Image Dataset Tool

ROSHAN SHRESTHA, University of Texas at Arlington

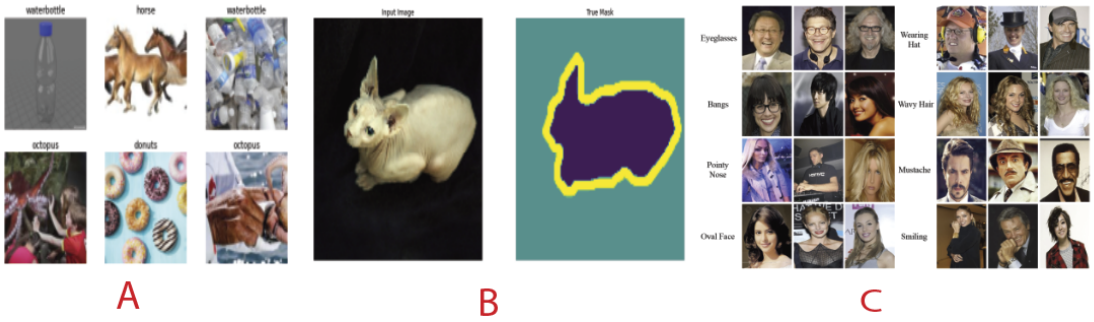

Fig. 1. Fig. A shows image labeling and Fig. B shows image mask segmentation. Fig. C is an example of an image dataset

Creating a custom image dataset helps people reach the variety, variability, and volume they desire in the dataset. In this paper, we explore image labeling and segmentation in order to understand what automation of those tasks look like. We present a user study of 5 participants which offers insight into the level of automation the tool achieves for image labeling and segmentation. In the study we carried out a qualitative analysis to identify what key common themes that people look for in an image dataset. The results helped us design a pipeline achieving higher efficiency and ease of use.

CCS Concepts: • **Human-centered computing Human computer interaction (HCI)**; *Haptic devices*; User studies.

Additional Key Words and Phrases: image dataset, automate, annotation, segmentation, mask

ACM Reference Format:

Roshan Shrestha. 2018. Image Dataset Tool. In *TEI '21: ACM International Conference on Tangible, Embedded and Embodied Interaction, June 03–05, 2018, Salzburg, Austria.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

With the advance in digital technology over the past years, there has been a huge surge in the availability of datasets. Structured and semi-structured dataset in particular has become critical in most domains and professional rules. With the rise of data science, millions of datasets have been published, some under an open license, in institutional repositories, online markets, and on social networks, in sectors from science and finance, to marketing and government. People use such datasets to improve services, design policies, generate business value, advance science, and make more informed decisions. The search for datasets often starts on the web. People get bombarded by a never-ending list of options. Despite the increased availability, they experience many difficulties in finding, accessing and assessing the datasets they desire. Relevance, usability, and quality are the three major aspects that matter when selecting a dataset. People have to make sense of each of these aspects to make an informed decision whether to use it for their task. A slight glance at the dataset is not enough to decide if it is useful for a task. From a user's perspective, making your own dataset is more efficient as it covers the relevance, usability and quality for the task at hand. There are a few tools available that help

with creating your own image datasets but these tools do not address the actual problem of automating the task at hand: the task being automation of labeling images and selecting the image masks.

Annotating images in the traditional way requires a lengthy period of time which is dependent on the size of the dataset. One way to decrease the time would be to use a group of people but this leads to variances in the annotated images which leads to bias in the dataset. We have tried using various tools like LabelIMG, VGG Image Annotator, MakeML and many more that help label images for datasets through a tedious but doable process. All these tools help people label images but the actual task at hand is still manual. People want to focus on their end goal rather than work around the clock to gather and arrange datasets.

In this paper, we aim to build a pipeline for an automated image dataset tool: automate image annotations such as labeling and selecting image masks. The image dataset tool uses ResNet-34 Model with CNN Learner) to automate labeling and U-Net model with MobileNetV2 for image segmentation. Our tool uses a machine learning model to automate the annotation process. Like all learning models, this model also requires a dataset to learn from. But our goal is to create a dataset. So, we have a chicken and egg problem here. Our approach for this problem is to hand annotate a hundred images, use these hundred annotated images to train the model. The FastAI pretrained model we use does not require an exponential number of images to learn from. This helps us set a slow learning curve for the model while also achieving the task of image annotation.

