Link-Rot in Web-Sourced Multimedia Datasets

Viktor Lakic[0000−0002−8382−7606], Luca Rossetto[0000−0002−5389−9465], and Abraham Bernstein[0000−0002−0128−4602]

Department of Informatics, University of Zurich, Zurich, Switzerland

Abstract. The Web is increasingly used as a source for content of datasets of various types, especially multimedia content. These datasets are then often distributed as a collection of URLs, pointing to the original sources of the elements. As these sources go offline over time, the datasets experience decay in the form of link-rot. In this paper, we analyze 24 Web-sourced datasets with a combined total of over 270 million URLs and find that over 20% of the content is no longer available. We discuss the adverse effects of this decay on the reproducibility of work based on such data and make some recommendations on how they could be mediated in the future.

Keywords: Link Rot · Dataset Rot · Online Datasets · Reproducibility.

1 Introduction

Multimedia datasets enjoy increasing popularity, largely driven by the ever increasing need for training data of large machine learning models. The Web offers a convenient source from which to compile such datasets. Such collections of Web content are then often distributed as a list of URLs, directly pointing to the original source of the individual contained content elements. While this distribution method is convenient, it comes with a substantial drawback. Since commonly, the authors of a collection are not the authors of its content, they have no control over the availability of the contained elements. When individual elements become unavailable, the links contained within the collection break and the collection degrades, calling into question the reproducibility of any results obtained by using the dataset.

In this paper, we analyze the link-rot of Web-sourced datasets of various types. To do this, we query the original sources of 24 different datasets published between 2009 and 2022 and observe the overall success rate as well as the different error responses. We find that of the roughly 270 million URLs queried, over 20% do no longer return the expected content.

The remainder of this paper is structured as follows: after providing a brief overview of related work in Section 2, Section 3 introduces the 24 different datasets that were used in this study. Section 4 outlines the methodology used for the analysis of the datasets and their online availability before Section 5 details the results. We offer some concluding remarks in Section 6.

2 Related Work

Hyperlinks that do not or no longer point to their intended resource are a well-known occurrence in the Web. While at the very least a source of annoyance to the everyday Web-user, the impact of such 'broken' links can be more far-reaching. The Hiberlink project [22] studied the extent of the preservation of linked online resources in scholarly publications. Zhou et al. [32] and Klein et al. [13] found that already in 2014, over 20% of surveyed scholarly articles included Web links and that this rate appeared to be increasing. Experiments presented in [31] estimate that roughly 36% of all URLs contained in published scientific papers no longer point to their intended resource.

Also, outside of scholarly publications, the decay of online references has undesirable consequences. Dividino et al. [7] observe increasing error responses across a large number of linked open data endpoints over time and Zittrain et al. [34] find that 50% of URLs in US Supreme Court opinions no longer point to their original content. As several of these studies only consider `400` and `500` HTTP responses as rotten links, it is safe to assume that the actual number is probably even higher as many sites use 'soft-404s', which respond with a `200 OK` code but display a more user-friendly `404` error message [16].

While several suggestions and recommendations have been made to address these issues over the years [3,21], the problem still persists.

With the increased activities in the area of machine learning research in recent years, the need for ever more training data has become apparent and the Web offers a convenient source of such data in various modalities. Early Web-sourced multimedia datasets such as ImageNet [6], have been dwarfed by more recent collections such as LAION-400M [24] or LAION-5B.[1] The latter are exclusively distributed as a collection of URLs to the individual image sources, since both the size of the collection as well as the uncertain licensing of the individual contained items makes an other form of distribution inconvenient.

There has so far been no comprehensive analysis on the rate of (link-)rot experienced by such collections and its consequence for reproducibility of work based on them.

3 Datasets

For our analysis, we use 24 different Web-sourced datasets originally released for different purposes and containing different kinds of documents. The complete list of all considered Datasets is shown in Table 1.

