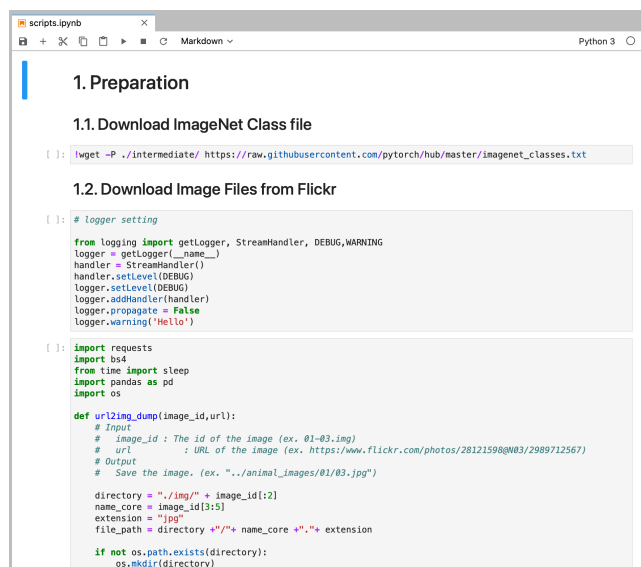


図 5 Jupyter Notebook スクリーンショット.

トは, Mendeley Data にて公開している. 今後の課題としては, データセットの観点では, タグ曖昧性解消システムの性能向上に寄与するように, 利用や認知状況の分析・調査, それらを反映した認知向上に向けた活動が挙げられる.

参考文献

- [1] Deng, J., et al.: ImageNet: a large-scale hierarchical image database, *Proc. of the 2009 IEEE Computer Vision and Pattern Recognition*, pp. 248—255, (2009).
- [2] Krizhevsky, A., Sutskever, I. and Hinton, G. E., ImageNet classification with deep convolutional neural networks, *Proc. of the 25th International Conference on Neural Information Processing Systems*, pp. 1097—1105, (2012).
- [3] Kroegen, A.: The road to BBIBFRAME: the evolution of the idea of bibliographic transition into a post-MARC future, *Cataloging & Classification Quarterly* 51(8), 873—890, (2013).
- [4] Li, Z., et al.: Collaborating CNN and SVM for automatic image annotation, *Proc. of the 2019 ACM International Conference on Multimedia Retrieval*, pp. 63—67, (2019).
- [5] Mikolov, T., et al.: Distributed representation of words and phrases and their compositionality, *Proc. of the 26th International Conference on Neural Information Processing Systems*, pp. 3111—3119, (2013).
- [6] Nurmikko-Fuller, T., et al.: A comparative analysis of bibliographic ontologies: implications for digital humanities, *Digital Humanities 2016: Conference Abstracts*, pp. 639—642, (2016).
- [7] Page, K. R., et al.: Realizing a Layered digital library: exploration and analysis of the live music archive through linked data, *Proc. of the 17th ACM/IEEE Joint Conference on Digital Libraries*, pp. 89—98, (2017).
- [8] Robertson, S. E., et al.: Okapi at TREC-3, *Proc. of the third Text Retrieval Conference*, pp. 109—126, (1996).
- [9] Villegas, M. and Paredes, R.: Overview of the ImageCLEF 2014 scalable concept image annotation task, *Proc. of the fifth Conference of the CLEF Initiative*, pp. 308—328, (2014).

- [10] 鈴木 規規, 池田 大輔, 画像アノテーション課題からテキスト分類課題へ～深層学習を用いたモダリティ変換の試み～, 火の国情報シンポジウム 2020, (2020).