# Distribution Shift nested in Web Scraping : Adapting MS COCO for Inclusive Data

Theophile Bayet, Christophe Denis, Alassane Bah, Jean-Daniel Zucker

# Distribution Shift nested in Web Scraping : Adapting MS COCO for Inclusive Data

**Theophile Bayet** [1 2] **Christophe Denis** [1] **Alassane Bah** [3 4] **Jean-Daniel Zucker** [2 5]

## Abstract

Popular benchmarks in Computer Vision suffer from a Western-centric bias that leads to a distribution shift problem when trying to deploy Machine Learning systems in developing countries. Palliating this problem using the same data generation methods in poorly represented countries will likely bring the same bias that were initially observed. In this paper, we propose an adaptation of the MS COCO data generation methodology that address this issue, and show how the web scraping methods nests geographical distribution shifts.

## 1. Introduction

Recent works points out the major role Machine Learning (ML) is poised to play in the upcoming challenges described in the Sustainable Development Goals (Rolnick et al., 2019), and particularly in the developing countries (DCs), since those are the most at risk facing these challenges. Yet, ML is facing many issues in the DCs, including lack of generalization of models (Recht et al., 2019), lack of heavy and reliable infrastructure and data scarcity (Cvitkovic, 2018). Worldwide deployment of ML systems proves to be a challenge in itself, as data-in-the-wild distribution often differs from the training distribution considered by the models used, thus provoking distribution shift (Koh et al., 2021).

Distribution shift has lead to harmful deployment of ML systems (Trivedi et al., 2019), and arise in the computer vision community from large benchmarks mostly made up of western data (Shankar et al., 2017). The MS COCO dataset (Lin et al., 2015) is an example of these benchmarks, with more than 200 000 images collected through web scraping and annotated with 91 stuff categories.

Works that have reported data generation in the DCs faced a data quality versus data volume trade-off. Indeed, crowd sourced approaches depend on users from diverse locations and thus need extended efforts to increase the amount of data (Atwood et al., 2020), while web scraping approaches benefits from large amount of data, but face data quality issues and need to add extra cleaning steps (Malobola et al.). Large data sets in the computer vision community all used web scraping approaches, and extended data curation pipelines, in order to ensure high quality (Deng et al., 2009; Lin et al., 2015; Benenson et al., 2019). As such benchmarks are to cover the broadest context possible, collecting and annotating new data is needed to cover the context of poorly represented geographic zones.

But what happens if the data generation method is not adapted to the context of the required data ? Will the information contained in the new generated data be relevant, and how can we measure relevance of such methods ? Theses questions are of upmost importance for DCs, as data generation bias might lead to harmful models. Without generation of inclusive data, understood as images drawn from locations and cultural context that are unseen or poorly represented (Atwood et al., 2020), countries in the Global South will not be able to make the most out of ML techniques.

In this work, we adapt the MS COCO data generation for inclusive data, and show how geographic bias is nested in the adapted methodology. Section 2 introduces previous works on data generation. In section 3 we introduce the material and methods used in this work. Section 4 hosts our results, that are discussed in section 5, where we also introduce our future works.

## 2. Related work

**Datasets:** Datasets and benchmarks are the root and the core of ML fast development. Pascal VOC, MS COCO and Imagenet (Everingham et al., 2010; Lin et al., 2015; Deng et al., 2009) all contributed to the computer vision boom, and new approaches such as the Inclusive Images dataset (Atwood et al., 2020) aim to spurr such a boom with inclusive datasets development. Our work aims to extend

[1]Sorbonne Université, LIP6, 75005 Paris [2]IRD, Sorbonne Université, UMMISCO, F-93143, Bondy, France [3]UCAD, IRD, UMMISCO, Dakar, Sénégal [4] Ecole Supérieure Polytechnique, UCAD, 15915 Dakar Fann, Sénégal [5]Sorbonne Université, INSERM, NUTRIOMICS, F-75013, Paris, France. Correspondence to: Theophile Bayet <theophile.bayet@ird.fr>.