In this work, we describe an image dataset tool that supports people in generating bespoke image datasets with bounding box labels. The tool collects images for the dataset using web scraping techniques, provides an interface for hand-specifying bounding boxes on target class labels. Using FastAI and other learning models, the tool automates image annotation. The image dataset tool examines the dataset, mainly the domain of each features and makes sure the dataset has evenly distributed images covering all features. If it doesn't get to an even distribution, it scrapes the web for more images.

User Study was the appropriate choice of evaluation for us as it reflects real life usage of the tool when people use it to create an image dataset for their projects. We conducted a user study of 5 participants where the participants created image datasets on a variety of topics of their choice. Only a small set of guidelines were provided to help the participants create their datasets. A short survey was conducted at the end of the user study.

2 RELATED WORK

The most closely related prior work is on tools for data transformation, extension and linking of large datasets, calculating relations and collecting tools.

2.1 Data Wrangling

Data wrangling is to transform and map raw data to some other format with the intent of making it more appropriate and valuable for a variety of purposes and analytics. Nikolay et al. [14] transformed, linked and extended massive data sets. They decompressed data, split data in chunks for faster processing and horizontal scalability of inputs, executed the data wrangling pipeline and imported the resulting data set. We aim to create a similar pipeline where our tool scrapes the web for the images, then processes the images first to train the learning model, then to use the model to

automate image annotation. Tan et al. [4] proposed a system that allows the creation of automated test dataset served for the evaluation of sensitive data leak analysis systems in Android applications and in Application Framework. Our focus is to automate the annotation pipeline while making sure we address the chicken and egg problem efficiently. Mathias et al. [11] presented a tool for a stream-based indexing and schema extraction of Linked Open Data (LOD) at web-scale. The image dataset tool focuses on utilizing the data wrangling pipeline to create automated image datasets. Our approach draws inspiration from these works in terms of extracting bounding regions through segmentation from the raw images.

2.2 Information Extraction

Extracting structured information from unstructured or semi-structured images has always proven to be a cumbersome task. Oren et al. [5] sketched the transformation of information extraction from a targeted method, appropriate for finding instances of a particular relationship in text, to an open ended method that scaled to the entire Web and can support a broad range of unanticipated questions. Our goal with the image dataset tool is to enable creation of datasets of all kinds of images covering fields of business to medical science. Fabian et al. [15] presented the SOFIE system that reconciles pattern based information extraction, entity disambiguation, and ontological consistency constraints into a unified framework. Oren et al. [6] presented an overview of KnowItAll's novel architecture and design principles, emphasizing its distinctive ability to extract information without and hand-labeled training examples. We propose a tool that focuses on user input and interactions for the automation of annotating image masks: segmenting bounding regions for the image dataset.

2.3 Calculate Relations

Eugene et al. [1] presented Snowball, a system for extracting relations from large collections of plain-text documents that requires minimal training for each new scenario. Ricardo et al. [2] presented their understanding behind semantic search, how they used shallow semantics to improve Web search and how the usage of search engines can capture the implicit semantics encoded in the queries and actions of people. Guha et al. [9] discussed the evolution and development of Schema.org and how it can contribute to the need to pull together data from different sources and hence the need for sharing vocabularies. The image dataset tool focuses on using various machine learning models like ResNet-34 and CNN learner on the bounding images selected by the user. We plan to implement an image classifier that predicts the class label and decides what kind of segmentation to use on the image.

2.4 Collection Tool

There are various methods of collecting web images. It is an important task to only use images that you have obtained consent for. Alon et al. [10] described a system that provided access to metadata about billions of datasets within an enterprise. They used ranking to identify important datasets. Dan et al. [3] described a dataset discovery tool that provided search capabilities potentially over all datasets published on the Web. The approach relied on on open ecosystem where dataset owners and providers published semantically enhanced metadata on their own sites. This metadata was aggreagated, normalized, and reconciled by the search engine to let users find datasets. Gabe et al. [7] implemented a BACnet/IP adapter for the data collection tools. They presented an open dataset of building point attributes for use in developing and evaluating data-driven metadata normalization methods for buildings. We propose a web scraper that takes website links, number of images and class label as inputs from the user and uses Google's webdriver to extract links and download the images.