The age range of the considered datasets spans roughly 13 years, the oldest one being from 2009 while the most recent one was published in 2022. The number of URLs per dataset ranges from 10^4 to 10^9, with the consequence that the two largest datasets together account for 81% of all URLs. Since our focus is on multimedia datasets, 15 of them contain references to videos, 6 contain images,

[1] https://laion.ai/blog/laion-5b/

Table 1. Datasets Overview sorted by number of URLs

Dataset	#URLs	Hosts	Content	Release
YouTube Speakers [23]	1‘111	YouTube	Video	2013
Evoked Expressions in Video [28]	5‘155	YouTube	Video	02.2021
RealEstate10K [33]	7‘255	YouTube	Video	05.2018
Columbia Consumer Video [11]	9‘317	YouTube	Video	04.2011
ActivityNet [8]	19‘994	YouTube	Video	06.2015
V3C [20]	28‘450	Vimeo	Video	10.2018
IACC [17]	39‘937	Archive.org	Video	2010
What’s Cookin’ [15]	47‘935	YouTube	Video	03.2015
PS-Battles [10]	102‘028	Mixed	Image	04.2018
Document Similarity Triplets [5]	120‘515	Wikipedia arxiv.org	Other	2014
NUS-WIDE [4]	257‘789	Flickr	Image	07.2009
YouTube-BoundingBoxes [18]	285‘410	YouTube	Video	02.2017
Kinetics 700-2020 [27]	643‘459	YouTube	Video	10.2020
CORD-19 [30]	1‘122‘751	Mixed	Other	03.2022
Sports-1M [12]	1‘133‘158	YouTube	Video	04.2014
AudioSet [9]	2‘084‘320	YouTube	Video	06.2017
Conceptual Captions [25]	3‘333‘652	Mixed	Image	07.2018
Google Metadata for Datasets [2]	3‘602‘027	Mixed	Other	05.2020
YouTube-8M [1]	5‘410‘112	YouTube	Video	09.2016
Open Images Dataset V6 [14]	9‘178‘275	Flickr	Image	02.2020
Wikilinks [26]	10‘888‘549	Mixed	Other	10.2012
ImageNet [6]	14‘197‘119	Mixed	Image	2011
YFCC100M [29]	100‘000‘000	Flickr	Video Image	02.2016
Web Video in Numbers [19]	121‘781‘244	YouTube Vimeo	Video	07.2017
Total	274‘299‘562			
Total unique	270‘828‘632			

with one of them containing both. The remaining 4 datasets are comprised of URLs referencing other resource types and are included for comparison. Out of the 15 datasets containing references to Web video, 11 exclusively use YouTube as a host. An additional two exclusively contain references to Vimeo[2] and the Internet Archive[3] respectively. The remaining one contains references to both YouTube and Vimeo. This uniformity in hosts is an almost exclusive property of the video datasets. Of the collections of URLs linking to images, only one of them uses one host exclusively; Flickr.[4] All the other image datasets as well as all datasets referencing other content have a diverse set of hosts from all over the Web.

[2] https://vimeo.com
[3] https://archive.org
[4] https://flickr.com

Across all the analyzed datasets, there are a total of 270‘828‘632 unique URLs, which means that out of the 274‘299‘562 ones that form the naive total, 3‘470‘930 are duplicates. For the analyses presented in Section 5, we treat all datasets independently and can therefore ignore this overlap.

4 Methodology

To query the URLs, we used a custom distributed setup with a central coordination node and multiple worker instances hosted on various cloud providers. Workers would request batches of URLs from the coordinator via a simple REST API, query all URLs in the batch and report their findings back. The URLs were queried between April and June 2022 with some additional queries in August.

For Vimeo and YouTube, the workers could make use of APIs provided by the platforms in order to efficiently query the status of videos. This enabled the worker to request the status of multiple videos within one request, thereby reducing the required number of requests. Using the APIs of these platforms was reasonable since the set of possible errors reasonably to be expected from one of these platforms was limited to a well-formed response about a video no longer being available. For this analysis, we did not distinguish between semantic reasons for the unavailability of a video, such as a video being deleted by the user, regionally restricted, set to private, or removed by the platform for copyright violations.