*Presented at ICML Workshop on Principles of Distribution Shift 2022*,

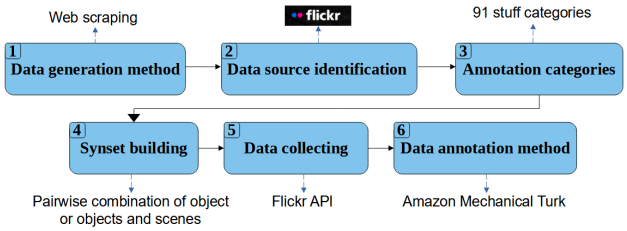


*Figure 1.* Illustration of the data generation method of the MS COCO dataset. Boxes represent the steps requiring a choice, and those choices are indicated with the dashed arrows.

such initiatives by adapting previously used data generation methodologies.

**Data collection methods:** Three data collection methods are commonly used to build datasets : web scraping, crowd sourcing and professional sourcing. The latter is the only method that depends on financial constraints, though it allows for a good data quality. Atwood et al. 2020 mixed it with crowd sourcing to build the Inclusive Image dataset. The computer vision benchmarks prefer web scraping approaches, aiming for large data volume collection.

**Web scraping process:** Web scraping relies on the choice of a data source, a request construction pipeline, and a collection algorithm. Preferred data sources are public image platforms like Flickr (Everingham et al., 2010; Lin et al., 2015) and search engines like google (Deng et al., 2009). The queries are often built using the annotations categories in order to collect adequate data, themselves following a precise construction pipeline. The Pascal VOC team chose its categories using a taxonomy inherited from older challenges and taxonomies (Everingham et al., 2010), and the MS COCO team built on Pascal VOC categories and common words from the english language. Imagenet authors used Wordnet, an electronic lexical database based on the english language (Fellbaum, 1998). Queries then consist of these annotations categories and their synonyms (Everingham et al., 2010), or combination of aforementioned categories (Deng et al., 2009; Lin et al., 2015). The collection algorithm sends the request to the data sources and gather the outputs.

**Data annotation:** Data annotation can either be done by experts (Everingham et al., 2010), ensuring consistent and accurate but time-consuming annotation procedure, or tasked to workers through Amazon Mechanical Turk(AMT) or other platforms (Lin et al., 2015; Deng et al., 2009). In both cases, it relies on a single tool used by all the annotators in order to ensure uniformity in the procedure.

**The MS COCO dataset:** The Microsoft Common Obects in Context dataset pursued the goal of "advancing the state-of-the-art in object recognition", by collecting "images of complex everyday scenes containing common objects in their natural context" (Lin et al., 2015). This goal is supported by the data generation method described in Figure 1, that relies on web scraping through the Flickr platform, using 91 "stuff" categories. The queries are built using pairwise combination of these categories, or combination of these categories with a scene list, and then fed to the Flickr API. Finally, the gathered data is annotated thanks to AMT.

## 3. Material and methods

All this work was done with an Acer Aspire without much computation capabilities.

We chose to try and replicate Lin et al.(2015) data generation method, as it is well documented and the MS COCO dataset is still used in challenges as of today (Gupta et al., 2019). Our adaptation consists in two modifications of the original method : we added a location term to the queries and switched from AMT to expert annotation.

### 3.1. Adapting queries for inclusive data collection

The query building pipeline of Lin et al. (2015) work consists in pairwise combination of object categories and scene / object category pairs. Example of those queries are **"dog +car"**, **"dog + shop"**, ... Inspired by the search engine logic that uses keywords, we added a geographic term in the queries in order to introduce local context and thus gather inclusive data.

