

画像データの索引付け

及川 昭文[†]

昨今は画像データを含むデータベースがいろいろな分野で増えてきている。また、画像そのものを検索の対象としている類のデータベースも多くなってきている。画像を適切に検索するためには、画像に対して索引語を付与する必要がある。これまでに作成した3つの画像データベースを例に、画像データの索引付けについて報告する。

Indexing to Images

Akifumi Oikawa[†]

In recent years databases including image data is increasing in various fields. In addition, many are getting that kind of database search for the images themselves. To find the proper image, we need to give a word or words to each image for the index. In this report indexing to images is explained through three image databases.

1. はじめに

画像データを含むデータベースが増えてきているが、一般的に画像を探すには(1)付与されているテキスト情報を手がかりにする場合、(2)画像を視覚的に分析し、その特徴を手がかりにする場合に大別される。後者については、すでに数多くの研究が「人文科学とコンピュータ」研究会でも報告されている。[1-5]しかし、いまだに本格的な実用に供された例はみあたらない。試行錯誤が続いている状態にあるといってもよいが、実用化にいたらない理由としては、以下のようなことが考えられる。

- 1) どのような観点から視覚的分析を行うかは、それぞれの研究者で異なっており、その結果が不特定多数の利用者の共感を得ることはきわめて困難である。
- 2) どのような視覚的分析を行うにしろ、データベースを作成する前の画像処理には時間と手間がかかり、データベースの作成コストが非常に高くなる。
- 3) たとえば浮世絵、風景写真といった限定された対象であれば可能であっても、博物館収蔵品のように多種多様なものが混在しているものすべてを対象に、同じ視覚的分析を行うことはほとんど不可能である。

したがって、実用化ということを第一の目標とするならば画像に対してキーワードを付与することがもっとも現実的である。本報告では、これまでに作成した3つの画像データベースについてどのようにして画像の索引化を行ったかについて述べる。

2. 画像データベースの作成

データベースはその性格から大きく2つに、すなわち(1)図書館におけるOPAC(Online Public Access Catalog)や博物館における収蔵品データベースのような参照用データベースと、(2)分析の元データとして利用するデータベースに大別される。もちろん、両者の間に明確な境界があるわけではないし、参照用データベースを分析の対象とすることも、分析用データベースを参照用として利用することも十分に可能である。ただ、一般的に前者は後者に比べて標準化ということを強く意識して項目の構成が検討されることになる。

今回の3つのデータベースはいずれも分析用として利用できることを目指して作成したものであるが、分析用とする場合にはデータベースだけでなく、そのプラットフォームである検索システムにも分析のための機能を備えている必要がある。[6,7]

[†] 総合研究大学院大学

The Graduate University for Advanced Studies

今回作成した3つのデータベースは、さまざまな分析が出来ることを念頭に置いて、項目の構成、キーワードの付与を検討した。以下、それぞれのデータベースの概要について説明する。

1) 鶏コレクション

本データベースは総研大の学内共同研究プロジェクト「ドメスティケーション・モデルの構築—博物学の視点から」（研究代表者：秋篠宮文仁・総研大特別研究員）における議論の中から提案されたものである。このプロジェクトは、「家畜は人間によってその生殖を管理されるようになり、さまざまな体型的、生理的、色彩の変異を起こし、それぞれがもつ特性を活かす形で人と共存するに至っている動物である。そのような家畜という文化的生き物を研究するには、生物学的側面と文化学的側面の双方から考察していくことが必要である」という視点で研究が進められてきた。古代から人と共生してきた「鶏」は、社会の中で家畜としてだけでなく、民芸品、玩具、置物、信仰用品、実用品などさまざまな形



図1：鶏コレクションの例

で人とかかわってきている。このかかわりを探求するためには、これらの民芸品などの実態を正確に記録しておく必要がある。また、すでに社会から消滅してしまったものも少なくなく、まだ現資料が存在しているうちに「情報」として記録することも重要な課題としてある。生き物でない「鶏」が、異なる民族、社会、文化の中でどのような形で現れているかを精査することによって、ドメスティケーション過程を推測する一つの方法である「鶏と人の関係」の観察ができるとともに、それぞれの民族、社会、文化の一側面に言及することができる。また、民芸品や置物の意匠、デザインを比較することによって、民族性、地域性を論ずることも可能である。このように本データベースは、さまざまな視点からの活用が可能であり、学術的にも文化的にも意味のあるデータベースということができる。