For all other platforms, a worker would query an URL directly via a HTTP request. In case the worker could not establish a HTTP connection, it would retry at a later point in time for a maximum of three tries. If the 3rd try failed to establish any connection, it would record the type of network error (i.e., name resolution failure, connection refusal, connection timeout, etc.) as a result for that URL. In cases where a HTTP connection could be established, the worker would first check the status code. If a well-formed HTTP error is reported, the worker accepts it as a response for that URL. If a `200 OK` status is returned, the worker validates the actually returned content. It checks if the content type corresponds to the expected data (e.g., if a request to an URL ending in ‘.jpg’ has the content type ‘image/jpg’) and if the content itself is valid (e.g., if the returned image can be decoded). This is done in order to identify both inherently broken documents as well as ‘soft-404s’ which are human-readable error responses delivered with a valid status code, rendering them no longer machine readable.

Once a worker has queried all URLs of its current batch, it sends the obtained results back to the coordinator and requests a next batch. The coordinator stores all received results persistently for subsequent analysis.

5 Results

This section discusses the results obtained from the analysis of the datasets introduced in Section 3.

Table 2. Total number of available and unavailable elements per dataset

Dataset	Total	Available	Unavailable	Availability
YouTube Speakers	1‘111	1‘111	0	100.0%
PS-Battles	102‘028	101‘676	352	99.7%
Document Similarity Triplets	120‘515	119‘347	1‘168	99.0%
IACC	39‘937	38‘132	1‘805	95.5%
RealEstate10K	7‘255	6‘871	384	94.7%
Kinetics 700-2020	643‘459	600‘265	43‘194	93.3%
EEV	5‘155	4‘743	412	92.0%
YouTube-BoundingBoxes	285‘410	261‘539	23‘871	91.6%
Open Images Dataset V6	9‘178‘275	8‘237‘848	940‘427	89.8%
AudioSet	2‘084‘320	1‘804‘406	279‘914	86.6%
YFCC100M	100‘000‘000	86‘388‘275	13‘611‘725	86.4%
What’s Cookin’	47‘935	40‘905	7‘030	85.3%
Conceptual Captions	3‘333‘652	2‘845‘237	488‘415	85.3%
ActivityNet	19‘994	16‘846	3‘148	84.3%
YouTube-8M	5‘410‘112	4‘557‘852	852‘260	84.2%
Sports-1M	1‘133‘158	923‘190	209‘968	81.5%
V3C	28‘450	22‘729	5‘721	79.9%
NUS-WIDE	257‘789	205‘471	52‘318	79.7%
Google Metadata for Datasets	3‘602‘027	2‘784‘720	817‘307	77.3%
CORD-19	1‘122‘751	863‘362	259‘389	76.9%
Web Video in Numbers	121‘781‘244	90‘691‘509	31‘089‘735	74.5%
Columbia Consumer Video	9‘317	6‘893	2‘424	74.0%
Wikilinks	10‘888‘549	6‘495‘786	4‘392‘763	59.7%
ImageNet	14‘197‘119	7‘144‘800	7‘052‘319	50.3%
Total	274‘299‘562	214‘164‘681	60‘134‘881	78.1%

Table 2 shows the number of available and unavailable elements for each dataset as well as the fraction of availability. The mean-average availability across all datasets is 78.1%, meaning 21.9% of all queried URLs did not return the content referenced when the dataset was compiled. Out of the 24 tested datasets, only one is still completely available from the original sources. The most ‘link-rotten’ dataset is ImageNet, which is also among the oldest among the tested ones. In the roughly 11 years between its publication and our analysis, almost half of its original sources became unavailable.

While the age of a dataset is certainly a contributing factor to the amount of Link-Rot it shows, it lacks clear explanatory power. This can be seen in Figure 1 which shows the amount of rot per dataset on the vertical and the time of its release on the horizontal axis. In this figure, the three datasets that are either composed of one type of content from two hosts or two content types from one host (i.e., Document Similarity Triples, YFCC100M, and Web Video in Numbers) are split in two to consider each host or type separately. The correlation between the age of an observed dataset and the amount of rot it experiences is 0.19 ($p = 0.348$). When only considering the video datasets sourced from YouTube,