*Geographic terms* are names of geographic zones defined by the M49 norm (United-Nations) and names of the countries in these zones. Our queries were then similar to those used to collect the MS COCO dataset, except for that location term behind each of these previous pair queries. Example of those queries are **"dog + car + Senegal"**, **"dog + shop + Western Europe"**, ... In total, the MS COCO dataset used at minimum 3640 queries on the Flickr platform. Our method lead to 270 more terms to combine with the previous combinations, leading to a minimum of 982 800 queries.

In order to fasten and lighten the data gathering phase, only the urls returned as outputs of the queries were collected. It still took more than four months of continual data gathering to collect the urls for the 23 geographic zones using one computer fetching queries in a serial fashion.

### 3.2. Annotation procedure

Instead of using AMT, all the annotation task was handled by the first author of this work, for financial and ethical reasons.

We annotated 400 images downloaded from the collected urls for each of the 23 geographic zones, totalling 9200 images. The images were chosen using a randomized ex-

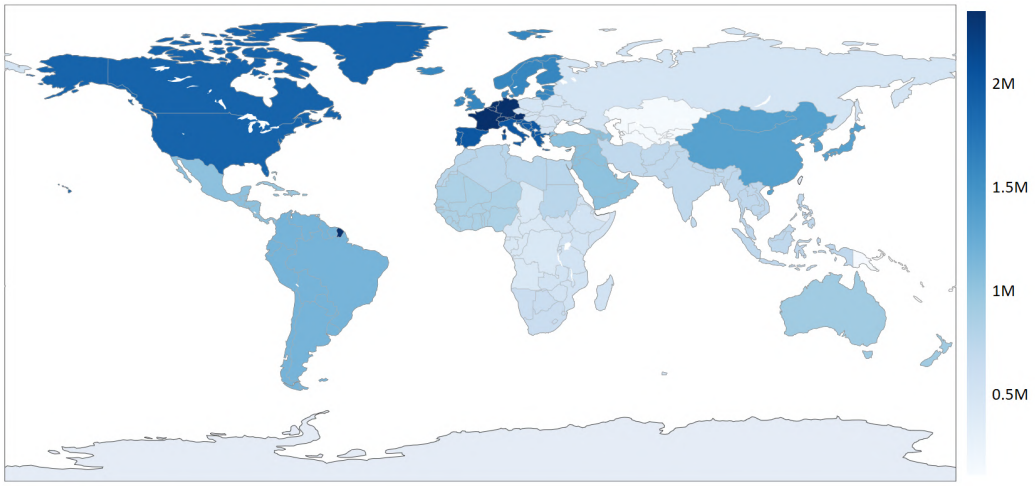


*Figure 2.* Geographic distribution of the number of urls collected using the modified queries. Western geographic zones are darker as they were favored by the data generation process

traction of urls in the scraping outputs for each geographic zone.

The annotation phase took approximately 46 hours, using the VIA ( VGG Image Annotator (Dutta et al., 2016)) tool, version 2.0.1. A custom file annotation attribute was used, allowing to annotate for the 91 categories of the COCO stuff categories (see Table 4 in Appendix A). Any complex annotation situation, such has having a hard time deciding whether an image should be given a specific label or not, was noted in a spare file, as well as the taken decision. This allowed for constant behavior for the whole process.

## 4. Results

### 4.1. Data gathering

Once all data is gathered, we observe that some geographic zones, namely Northern, Western and Southern Europe as well as Northern America generated more than 1.5 millions urls, when every African zone generated less than 1 million urls and Central Asia, Melanesia and Micronesia generated less than 100,000 urls (see Figure 2).

The same tendencies appeared in the average number of urls generated per country in each geographic zones. Northern America has the lead by far with more than 250,000 urls generated per country, followed by Western Europe (more than 200,000) and Eastern Asia (a bit more than 150,000). Central Asia, Melanesia and Micronesia are once again undermined with less than 10,000 urls generated per country. All zones of Africa generated between 20,000 (Eastern Africa) and 90,000 (Southern Africa) urls per country, as shown in Figure 5 in Appendix B.