2) ブックカバー

ほとんどの図書館では、購入した本を本棚に配架する際、ブックカバーについては本から外し廃棄している。著名なイラストレーターやブックデザイナーがデザインした、美術的にも十分価値のあると思われるブックカバーであっても、すべて捨てられてしまっている。本そのものについては、書誌情報が目録データベースとして詳細に記録されているが、ブックカバーについてはその作者の名前さえ、ほとんどの場合記録されていないのが現状である。このような現状を改善し、文化資産としての価値を有するブックカバーをいかにして保存、活用するかという課題で、2007年12月から2008年3月まで、ケベック大学モントリオール校のレイモンド・ベジナ教授を総研大



図2：ブックカバーの例

の外国人招聘研究員として招き共同研究を行った。ブックカバーそのものを図書と同じように保存することは現状では困難であるから、まずブックカバーを電子化し、バーチャルなブックカバー・ライブラリを作るべきであるという結論に達し、データベースの項目などについて検討した。本データベース作成の目的は、現状のままでは忘れ去られてしまうブックカバーをデータベースという形で保存し、それぞれのデザインに対して索引を付与し、さまざまな活用を可能にすることである。データベース作成作業を開始するとともに、日本図書設計家協会（出版物のデザインに関する仕事を業務とする装丁家、装画家の集まり；<http://www.tosho-sekkei.gr.jp/>）に協力依頼を行い、快諾を得た。現在は2名のデザイナーの作品を中心にデータベース化を進めているが、今後はこの協会の会員（約200名）に順次作品の提供を依頼していくことにしている。

3) 絵はがき

福岡県立嘉徳高等学校の郷土部（注）の部室に保管されていた絵はがきの画像データベースである。これらの絵はがきがどのような経緯で郷土部で保管される



図 3 : 絵はがきの例

ようになったかは不明だが、「福岡県立嘉穂中学校生徒文庫印」の印がおされたものがあることから新制高校（1947年より）になる前から収集されたものと想定される。また、現在までに整理したもののひとつに「大正七年五月撮影」（1918年）とある絵はがきがあり、大正から昭和初期に収集されたものではないかと考えている。日本だけでなく、韓国、中国、台湾などの絵はがきもあり、現在では入手が困難な絵はがきが多く含まれており、貴重な文化資産として活用できるのではないかと期待している。

註）筆者が在学中に所属していた考古学のクラブである。

表 1 : 主な検索項目

鶏コレクション

ID, 種別（玩具、置物、絵馬、人形、小物入れ、食器など）、素材（木、金属、陶磁器、ガラス、布など）、重さ、高さ、幅、奥行き、キーワード、国名等、地名等、制作者

ブックカバー

ID, 装丁者、挿画者、イメージ分類（抽象、具象、タイポグラフィ）、キーワード、判型、ジャンル、タイトル、シリーズ名、著者、出版社、出版年

絵はがき

ID, キャプション、キーワード、国・地域、都市、発行所、年代、備考

3. 統制された索引付け

画像に対して統制された索引付けの方法としては、分類表にもとづいた索引付けと統制された語による索引付けの2つに分けることができる。

3.1 分類表にもとづいた索引付けの例

あらかじめ作成されている分類表にもとづいて、索引付けを行うもので、たとえば、図書の日録を作成する場合、表 2 のような「日本十進分類法」にしたがって分類記号を付与している。

画像をこの日本十進分類表にしたがって索引付けすることはできないが、

画像を対象とした分類表のひとつに Iconclass というものがある。この Iconclass のホームページ [8] には

Welcome to the Iconclass website. Iconclass is a classification system designed for art and iconography. It is the most widely accepted scientific tool for the description and retrieval of subjects represented in images (works of art, book illustrations, reproductions, photographs, etc.) and is used by museums and art institutions around the world.

と書かれており、愛知県美術館 [9] ではこの Iconclass コードを指定して所蔵作品の検索ができるようにしている。Iconclass は 7 つのレベルから構成され、第 1 レベルは表 3 のように 10 のカテゴリーに分類されている。

表 2 : 日本十進分類表の例

0 類 総記
000 総記
002 知識・学問・学術
.....
007 情報科学
007.1 情報理論
007.11 サイバネティクス
007.13 人工知能・パターン認識
007.2 歴史・事情
007.3 情報と社会・情報政策
007.35 情報産業・情報サービス
007.4 情報源
007.5 ドキュメンテーション・情報管理
007.53 索引法
.....