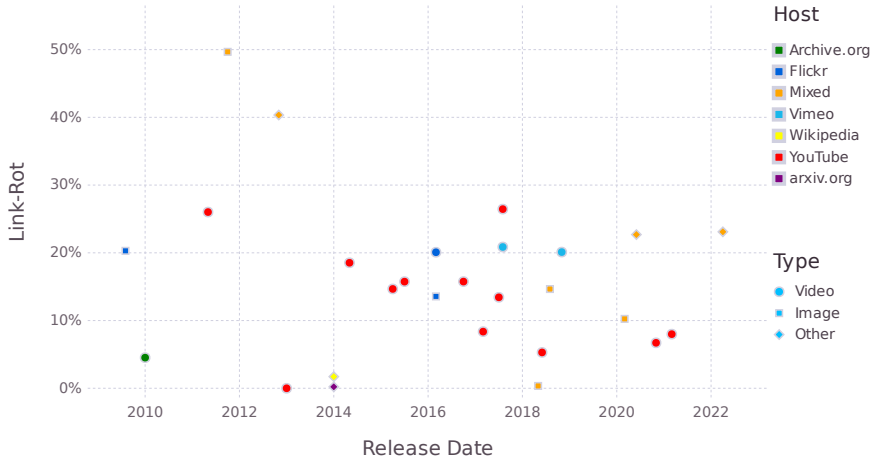

Fig. 1. Amount of Link-Rot relative to Dataset Release Date by Type and Host

which from the largest subset, the correlation grows to 0.35 ($p = 0.257$). The two datasets with the worst availability are both sourced from the Web at large and are both roughly a decade old at the time of sampling. Their availability is much lower when compared to other datasets of similar age that were sourced from one single hosting platform. It therefore stands to reason that the general decay rate in the Web might be higher compared to that of a dedicated content hosting platform (correlation: 0.73, $p = 0.064$). This makes intuitive sense, since a dedicated platform might have a stronger interest to guard against unintended content loss.[5] In contrast to more broadly sourced collections, datasets sourced from only one source do however have a single point of failure, as they would become completely unavailable if the one host they are sourced from would become permanently unavailable (however unlikely this might be for these particular platforms).

In contrast to the video hosting platforms, the sources used for the document and image datasets can all be delivered with a simple HTTP requests and are hence easier to obtain. Since communication is happening directly via HTTP rather than some platform-specific mechanism, a different set of failure states needs to be handled in case the request does not return the expected content. Table 3 shows the different errors that the workers encountered when querying the content of the various datasets.

The first failure case listed in the table consists of a request returning a well-formed HTTP response containing *invalid content*. This predominantly oc-

[5] This is to be seen as a general trend rather than a definite insight, as due to lack of release dates for individual URLs and the resulting crudity of the analysis, none of the correlations pass any reasonable threshold for statistical significance.

curs in the form of so-called 'soft-404s' which is a human-readable error page delivered with a `200 OK` status code. This category also include requests that return incomplete or otherwise unreadable files. These kinds of invalid responses can often be observed in *Conceptual Captions* and *ImageNet* which are both composed of images sourced from all over the Web.

The next category of problems encompasses everything happening one or more layers below HTTP, resulting in no valid HTTP connection being established. This includes connection refusals, connection timeouts, name resolution errors, etc. These errors predominantly occur when querying datasets from a large range of hosts.

The remaining categories consist of various well-formed HTTP errors. These can primarily be divided into two groups: client-side errors (`HTTP 4XX`) and server-side errors (`HTTP 5XX`). All error codes outside of these ranges are grouped into the *Other* column in this table. Those are largely composed of status codes that are not officially defined in the HTTP specification.[6]

Among the client side errors, the `404 Not Found` is the most common one, which is to be expected, as it is the appropriate reply for a request for a resource that does not exist. This is in contrast to `410 Gone`, which describes a resource that was valid once but no longer exists. Other common responses in this category include `401 Unauthorized` and `403 Forbidden`, both dealing with requests for which additional permissions would be required, as well as `400 Bad Request` that here appears to be used as a more generic catch-all error response. Several other error codes in the 400-range were observed as well, many of them lacking an official definition.[7]

In the remaining category of server-side errors, the most commonly observed is the generic `500 Internal Server Error`. Other commonly observed errors from this category include `502 Bad Gateway`, `503 Service Unavailable`, and `504 Gateway Timeout`. The less commonly observed status codes include a wide range, including many for which no official definition exists.[8]

An obvious solution to the decay of a dataset though link-rot is to produce an archived copy of all of its content, before it has a chance to decay. To do this, it is however necessary that all content elements of a dataset are licensed in such a way that allows their redistribution. Table 4 shows that out of the analyzed datasets, only 5 exclusively contain content that comes with such permissive licenses and for only 8 of the datasets, we were able to confirm that archival copies independent from the contents original source are available.