Some of the categories generated significantly more data during the fetching phase than others, namely categories ”person” (15.54 percent of the urls) and ”book” (15.55 percent of the urls). Next are the ”vehicle” categories (”bicycle”, ”car”, ”motorbike”, ”airplane”, ”bus”, ”train”, ”truck”, ”boat”) that generated 17.3 percent of the urls when all summed up. Table 1 in Appendix C shows the number of urls collected for the top ten categories.

Some urls were fetched in multiple geographic zones, indicating that the methodology suffers from lack of good geographic context generation capability. A total of 360,974 urls were shared between the geographic zones representing 1.12 percent of the total gathered data. Some example of these images can be found in Figure 7 in Appendix D.

### 4.2. Observations on annotations

We found that the distribution of the labels given to the images during the annotation phase does not mirror the category data generation distribution. Instead, it put the emphasis on people-related labels (”person”, ”shoe”, ”hat”, ”eyeglasses”), with the label ”window” over represented, and the label ”book” almost not used despite its importance in the data generation process. This distribution is shown in Table 2 in Appendix C.

Among the annotated images, a lot were uncorrelated with the categories used to label the images, reaching the surprising number of 5,550 inadequate images (example of those images are shown in Figure 6 in Appendix D). The distribution of those images in the geographic zones is detailed in Figure 3. This high volume of inadequate data highlights the particularity of web scraping among data generation methods : large volumes of low precision data.

Complex annotation situations encountered during this an-

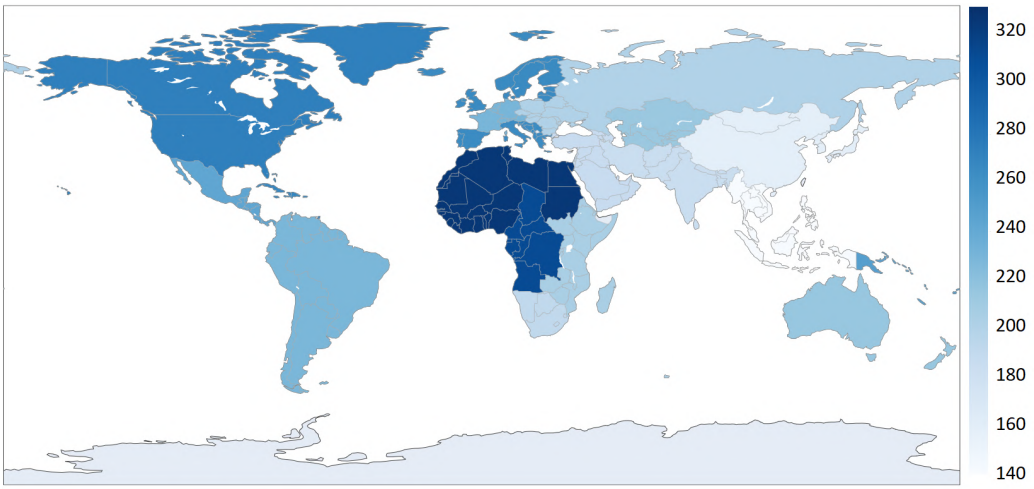


*Figure 3.* Geographic distribution of the number of not annotated images from the 400-batch annotation round, due to absence of correlation between the image and the annotation categories. Northern, Western and Middle Africa are darker, which mean they were undermined by the data collection process compared to other geographic zones.

notation phase were listed throughout the process and can be found in Figure 3 in Appendix E. Some of these complex situations come from cultural or technological differences, that render some of the categories inadequate for the context of the image. For example, baskets made of wickers or other materials wouldn't be counted as backpacks or handbags, or hats when they are worn on the head. Chopsticks would not fit as fork or knife either even though they are common cutlery in Asia.

## 5. Discussion and Future Works

We aimed through this work to adapt Lin et al.(2015) data generation methodology for inclusive data. As our adaptation lead to only 1.2 percent of the total data shared between geographic zones, we can sincerely believe we achieved our goal. Yet, there is to date no metric that allows to measure the suitability between data and geographic context, nor do we have access to verified geographically-tied datasets, when both would be needed to properly assess our performance.