表 3 : Iconclass の第 1 レベルの分類

- 0 Abstract, Non-representational Art (抽象画)
- 1 Religion and Magic (宗教, 魔法)
- 2 Nature (自然)
- 3 Human being, Man in general (人)
- 4 Society, Civilization, Culture (社会, 文明, 文化)
- 5 Abstract Ideas and Concepts (抽象的な思想や概念)
- 6 History (歴史)
- 7 Bible (聖書)
- 8 Literature (文学)
- 9 Classical Mythology and Ancient History (神話, 古代史)

第2レベルは最大9(数字の1-9)に細分され、第3レベルは最大25("J"のぞく英大文字)に分けられている。たとえば2 Nature - 25 Earthでは25A maps, atlases, 25B continents represented allegorically, 25C geological phenomena, ……のように分類されており、以下、第4レベルから第7レベルまでは、それぞれ数字で細分されている。このように細分化されるとイメージをどのような分類にあてはめるかがわかりにくくなるが、Iconclassではオンラインで対応するイメージを表示して各レベルの語を検索することができるブラウザ[10]を提供している(図4参照)。

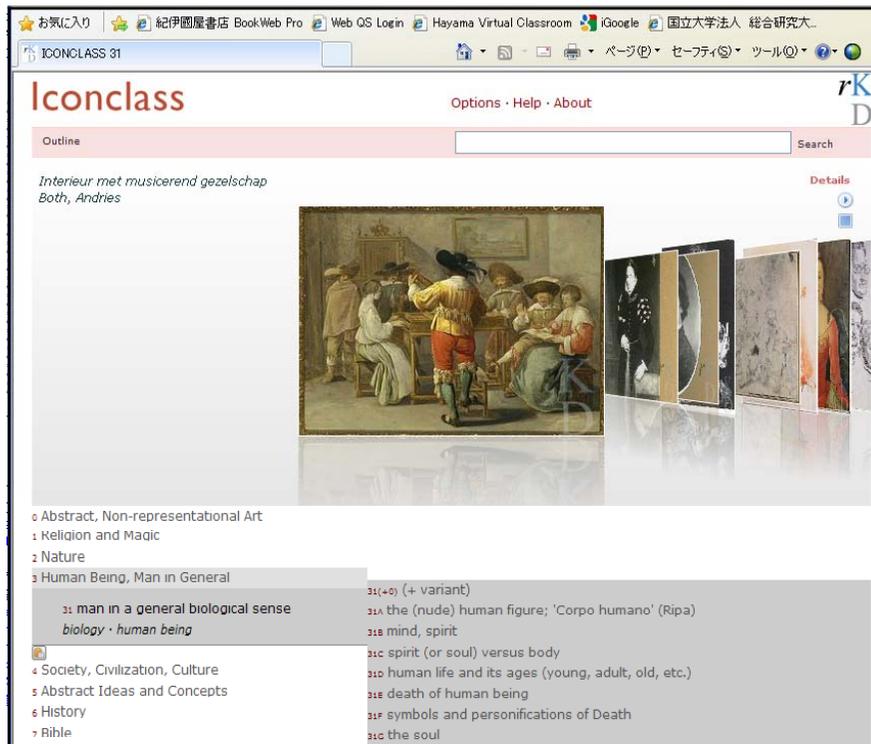


図4: Iconclass 2100 Browser

3.2 統制された語による索引付けの例

分類表ほど統制されていないが、「上位語、下位語、同義語」といった語と語の関係をもとめたシソーラスにもとづいて索引語を設定することもある。芸術の

分野で著名なものとしては米国 Getty Museum の作成した "The Art & Architecture Thesaurus (AAT)®" [11] がある。これには約 34,000 の概念から構成され、約 13 万の用語についての説明がなされている。

AAT も Iconclass と同様に、オンラインでの用語検索機能を提供しており、たとえば弥生時代の「Yayoi」を検索してみると(図5)、図6、7のように説明が表示される。