Fujimoto et al. [8] attempted to make a dataset of comic books. They mainly focused on the images and their semantics in the comic books. Their end goal was to use the dataset to process comic books, translate the semantics of the images in various languages. Mousselly et al. [13] generated a dataset of 14 million geotagged photos crawled from Flickr with their metadata. Their end goal was to enable search-based automatic image annotation for reverse geotagging by exploiting collective knowledge presented by user-tags in the uploaded photos to predict tags for new unlabeled images. Loni et al. [12] created a dataset of more than 10000 fashion images. Their goal was to use the fashion dataset for clothing recommendation, trend analysis, for multimedia analysis. Similarly, datasets created using the image dataset tool can be used for a variety of purposes from business to fashion, medical science and more.

3 FRAMING

Our image dataset tool is hylomorphic as it has various designs as its backbone. Since most people prefer online datasets as their first choice due to the ease of access and use, the image dataset tool fulfils the needs of those who intend to create new things with new designs and the datasets available online do not suffice them. So, the tool has a pro-C level of use. The tool was originally created with learning models in mind but is not limited to any field of study. The tool has an inspirationalist process of creation as people can break away from normal trends of creating an image dataset and bring their own twist within the dataset they create. The tool has a running shoes type of creative support as it enables people to save time by creating their own image dataset which ultimately helps them advance further into their goals.

4 FORMATIVE STUDY/SYSTEM DESIGN MOTIVATION

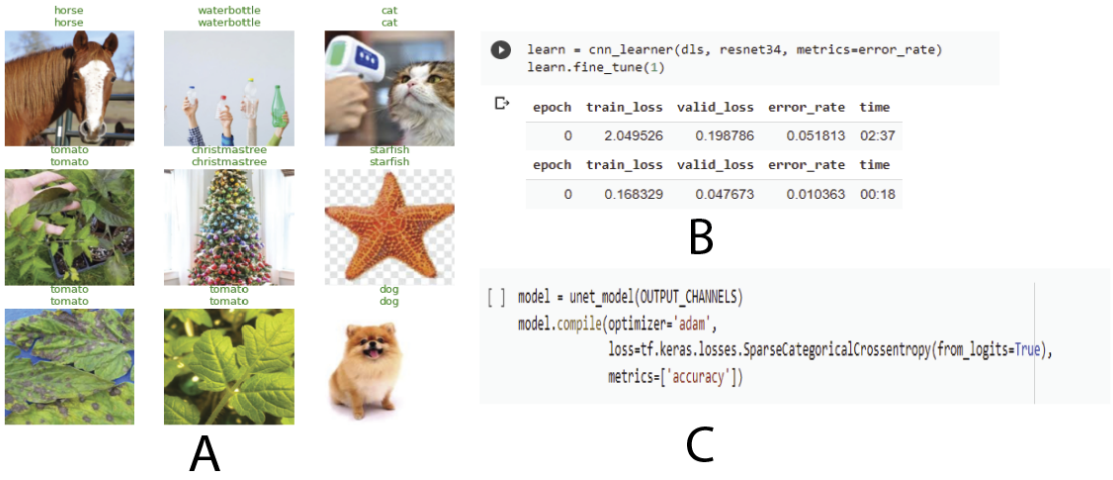

Fig. 2. Fig A. Automated Labeling Results, Fig B. CNN Learner, Fig C. UNET model

To guide our design to automate generating image datasets along with input from the user, we start out by scraping web images. We use packages like Selenium, Beautiful Soup, and Google webdriver to automate the download of n images. The tool starts out by opening an automated session of google chrome, goes to the search link provided and starts saving links to the first n images it finds. After the links are collected in a .csv file, the tool downloads the images

and saves them in a folder. With Google's new update, it prevents users from download a large number of links at once. So, to overcome this, our webcrawler goes through each image and saves its link.