Nonetheless, this work showed how using location terms altered the outputs of the MS COCO web scraping method, and how some geographic areas (small islands, Africa and Central Asia) are undermined by web scraping methods. The fact that these tendencies also appears in the mean number of urls collected per country for each geography zone shows that it is not due to some geographic zone having more countries than others (and thus more request). From better data volume collection for those undermined locations to more precise data fetching, there is space for improvement in web scraping methods.

We did not address a number of downsides of the web scraping data generation method in this work, such as bias coming from the data source, from the requests language, or from the categories that are anchored in western context (Noble, 2018; Prabhu & Birhane, 2020; Leavy et al.). The fact that only one annotator annotated the whole database ensure some data quality, but also undermines the diversity and increase risk of bias inference in the annotations. When annotating bigger volumes, using multiple experts or crowd sourcing should be preferred.

Future works will focus on addressing these issues, as well as exploring new data sources and crafting the metrics needed to measure the improvements in fighting geographic bias in data generation. Ongoing works include exploration of the generalization gap for models trained on the MS COCO dataset, and tested on our newly annotated data.

## 6. Conclusion

We proposed a data generation methodology that allows inclusive data generation for geographic zones all over the globe, by adapting the MS COCO query building pipeline. Addition of geographic terms to the original queries proved to be efficient to gather contextualized data.

We also brought to light how distribution shift is nested in the web scraping method, and showed the need for improvements of these methods, as well as new metrics to measure those improvements. Better scraping method will result in better data generation for the global South, which will ease the deployment of ML systems in DCs. Preliminary results show that the proposed adaptation of the method from MS Coco is a first step in this direction.

## Acknowledgements

We would like to thank the anonymous reviewers of this work for their usefull feedback. We would alo like to thank the PDI MSC grant for funding this work.

## References

Atwood, J., Halpern, Y., Baljekar, P., Breck, E., Sculley, D., Ostyakov, P., Nikolenko, S. I., Ivanov, I., Solovyev, R., Wang, W., and Skalic, M. The Inclusive Images Competition. In Escalera, S. and Herbrich, R. (eds.), *The NeurIPS '18 Competition*, pp. 155–186. Springer International Publishing, Cham, 2020. ISBN 978-3-030-29134-1 978-3-030-29135-8. doi: 10.1007/978-3-030-29135-8_6.

Benenson, R., Popov, S., and Ferrari, V. Large-scale interactive object segmentation with human annotators. *arXiv:1903.10830 [cs]*, April 2019.

Cvitkovic, M. Some Requests for Machine Learning Research from the East African Tech Scene. *arXiv:1810.11383 [cs, stat]*, November 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. pp. 8, 2009.

Dutta, A., Gupta, A., and Zissermann, A. VGG Image Annotator (VIA), 2016. URL http://www.robots.ox.ac.uk/~vgg/software/via/.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int J Comput Vis*, 88(2):303–338, June 2010. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-009-0275-4.

Fellbaum, C. (ed.). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, Mass, 1998. ISBN 978-0-262-06197-1.

Gupta, A., Dollár, P., and Girshick, R. LVIS: A Dataset for Large Vocabulary Instance Segmentation. *arXiv:1908.03195 [cs]*, September 2019.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv:2012.07421 [cs]*, July 2021.

Leavy, S., O'Sullivan, B., and Siapera, E. Data, Power and Bias in Artificial Intelligence. pp. 5.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*, February 2015.

Malobola, L. B., Rostamzadeh, N., and Mohamed, S. Se-Shweshwe Inspired Fashion Generation. pp. 5.

Noble, S. U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York, 2018. ISBN 978-1-4798-4994-9 978-1-4798-3724-3.