図5: AAT の検索画面

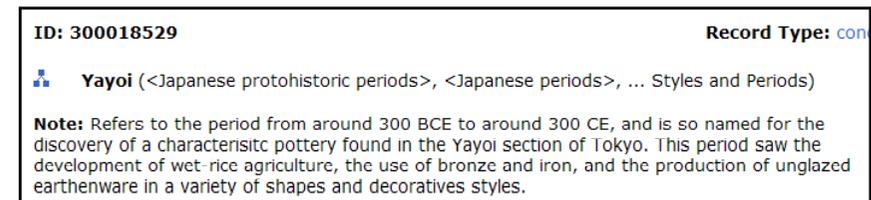


図6: "Yayoi"の説明画面

3.3 統制された索引付けの問題点

分類表やシソーラスを利用する利点は、一定の標準化がなされており情報の共有化に有効であるということである。しかし、一方では以下のような問題が指摘されている。

- 1) 「分類表」にしる「シソーラス」にしる、実用に耐えうるものを作成するには、組織的な体制が必要であり、膨大な時間と多くの専門家を要する。したがって、その作成コストは高額であり、また維持・保守のためのコストにも十分に配慮しておかなければならない。
- 2) 新しい語(用語)に即座に対応することは困難である。
- 3) 利用者である研究者の視点とずれている場合がある。
- 4) 索引付けを行う担当者は、「分類表」や「シソーラス」に習熟している必要があり、そのような要員の教育・研修には時間がかかる。

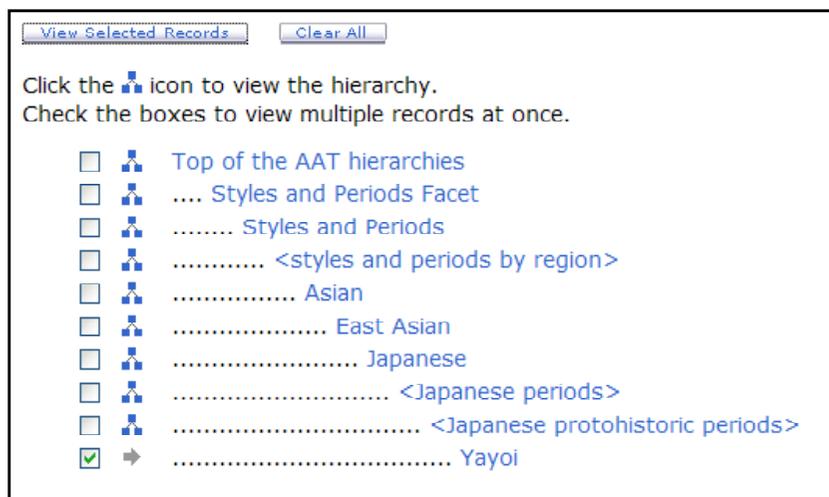


図 7: "Yayoi"の概念構造

4. 自由語による索引付け

今回作成した3つのデータベースは、総研大の ACI Sokendai から公開されている [12]. この ACI Sokendai のプラットフォームとなっている SOARS (Sokendai Academic Resources System) には、いくつかのデータベース分析機能が準備されており、この機能を活用して索引付けを行った。

イメージからの索引付け

各データベースの項目構成 (表 1 を参照) からわかるように、それぞれ検索の手がかりとなる項目が設定されている。たとえば、鶏コレクションでは「種別、素材、大きさ、注記事項、国名等、地名等」などがそうである。したがって、索引として付与する語は、これらの項目に含まれないものを画像そのものをみながら検討することになる。

図 8 は絵はがきの例であるが、ここでは「街路、建物、路面電車、通行人」が索引語として設定されている。どのような語を索引語とするかは、入力者それぞれの判断に任せられている。ただし、同じような意味を持つ「道路、道、通り、街路」は、基本的に「街路」に統一するといった作業は適時おこなった。

鶏コレクション、ブックカバーの場合は、多数の検索項目があるため、「キーワード」として索引語が付与されているレコードは少なくなっている。とくに、ブックカバーの場合は、現在までに登録している画像がほとんど「抽象画」であ



図 8: 絵はがきの例
キーワード: 街路, 建物, 路面電車, 通行人

るため、そこから検索語を付与することはほとんどできなかった。絵はがきの場合は、キャプション以外にはないので (「国・地域」, 「都市」, 「年代」はキャプションから抜き出している), ほとんどのレコードに索引が語が付与されている。