4.1 Annotation-Labeling

We use Python FastAI to generate an image classifier (Fig 2B). A pretrained fastAI model object is imported and tuned to the collection of images downloaded through the webscraper. We use CNN learners with Resnet34 model from FastAI to automate image labeling. Since we use a FastAI pretrained model which does not require a large number of images to train on, the 100 images annotated by the user are more than enough to train the model. This helps us solve the chicken and egg problem we faced. The FastAI model also saves computation time as it averages to less than five minutes (Fig 6.).

4.2 Annotation-Masks

We use U-Net model with MobileNetV2 from FastAI to automate the image segmentation task (Fig 2C). The learning model requires true masks for the images it learns on. This requires the user to selects mask for the first 100 images used the train the learning model. The accuracy of the image masks generated by the model is about 95 percent.

Overall, the image dataset tool has minimal training time. From the user studies conducted, we were able to conclude that the labeling accuracy is 95 percent.

5 ARTIFACT PROCESS / INTERACTION OR SYSTEM IMPLEMENTATION

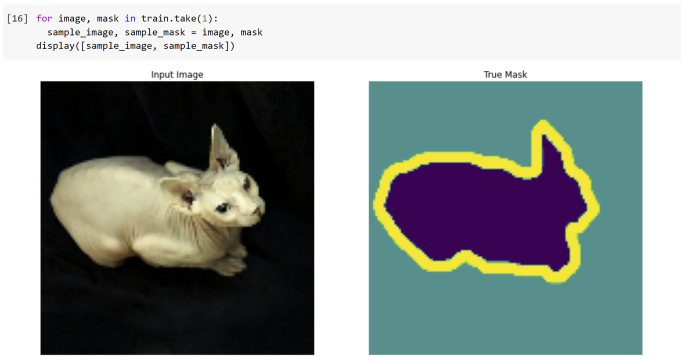

Fig. 3. Original Image (Left) and True Bounding Region (Right)

The main goal of the image dataset tool is to automate image annotation i.e. automation of labeling images and mask generation. It uses machine learning models to perform the annotations: ResNet-34 with CNN learner for labeling and U-Net with MobileNetV2 for image segmentation. Fig 2. shows the results of automated image labeling. Fig 3. shows the segmented image on the right - a poor bounding region obtained at epoch 0. By epoch 20 on Fig 5. the tool generates a segmented image that is very similar to the true bounding region.

The images in the dataset created by the tool lack the auto generated mask. The FastAI model used requires true image

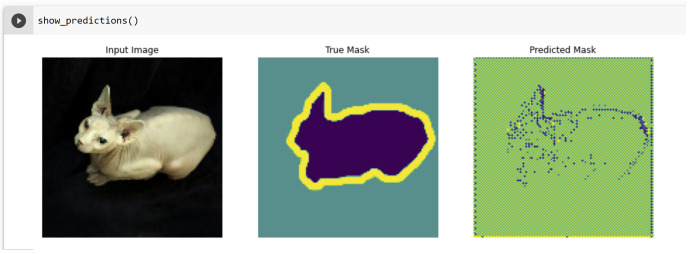

Fig. 4. Segmentation epoch = 0 shows a poor prediction of the bounding region

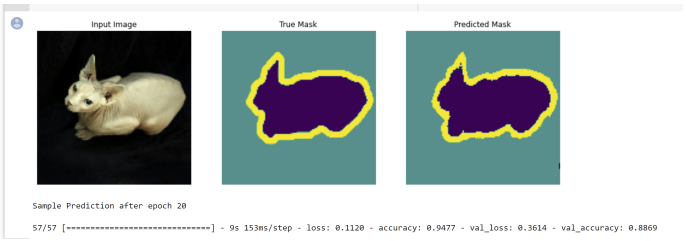

Fig. 5. Segmentation epoch = 20. The predicted bounding region is very close to the true bounding region.

masks for the images used to train the model. Since selecting image masks is a lengthy process, we look forward to automating this step.

6 EVALUATION

We chose user study as a method of evaluation because it reflects the real life usage of the tool. A user study would enable us to learn about user expectations from the tool and address any difficulties that users face when using the tool to create an image dataset of their choice.