[6] Observed status codes below 400 and above 600: 101, 300, 301, 302, 303, 304, 307, *600*, *617*, *651*, *670*, *724*, *750*, *903*, *999*. Italicized codes have no generally accepted definition.

[7] Observed error codes between 400 and 500: 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 412, 414, 415, 416, 417, 418, *419*, 420, 421, 422, 423, 424, 426, 429, 444, *445*, *447*, 449, *451*, *456*, *463*, *465*, *470*, *471*, *473*, *477*, *478*, *479*, *490*, *493*, *498*. Italicized codes have no generally accepted definition.

[8] Observed error codes between 500 and 600: 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 511, *512*, *520*, *521*, *522*, *523*, *524*, *525*, *526*, *529*, *530*, *533*, *534*, *535*, *543*, *555*, *556*, *567*, *591*. Italicized codes have no generally accepted definition.

Table 3. Distribution of reason for missing elements from Image and Document datasets. Most relevant HTTP error status codes are listed explicitly, others are shown in aggregate.
Invalid Content counts instances of valid HTTP responses that delivered either invalid or incomplete documents or documents of a different datatype. *No HTTP Connection* counts instances where no HTTP connection could be established to the remote server. Relevant error codes include: `400 Bad Request`, `401 Unauthorized`, `403 Forbidden`, `404 Not Found`, `410 Gone`, `500 Internal Server Error`, `502 Bad Gateway`, `503 Service Unavailable`, and `504 Gateway Timeout`.
Other codes include both officially defined HTTP status codes as well as such for which no official definition and agreed-upon semantics exists.

Dataset	Unavailable	Invalid Content	No HTTP Connection	HTTP 4XX						HTTP 5XX					Other
				400	401	403	404	410	Other	500	502	503	504	Other	
PS-Battles	352	30	36	2	1	50	197	2	12	8	2	7	0	5	0
Document Similarity Triplets	1‘168	128	0	0	0	0	1‘040	0	0	0	0	0	0	0	0
IACC	1‘805	1	86	0	0	91	101	0	0	1‘526	0	0	0	0	0
NUS-WIDE	52‘318	0	0	0	0	0	35‘141	17‘177	0	0	0	0	0	0	0
Conceptual Captions	488‘415	66‘228	113‘993	13‘933	3‘666	63‘818	187‘610	3175	23‘549	5‘467	1420	4‘669	25	759	9
CORD-19	259‘389	0	10‘967	0	7	176‘583	5‘618	3‘452	2‘142	124	8	60‘488	0	0	0
Google Metadata for Datasets	817‘307	0	49‘611	254	1	540‘485	169‘119	5‘613	49‘189	886	51	2‘098	0	0	0
Open Images Dataset V6	940‘427	1	13	0	0	0	734‘184	206‘171	0	19	39	0	0	0	0
Wikilinks	4‘392‘763	0	1‘486‘965	12‘289	7‘351	679‘336	1‘605‘625	132‘304	356‘554	50‘951	3‘763	35‘848	902	17‘841	1‘659
ImageNet	7‘052‘319	720‘890	1‘785‘527	27‘642	4‘635	462‘426	3‘377‘523	406‘918	181‘403	29‘716	5‘784	40‘977	2‘291	4‘874	347
YFCC100M	13‘611‘725	3	2‘586	0	0	1	9‘860‘215	3‘746‘659	0	294	1‘966	0	1	0	0

Table 4. Availability of dataset content under a permissive license and availability of a complete copy of the linked content. *Google Metadata for Datasets*, *Web Video in Numbers*, and *Wikilinks* are omitted from this table, as these datasets are only concerned with the links themselves and not the actual content they point to.