Prabhu, V. U. and Birhane, A. Large image datasets: A pyrrhic win for computer vision? *arXiv:2006.16923 [cs, stat]*, July 2020.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet Classifiers Generalize to ImageNet? *arXiv:1902.10811 [cs, stat]*, June 2019.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y. Tackling Climate Change with Machine Learning. *arXiv:1906.05433 [cs, stat]*, November 2019.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv:1711.08536 [stat]*, November 2017.

Trivedi, A., Mukherjee, S., Tse, E., Ewing, A., and Ferres, J. L. Risks of Using Non-verified Open Data: A case study on using Machine Learning techniques for predicting Pregnancy Outcomes in India. *arXiv:1910.02136 [cs, stat]*, October 2019.

United-Nations, S. D. UNSD — Methodology. URL https://unstats.un.org/unsd/methodology/m49/.

## A. Annotation tool

The annotations were generated in the COCO format thanks to the VIA tool (VGG annotation tool,(Dutta et al., 2016)). A custom file annotation attribute was designed in order to ease the annotation of the images with the 91 ”stuff” categories.

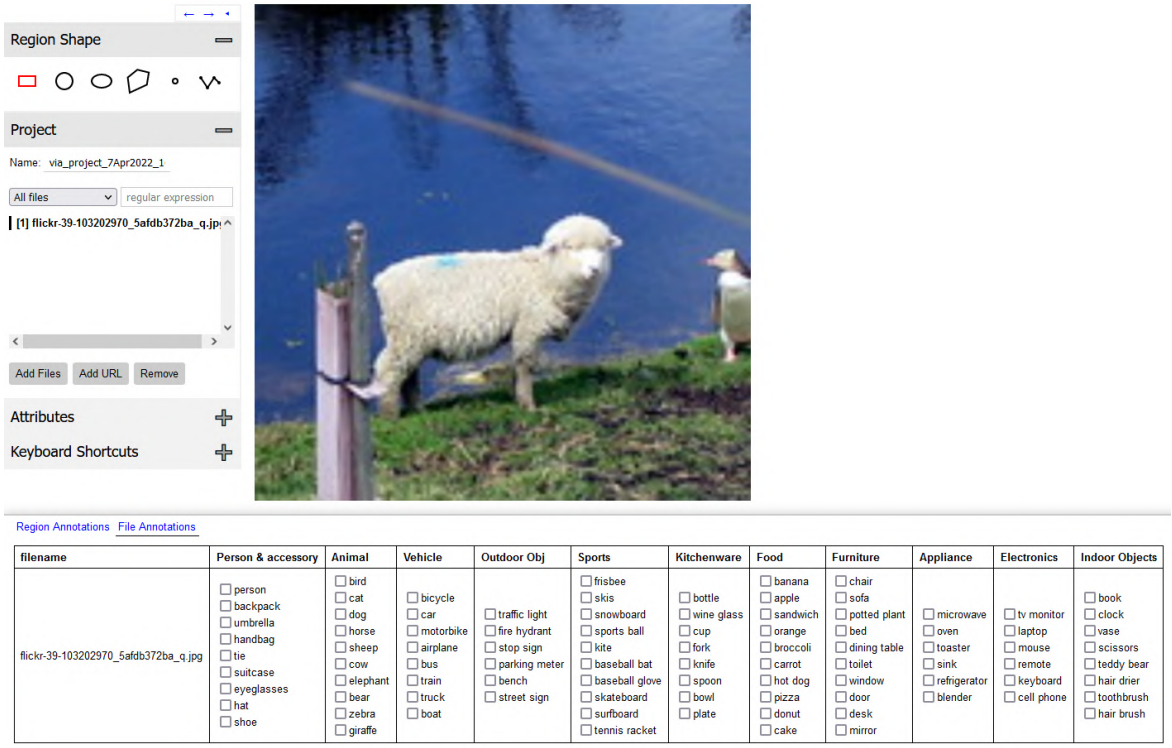


*Figure 4.* Illustration of the annotation tool and the custom file annotation attributes, that allows to annotate all the 91 classes of Ms COCO on one image.