索引語のチェック

複数の人間が、また長期間にわたって索引語の付与作業をおこなうことになり、語の表記の違い (例えば、漢字とかな), 入力ミスなどで整合性がとれないことや、不適切な語を付与することが起こりがちである。このような問題を解決するには、適時索引語リストを作成し、そのリストを精査し誤りを修正していくことが必要である。SOARS にはそのようなチェックを支援する便利な機能が備わっており、次のような手順でリストを作成する。(この手順については、オンラインマニュアルに詳細が記載されている [13])

ステップ 1: 調べたいデータベースを選択し、検索語の欄を空白のまま検索キーをクリックする。

ステップ 2: 全レコードが検索されるので、検索結果一覧表示で「頻度統計」をクリックする (図 10)。

ステップ3: 集計対象の項目に「キーワード」を選択し,実行ボタンをクリックする(図10).

ステップ4: 集計表が表示される(図11).

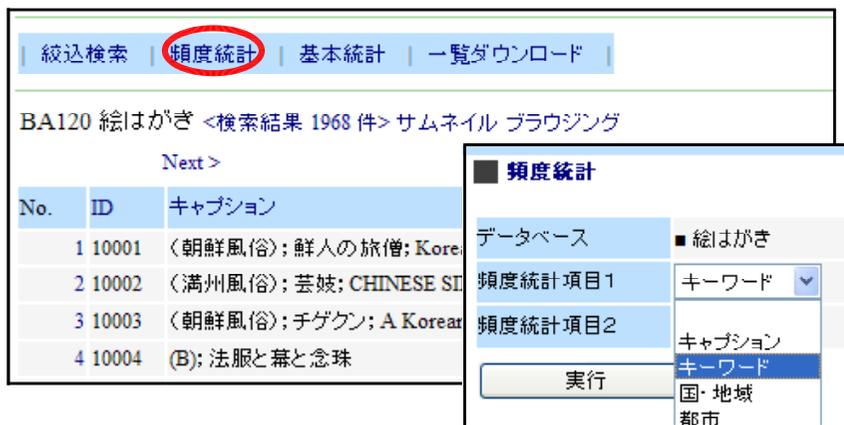


図10: 「頻度統計」をクリックし,次に「キーワード」を選択

図11が頻度を集計した表であるが,キーワードの文字コード順にソートしたり,出現頻度順にソートしたり,あるいはCSVファイルとしてダウンロードしたりすることができる。(詳細は「利用の手引き」[14]を参照のこと)

この頻度表は,はじめてデータベースを検索する前にあらかじめ作成しておけば,どのような語が索引語として設定されているかを知ることができ,検索の効率化に役立つ。

図11はチェックが終了した段階での集計表であるが,チェックが終了しない時点では図12のような頻度表が作られる。これは鶏コレクションの例であるが,「雄,オス」「雌,メス」「ヒヨコ,ひよこ」のように,統一すべき表記が異なっている。また,「台,台付き,台付き(木製),台(竹)」のように表記の乱れなども,この頻度表を作成することに

キーワード	出現頻度
	6077
1 建物	287
2 風景	111
3 鳥居	95
4 通行人	77
5 庭園	75
6 市街	71
7 橋	70
8 池	61
9 灯籠	58
10 比叡山	58

図11: 頻度表

よって容易に発見できる。「頻度表作成」→「チェック」→「修正」のサイクルを何度か繰り返すことによって,設定した索引語の多くの誤りを修正することができる。

問題点

付録1に現在(2009年11月)までに設定した索引語の一部(頻度が上位のもの)を掲載している。図13-15は,これらの索引語で検索した結果であるが,おおむね適切な結果となっていると考えている。

絵はがきの中の「阿蘇山」を「国・地域」あるいは「都市」といった地名の項目に設定した方がよいと思われるものもあるが,しばらくいろいろな利用や分析をおこなった後に再検討することにしていく。ただ,これまでにわかった問題点としては,以下のようなことがある。

- 1) 検索した結果の画像をブラウジングしていて,異質な画像や当然含まれているべき画像が含まれていない場合がある。理由は単純で,索引語のつけ間違いやつけ忘れであるが,このようなミスを少なくする工夫が必要である。
- 2) 入力する者の知識不足に起因するミスが少なからず起こっており,何らかの対応策が必要である。
- 3) 索引語の語彙数が増えた場合,簡単な分類表を作成するなどの整理が必要と思われる。