Dataset	Permissive licence	Copy available
ActivityNet		✓
AudioSet		
Columbia Consumer Video		
Conceptual Captions		
CORD-19		✓
Document Similarity Triplets	✓	
Evoked Expressions in Video		
IACC	✓	✓
ImageNet		✓
Kinetics 700-2020		✓
NUS-WIDE		
Open Images Dataset V6	✓	✓
PS-Battles		
RealEstate10K		
Sports-1M		
V3C	✓	✓
What's Cookin'		
YFCC100M	✓	✓
YouTube Speakers		
YouTube-8M		
YouTube-BoundingBoxes		

6 Conclusion

We presented an analysis of 24 Web-sourced datasets and found that of the combined 270 million URLs, over 20% no longer point to the intended content. Only one of the 24 datasets was unaffected by this link-rot while the most affected has decayed by almost half. We found that datasets sourced from the Web at large experience the largest amount of link-rot and the greatest diversity in observed errors during querying. While the amount of link-rot of a dataset can only increase over time, age is only one factor among several. Only for a small subset of the studied datasets were we able to confirm that a complete archival copy exists and is available. Even fewer datasets are exclusively composed of content that would easily allow for such re-distribution.

Since it is not foreseeable that the need for diverse datasets of various modalities will decrease, it is to be expected that the observed link-rot will have increasingly adverse effects of the reproducibility of scientific work that makes use of such data. Authors of such datasets, especially when sourcing the content from the Web, should therefore take care that the content of their collections is licensed in such a way that enables re-distribution. Further, they should endeavor

to keep a complete reference copy of all the content of their dataset and make at least those parts of it available upon request that can no longer be obtained from the original sources. In cases where the collections become too large to be easily handled, it might become necessary to employ distributed storage solutions so that the ones interested in using a dataset can contribute to its hosting, thereby ensuring future availability while sharing the incurred resource requirements.

Acknowledgements This work has been partially supported by the Swiss National Science Foundation, Project "MediaGraph" (Grant Number 202125).

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. CoRR **abs/1609.08675** (2016), http://arxiv.org/abs/1609.08675
2. Brickley, D., Burgess, M., Noy, N.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: The World Wide Web Conference. p. 1365–1375. WWW '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3308558.3313685
3. Burnhill, P., Mewissen, M., Wincewicz, R.: Reference rot in scholarly statement: threat and remedy. Insights **28**(2) (2015). https://doi.org/10.1629/uksg.237
4. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.T.: Nus-wide: A real-world web image database from national university of singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval. CIVR '09, Association for Computing Machinery, New York, NY, USA (July 8-10, 2009). https://doi.org/10.1145/1646396.1646452
5. Dai, A.M., Olah, C., Le, Q.V.: Document embedding with paragraph vectors. In: NIPS Deep Learning Workshop (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009). https://doi.org/10.1109/CVPR.2009.5206848
7. Dividino, R.Q., Kramer, A., Gottron, T.: An investigation of HTTP header information for detecting changes of linked open data sources. In: The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers. Lecture Notes in Computer Science, vol. 8798, pp. 199–203. Springer (2014). https://doi.org/10.1007/978-3-319-11955-7_18
8. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (June 2015). https://doi.org/10.1109/CVPR.2015.7298698
9. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE, New Orleans, LA (2017). https://doi.org/10.1109/ICASSP.2017.7952261