## B. Volume of urls collected

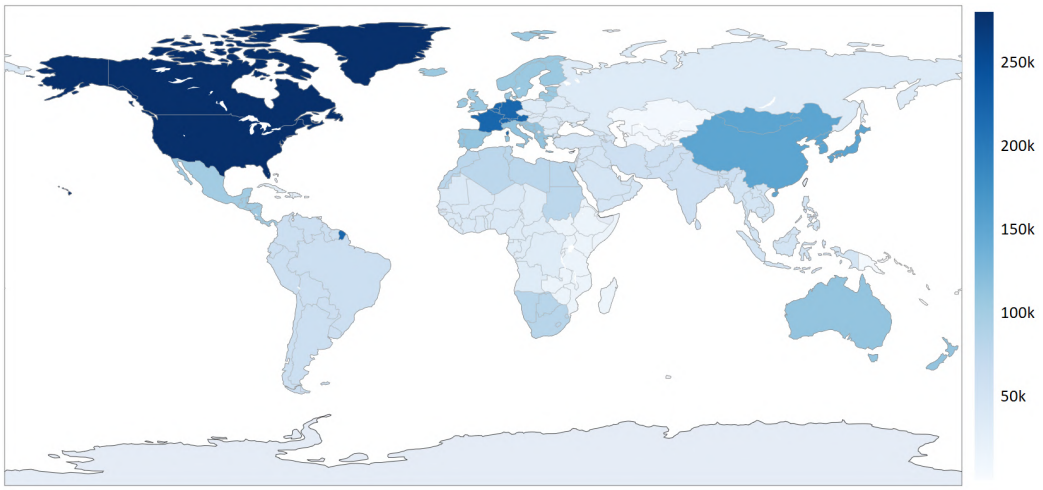


*Figure 5.* Geographical distribution of the number of urls collected for each geographic zone, using the modified queries, averaged with the number of countries per geographic zone.

Figure 2 shows how the volume of data collected differs from one geographic zone to one another. But one could wonder if

this variation is not due to some geographic zones being composed of more countries, and thus having more queries fueling them. To prove our point, we calculated the average number of urls collected per country, for each geographic zone, as shown in Figure 5. As roughly same tendencies as the previous figure are observed, we conclude that the data generation method suffer from geographic bias.

## C. Categories and labels distribution in the database

We show here the distribution of the categories among the collected data in Table 1, and the distribution of the labels given to random images selected in the collected data in each geographic zone in Table 2.

Even though each category was used the same amount of time in the queries, the collected data is not evenly distributed into each category as seen in Table 1.

Categories ”book” and ”person” are far ahead in terms of number of urls collected with both around 5 million urls, while the fifteen categories that collected the lesser amount of data are under 10 thousand urls per categories.

*Table 1.* Number of urls collected using the modified querries, for the top ten categories.

| CATEGORY | NUMBER OF URLS COLLECTED | PERCENTAGE OF THE DATA COLLECTED (%) |
|---|---|---|
| BOOK | 5,001,158 | 15.56 |
| PERSON | 4,998,124 | 15.55 |
| CAR | 1,677,570 | 5.22 |
| BOAT | 1,397,981 | 4.35 |
| BIRD | 1,315,220 | 4.09 |
| TRAIN | 1,293,387 | 4.02 |
| WINDOW | 1,143,959 | 3.56 |
| DOOR | 1,143,024 | 3.55 |
| HORSE | 1,062,142 | 3.30 |
| PLATE | 886,073 | 2.76 |

The label ”person” is bar fay the most given during the annotation phase, highlighting the person-centered bias of the Flickr platform data. Up to 8 labels were not used at all during the annotation phase, and 36 labels were used less than 10 times when annotating the images.