これらの問題はいずれも単独で解決できるものではなく,「自由語による索引付け」全体の問題として検討していく必要がある。

5. おわりに

分析の対象となるようなデータベースを作成する場合,重要なことは作りやすい,作り直しが容易であるといったことのほかに,短時間で作ることができるということがある。分類表やシソーラスを作成してから索引語を付与するのは,膨大な時間と人手がかかり,多くの研究者にとっては実行しがたい方法である。また,分類表やシソーラスを作成するには,しっかりとした組織と財政的な裏付けがなければならず,研究者個人や少数のグループで実現することはほとんど不可能である。

筆者自身にとってはそれなりに効果的であるが,今回のような画像から連想す

1 雄
2 つがい
3 雌
4 台
5 人
6 オス
7 長尾鶏
8 ヒヨコ
9 メス
10 太鼓
11 張り子
12 子供
13 親子
14 俵
15 チャボ
16 クリスタル
17 簪
18 籠
19 ひよこ

図12: 頻度表

る語を索引語とすることが、不特定多数の利用者にとって検索や分析に効果的であるかどうかは、現時点では不明である。今後利用者へのアンケート調査などを実施して、より効率的、効果的な索引付けの手法を求めていきたい。



図 13: 「親子」で検索した結果 (鶏コレクション)



図 15: 「路面電車」で検索した結果 (絵はがき)

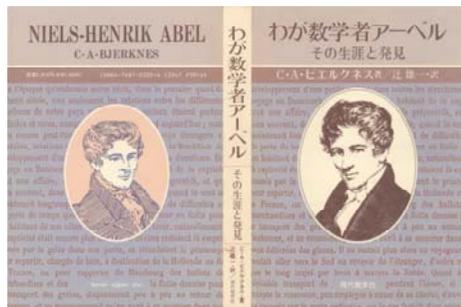


図 14: 「顔」で検索した結果 (ブックカバー)

参考文献

- 1) 黒川雅人 (日本 IBM), 洪政国 (日本 IBM), 杉田繁治 (民博): 形状情報を用いた画像の類似検索システムの試作 -民族学標本を例にして-, 情報処理学会研究報告 CH-5, pp.25-31 (1990)
- 2) 田中昭二 (ATR), 石若通利 (ATR), 井上正之 (ATR), 井上誠喜 (ATR): 画像からの感性要因の抽出 -注目領域の抽出法-, 情報処理学会研究報告 CH-33, pp.49-54 (1996)
- 3) 木本晴夫 (NTT): 感性語による画像検索とその精度評価, 情報処理学会研究報告 CH-38, pp.31-38 (1998)
- 4) 伊藤広和 (立命館大), 八村広三郎 (立命館大): キャプションと画像特徴に基づく類似画像検索, 情報処理学会研究報告 CH-45, pp.33-40 (1999)
- 5) 谷口明正 (立命館大), 八村広三郎 (立命館大): XML による画像の内容記述と検索, 情報処理学会研究報告 CH-46, pp.7-14 (2000)
- 6) 及川昭文 (総研大), 藤沢桜子 (人間文化研究機構), 洪政国 (日本 IBM), 山本啓史 (オーストラリア国立大): 研究支援機能を強化したデータベース・システムの開発, 情報処理学会 人文コンピュータ 2007, pp.213-220 (2007)

- 7) 及川昭文 (総研大), 山元啓史 (オーストラリア国立大): 研究者のためのマイ・データベース・システムの開発, 第 13 回公開シンポジウム「人文科学とデータベース」予稿集, pp.25-34 (2007)
- 8) <http://www.iconclass.nl/>
- 9) <http://search-art.aac.pref.aichi.jp/p/index.php>
- 10) <http://www.iconclass.nl/iconclass-2100-browser>
- 11) http://www.getty.edu/research/conducting_research/vocabularies/
- 12) <http://aci.soken.ac.jp>
- 13) http://aci.soken.ac.jp/fyi/manual_03.pdf
- 14) http://aci.soken.ac.jp/fyi/manual_01.pdf

付録1: 索引語一覧 (括弧内の数字は頻度)

鶏コレクション

雄(133), 雌(91), つがい(89), 台(32), ヒヨコ(29), 人(25), 長尾鶏(19), 太鼓(15), 張り子(12), 子供(11), 俵(10), 親子(9), チャボ(9), クリスタル(9), 簪(8), 籠(7), 卵(6), スワロフスキー(6), 蓋付き(5), 背中にたまご(5)