10. Heller, S., Rossetto, L., Schuldt, H.: The ps-battles dataset – an image collection for image manipulation detection. CoRR **abs/1804.04866**, arXiv:1804.04866 (Apr 2018). https://doi.org/10.48550/ARXIV.1804.04866
11. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: Proceedings of ACM International Conference on Multimedia Retrieval (ICMR), oral session. pp. 1–8. ICMR '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1991996.1992025
12. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. pp. 1725–1732. IEEE Computer Society (2014). https://doi.org/10.1109/CVPR.2014.223
13. Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., Tobin, R.: Scholarly context not found: One in five articles suffers from reference rot. PLOS ONE **9**(12), 1–39 (12 2014). https://doi.org/10.1371/journal.pone.0115253
14. Krasin, I., et al.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. (2017)
15. Malmaud, J., Huang, J., Rathod, V., Johnston, N., Rabinovich, A., Murphy, K.: What's cookin'? interpreting cooking videos using text, speech and vision (2015). https://doi.org/10.48550/ARXIV.1503.01558
16. Meneses, L., Furuta, R., Shipman, F.: Identifying "soft 404" error pages: Analyzing the lexical signatures of documents in distributed collections. In: Theory and Practice of Digital Libraries - Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7489, pp. 197–208. Springer (2012). https://doi.org/10.1007/978-3-642-33290-6_22
17. Over, P., Awad, G., Smeaton, A.F., Foley, C., Lanagan, J.: Creating a web-scale video collection for research. In: Proceedings of the 1st workshop on Web-scale multimedia corpus. p. 25–32. WSMC '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1631135.1631141
18. Real, E., Shlens, J., Mazzocchi, S., Vanhoucke, V., Pan, X.: Youtube-boundingboxes: A large high-precision human-annotated dataset for object detection in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7464–7473 (July 2017). https://doi.org/10.48550/ARXIV.1702.00824
19. Rossetto, L., Schuldt, H.: Web video in numbers - an analysis of web-video metadata (2017). https://doi.org/10.48550/ARXIV.1707.01340
20. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C - a Research Video Collection. In: International Conference on Multimedia Modeling. pp. 349–360. Springer, Springer, Heidelberg, Germany (Januar 2019). https://doi.org/10.1007/978-3-030-05710-7_29
21. Sanderson, R., Phillips, M., de Sompel, H.V.: Analyzing the persistence of referenced web resources with memento. CoRR **abs/1105.3459** (2011), http://arxiv.org/abs/1105.3459
22. Sanderson, R., Van de Sompel, H., Burnhill, P., Grover, C.: Hiberlink: Towards time travel for the scholarly web. In: Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts. p. 21. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2499583.2500370

23. Schmidt, L., Sharifi, M., Lopez-Moreno, I.: Large-scale speaker identification. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1650–1654. IEEE (2014). https://doi.org/10.1109/ICASSP.2014.6853878
24. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: open dataset of clip-filtered 400 million image-text pairs. CoRR **abs/2111.02114** (2021), https://arxiv.org/abs/2111.02114
25. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1238, https://aclanthology.org/P18-1238
26. Singh, S., Subramanya, A., Pereira, F., McCallum, A.: Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Tech. Rep. UM-CS-2012-015, University of Massachusetts, Amherst (2012)
27. Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., Zisserman, A.: A short note on the kinetics-700-2020 human action dataset (2020). https://doi.org/10.48550/ARXIV.2010.10864
28. Sun, J.J., Liu, T., Cowen, A.S., Schroff, F., Adam, H., Prasad, G.: Eev: A large-scale dataset for studying evoked expressions from video. arXiv preprint arXiv:2001.05488 **abs/2001.05488** (2021). https://doi.org/10.48550/ARXIV.2001.05488
29. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (jan 2016). https://doi.org/10.1145/2812802
30. Wang, L.L., et al.: Cord-19: The covid-19 open research dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Association for Computational Linguistics (Jul 2020), https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1
31. Zhou, K., Grover, C., Klein, M., Tobin, R.: No more 404s: Predicting referenced link rot in scholarly articles for pro-active archiving. In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. p. 233–236. JCDL '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2756406.2756940
32. Zhou, K., Tobin, R., Grover, C.: Extraction and analysis of referenced web links in large-scale scholarly articles. In: IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014. pp. 451–452. IEEE Computer Society, New York, NY, USA (2014). https://doi.org/10.1109/JCDL.2014.6970220
33. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. ACM Trans. Graph. (Proc. SIGGRAPH) **37**(4), 1–12 (07 2018). https://doi.org/10.1145/3197517.3201323
34. Zittrain, J., Albert, K., Lessig, L.: Perma: Scoping and addressing the problem of link and reference rot in legal citations. Legal Information Management **14**(2), 88–99 (2014). https://doi.org/10.1017/S1472669614000255