We conducted a user study with 5 participants. Since we do not have a complete package yet, users would have to install various dependencies required to run the tool. So, we set up a single PC where all the tests were conducted. All 5 participants were able to create an image dataset using the tool. Some favored using the tool and found it useful for their projects while some were more inclined to use datasets found online. In particular, Participant 3 created a dataset of Christmas Trees. "The dataset works when I test a Christmas tree. But when I tested the image of a man, the result came out to be Tomato". This discovery by participant 3 was really helpful. The tool was combining all the datasets created by the participants. As more participants tried it, the size of the dataset was increasing. To resolve this, we made some minor changes in the automated pipeline. Fig 6.C shows the results of participant 3.

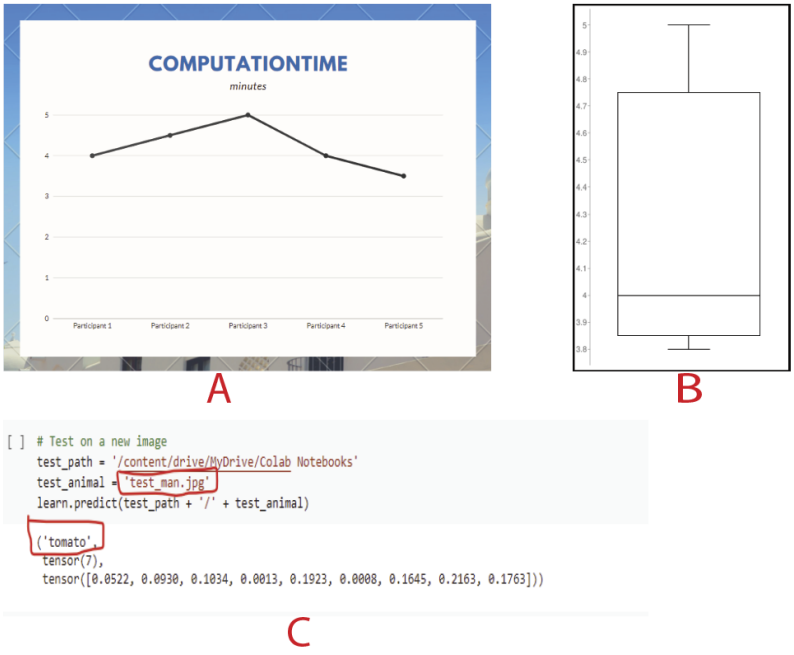

Fig. 6. User Study Results A. Computation Time B. Box Plot C. Participant 3 results

Fig 6. A and B show that the avearage computation time for the participants was about 4 minutes. Once the user study was complete, we conducted a small survey asking participants if they found the tool useful, if they would use it on their projects, and what their expectations were. After the survey, we found that most of the participants were expecting a GUI based tool and an even more automated pipeline. To put it all together, the user study was fruitful and pushed the image dataset tool to improve.

7 DISCUSSION

The image dataset created by the tool can be saved in various formats like .csv, .json and more. We tried various approaches including convolution and edge detection methods to automate image segmentation. So far the U-Net model works the best for the tool but needs true mask as input to train the learning model. We plan to work on image segmentation such that users do not have to spend their time selecting image masks for the model. For future versions of the tool, we will be making a full fledged tool with a simple GUI and an even more automated pipeline to save time. We also aim to decrease the average computation time down to 3 minutes.

8 CONCLUSION

With the increasing demand for custom image datasets, tools to help people create their own datasets are necessary. We have presented Image Dataset Tool, an automated image dataset generation tool that enables automated image labeling and segmentation. The core components of our tool are the webscraper, the classifier learning FastAI models, and the image segmenting models, which simultaneously achieve speed, control, and ease of use.

9 ACKNOWLEDGEMENTS

We are grateful to all our user study participants and a big Thanks to Google Colab and Fast AI. The tool was created using Google Colab and Fast AI.