ブックカバー

人(53), 顔(30), 車(12), 馬(8), 建物(7), 地図(5), 風景(4), 鉛筆(4), 木(4), 学生(4), 子供(4), 船(3), 文字(3), 手(3), 将棋駒(3), 女(3), 鳥(2), 馬車(2), 道具(2), 目(2), 樹木(2), 工場(2), 山(2), 女性(2), 原稿用紙(2), 原稿(2), ソリ(2), グラフ(2)

絵はがき

建物(287), 風景(111), 鳥居(95), 通行人(77), 庭園(75), 市街(71), 橋(70), 池(61), 灯籠(58), 比叡山(58), 溪谷(54), 人々(53), 門(50), 舟(50), 男(50), 鍾乳洞(49), 秋芳洞(47), 川(46), 温泉(44), 社殿(42), 公園(42), 路面電車(40), 船(39), 岩(36), 広場(32), 巖島神社(32), 街路(31), 本堂(31), 製鉄所(29), 境内(29), ケーブルカー(29), 名所旧跡(27), 子供(26), 参道(26), 噴煙(25), 阿蘇山(24), 港(24), 深耶馬溪(24), 女(24), 人力車(24), 長門峡(22), 海(22), 石垣(21), 樹木(21), 山(21), 太宰府神社(21), 塔(21), 噴気(21), 駅舎(20), 自動車(20), 石段(20), 伊勢神宮(20), 馬車(19), 自転車(19), 滝(19), 東大寺(19), 耶馬溪(18), 日光東照宮(18), 英彦山(17), 楼門(17), 延暦寺(17), 人(17), 浅草(16), 泉岳寺(16), 拝殿(16), 寒霞溪(16), 鉄橋(15), 桜(15), 本殿(15)

付録2: キーワードの KWIC リスト (絵はがき)

このような KWIC リストを作成して, 索引語の整合性, 入力ミスなどのチェックもおこなった。「路面電車」が「電路面電車」となっている入力ミスがあるのがわかる。

路面電車 ; トンネル ; 市街 ; 建物
朝鮮銀行 ; 建物 ; 路上 ; 通行人 ; 路面電車
鐘路通 ; 路上 ; 建物 ; 通行人 ; 路面電車 ; 電柱 ; 自転車
帝国劇場 ; 建物 ; 堀 ; 路面電車
駅舎 ; 広場 ; 路面電車 ; 通行人
道路 ; 橋 ; 樹木 ; 路面電車 ; 鳥居 ; 通行人
拓殖株式会社 ; 建物 ; 路上 ; 通行人 ; 電 路面電車 ← 「路面電車」が「電路面電車」
街路 ; 桜 ; 路面電車 ; 堀
難波橋 ; 路面電車 ; 市街
鉄橋 ; 路面電車 ; 通行人 ; 自転車
道路 ; 橋 ; 路面電車 ; 樹木
市街 ; 長堀橋 ; 交叉点 ; 路面電車 ; 自動車
市街 ; 三菱ビル ; 住友ビル ; 建物 ; 路面電車
馬場先通り ; 市街 ; 建物 ; 路面電車 ; 自動車 ; 自転車
銀座通り ; 市街 ; 建物 ; 路面電車
銀座通り ; 市街 ; 建物 ; 路面電車 ; 自転車
京橋通り ; 市街 ; 建物 ; 路面電車 ; 自転車
銀座 ; 市街 ; 建物 ; 路面電車 ; 人力車
日本橋 ; 市街 ; 建物 ; 路面電車
日本橋 ; 市街 ; 建物 ; 路面電車 ; 人々
街路 ; 建物 ; 路面電車 ; 自動車
街路 ; 建物 ; 路面電車 ; 通行人
街路 ; 建物 ; 路面電車 ; 通行人 ; 自転車
市街 ; 路面電車 ; 建物
市街 ; 路面電車 ; 通行人
七つ寺 ; 三重塔 ; 市街 ; 路面電車
日本橋 ; 三越呉服店 ; 市街 ; 路面電車 ; 自動車
大正橋 ; 市街 ; 路面電車 ; 自動車
松坂屋呉服店 ; 建物 ; 市街 ; 路面電車
松坂屋呉服店 ; 建物 ; 市街 ; 路面電車
征清記念碑 ; 市街 ; 路面電車
帝都復興祭 ; 市街 ; 路面電車 ; 花電車