*Table 2.* Number of labels used during the annotation phase, for the top ten labels.

| CATEGORY | OCCURRENCE OF LABEL | PERCENTAGE OF GIVEN LABELS (%) |
|---|---|---|
| PERSON | 2,042 | 21.86 |
| WINDOW | 1,026 | 10.99 |
| SHOE | 658 | 7.05 |
| HAT | 630 | 6.75 |
| EYEGLASSES | 455 | 4.87 |
| CAR | 450 | 4.82 |
| DOOR | 422 | 4.52 |
| CHAIR | 317 | 3.40 |
| BIRD | 264 | 2.83 |
| STREET SIGN | 232 | 2.48 |

## D. Example of images

In this appendix, we show examples of images that have specific characteristics : images that were not given any label during the annotation phase (Figure 6), and images that were found in multiple geographic zones (Figure 7).

The images considered inadequate that were not given any label such as those in Figure 6 either contained no category used to annotate the images, or were judged offensive or unsafe by the annotator.

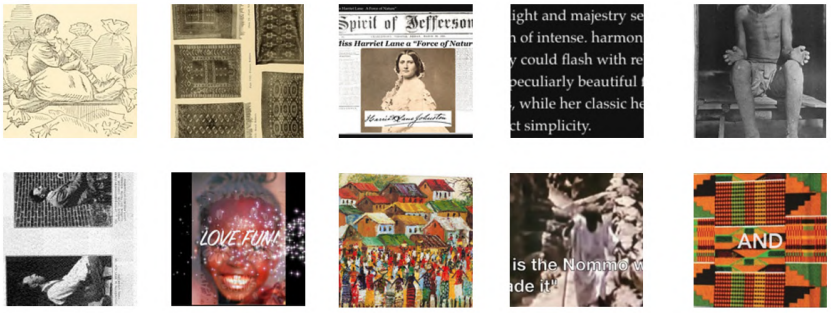


*Figure 6.* Example of images that were not given any label.


*Figure 7.* Example of images that were shared by multiple geographic zones.

## E. Situations met during the annotation phase

*Table 3.* difficulties met during annotation process and chosen behavior to overcome them

| ANNOTATION SITUATION | CHOSEN BEHAVIOR |
|---|---|
| STATUES OF PEOPLE COULD BE ANNOTATED AS 'PERSON' | STATUES WERE NOT CONSIDERED PEOPLE |
| CARTS ARE USED AS USUAL TRANSPORTS AROUND THE WORLD | CARTS WERE NOT CONSIDERED CARS |
| WIDE VARIETY OF SHOES, FLIP-FLOPS AND BOOTS CAN BE LABELLED 'SHOE' | EXCEPT FOR BEING BAREFOOT, ANYTHING WORN ON FEET WAS CONSIDERED A SHOE |
| WOVEN BASKETS COULD BE LABELLED AS HANDBAGS, BACKPACKS, OR NEITHER WHEN CARRIED ON THE HEAD | WOVEN BASKETS WERE NOT GIVEN ANY LABELS |
| A WIDE VARIETY OF VEIL, FABRICS AND OTHER HEAD-WORN MATERIALS COULD BE CONSIDERED AS A HAT | VEIL, FABRICS AND ANYTHING WORN ON THE HEAD WAS CONSIDERED A HAT (EXCEPT FOR CARRYING PURPOSES, LIKE WOVEN BASKETS) |
| CHOPSTICKS DO NOT FIT IN ANY CUTLERY CATEGORY | CHOPSTICKS WERE NOT ATTRIBUTED ANY LABEL |
| WINDOWS ARE A COMPONENT OF A LOT OF OBJECTS (CARS, TRUCKS, DOORS, ...) | WINDOWS WERE ALWAYS LABELED WHEN PRESENT |
| CHAIRS ARE A COMPONENT OF OTHER CATEGORIES (CARS, TRAINS, BUS, ...) | WHEN VISIBLE, CHAIRS WERE LABELED |
| UMBRELLA AND PARASOL ARE PRETTY COMMON | PARASOL WERE LABELED AS UMBRELLAS |