REFERENCES

[1] Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting Relations from Large Plain-Text Collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries* (San Antonio, Texas, USA) *(DL '00)*. Association for Computing Machinery, New York, NY, USA, 85–94. https://doi.org/10.1145/336597.336644

[2] Ricardo Baeza-Yates, Massimiliano Ciaramita, Peter Mika, and Hugo Zaragoza. 2008. Towards Semantic Search. In *Proceedings of the 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems* (London, UK) *(NLDB '08)*. Springer-Verlag, Berlin, Heidelberg, 4–11. https://doi.org/10.1007/978-3-540-69858-6_2

[3] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1365–1375. https://doi.org/10.1145/3308558.3313685

[4] Nguyen Tan Cam, Nghi Hoang Khoa, Le Duc Thinh, Van-Hau Pham, and Tuan Nguyen. 2019. Proposing Automatic Dataset Generation System to Support Android Sensitive Data Leakage Detection Systems. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence* (Bali, Indonesia) *(ICCAI '19)*. Association for Computing Machinery, New York, NY, USA, 78–83. https://doi.org/10.1145/3330482.3330522

[5] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open Information Extraction from the Web. *Commun. ACM* 51, 12 (Dec. 2008), 68–74. https://doi.org/10.1145/1409360.1409378

[6] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artif. Intell.* 165, 1 (June 2005), 91–134.

[7] Gabe Fierro, Sriharsha Guduguntla, and David E. Culler. 2019. Dataset: An Open Dataset and Collection Tool for BMS Point Labels. In *Proceedings of the 2nd Workshop on Data Acquisition To Analysis* (New York, NY, USA) *(DATA'19)*. Association for Computing Machinery, New York, NY, USA, 40–42. https://doi.org/10.1145/3359427.3361922

[8] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2016. Manga109 Dataset and Creation of Metadata. In *Proceedings of the 1st International Workshop on CoMics ANalysis, Processing and Understanding* (Cancun, Mexico) *(MANPU '16)*. Association for Computing Machinery, New York, NY, USA, Article 2, 5 pages. https://doi.org/10.1145/3011549.3011551

[9] R. V. Guha, Dan Brickley, and Steve Macbeth. 2016. Schema.Org: Evolution of Structured Data on the Web. *Commun. ACM* 59, 2 (Jan. 2016), 44–51. https://doi.org/10.1145/2844544

[10] Alon Halevy, Flip Korn, Natalya F. Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. 2016. Goods: Organizing Google's Datasets. In *Proceedings of the 2016 International Conference on Management of Data* (San Francisco, California, USA) *(SIGMOD '16)*. Association for Computing Machinery, New York, NY, USA, 795–806. https://doi.org/10.1145/2882903.2903730

[11] Mathias Konrath, Thomas Gottron, Steffen Staab, and Ansgar Scherp. 2012. SchemEX - Efficient Construction of a Data Catalogue by Stream-Based Indexing of Linked Data. *Web Semant.* 16 (Nov. 2012), 52–58. https://doi.org/10.1016/j.websem.2012.06.002

[12] Babak Loni, Lei Yen Cheung, Michael Riegler, Alessandro Bozzon, Luke Gottlieb, and Martha Larson. 2014. Fashion 10000: An Enriched Social Image Dataset for Fashion and Clothing. In *Proceedings of the 5th ACM Multimedia Systems Conference* (Singapore, Singapore) *(MMSys '14)*. Association for Computing Machinery, New York, NY, USA, 41–46. https://doi.org/10.1145/2557642.2563675

[13] Hatem Mousselly-Sergieh, Daniel Watzinger, Bastian Huber, Mario Döller, Elöd Egyed-Zsigmond, and Harald Kosch. 2014. World-Wide Scale Geotagged Image Dataset for Automatic Image Annotation and Reverse Geotagging. In *Proceedings of the 5th ACM Multimedia Systems Conference* (Singapore, Singapore) *(MMSys '14)*. Association for Computing Machinery, New York, NY, USA, 47–52. https://doi.org/10.1145/2557642.2563673

[14] Nikolay Nikolov, Michele Ciavotta, and Flavio De Paoli. 2018. Data Wrangling at Scale: The Experience of EW-Shopp. In *Proceedings of the 12th European Conference on Software Architecture: Companion Proceedings* (Madrid, Spain) *(ECSA '18)*. Association for Computing Machinery, New York, NY, USA, Article 32, 4 pages. https://doi.org/10.1145/3241403.3241437

[15] Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. 2009. SOFIE: A Self-Organizing Framework for Information Extraction. In *Proceedings of the 18th International Conference on World Wide Web* (Madrid, Spain) *(WWW '09)*. Association for Computing Machinery, New York, NY, USA, 631–640. https://doi.org/10.1145/1526709.1526794