Creating A Web-Scale Video Collection For Research

Paul Over, George Awad
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940
USA
{over, gawad}@nist.gov

Alan F. Smeaton, Colum Foley, James Lanagan
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Glasnevin, Dublin 9, Ireland
{alan.smeaton, colum.foley, jlanagan}@computing.dcu.ie

ABSTRACT

This paper begins by considering a number of important design questions for a large-scale, widely available, multimedia test collection intended to support long-term scientific evaluation and comparison of content-based video analysis and exploitation systems. While the collection presented here is not quite web-scale, it is to our knowledge the largest video collection created to date. It is therefore of use in expanding the scale of any evaluation of multimedia collections and systems. Such exploitation systems would include the kinds of functionality already explored within the annual TREC Video Retrieval Evaluation (TRECVid) benchmarking activity such as search, semantic concept detection, and automatic summarization. We then report on our progress in creating such a multimedia collection from publicly available Internet Archive videos with Creative Commons licenses (IACC.1), which we hope will be a useful approximation of a web-scale collection and will support a next generation of benchmarking activities for content-based video operations. We also report on some possibilities for putting this collection to use in multimedia system evaluation. It is the intended that this collection be partitioned and used within the TRECVid 2010 evaluations, and in subsequent years to that.

Categories and Subject Descriptors

H.5.1 [**Information Systems**]: Information Systems Applications—*Miscellaneous*

General Terms

Experimentation,Measurement,Standardization

Keywords

Evaluation, Benchmarking, Video Retrieval

1. INTRODUCTION

Experience across a variety of information seeking fields indicates that good test collections can promote progress in related technologies. For example, the Text Retrieval Conference (TREC) text collections have promoted progress in information retrieval, the National Institute of Standards and Technology's speech collections for automatic speech recognition, the TREC Video Retrieval Evaluations (TRECVid) collections for multimedia segmentation, feature detection, search, and summarization [4] and so on for activities such as ISMIR (International Society for Music Information Retrieval), CLEF (Cross-Language Evaluation Forum) and INEX (INitiative for the Evaluation of XML Retrieval). While the various test collections, associated with different benchmarking activities like these, have all served useful purposes, they are all challenged by issues of size and of the relevancy of their data to what people are actually searching among. Thus truly useful large and complex multimedia collections are, for a variety of reasons, difficult to develop and so are understandably rare.

At its core a good test collection for information retrieval research will be composed of "found objects" in the sense that these objects were produced as a whole for some other purpose, e.g., materials associated with the US tobacco company suits, or a crawl of the Internet, or self-selected videos for sharing, yet despite this they are useful for research. Understanding the nature of such objects may require a lot of work. In addition, most test collections are to some extent designed and then artificially fabricated, for example in the way the found objects are somehow chosen to become components of the collection. Good test collection design should start from some idea of how the collection will be used and the same understanding will be needed as one tries to evaluate the adequacy of a collection's "found" characteristics.

This paper considers a number of important design questions for a large-scale, widely available, multimedia collection intended to support long-term scientific evaluation and comparison of video analysis and content-based exploitation systems. We report on our efforts and progress in creating such a collection and how we intend to put it to use. As previously stated, this collection is aimed as the TRECVid community and takes the role of a large if not quite web-scale multimedia collection.[1]

WSMC'09, October 23, 2009, Beijing, China.

[1]Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

2. DESIGN CONSIDERATIONS

2.1 Requirements

Good test collection design, indeed any kind of good design, begins with an understanding of requirements, which in turn are dependent on the kinds of research questions which might be answered when using the test collection. The first requirement we consider is of size. At a high level, the unconstrained notion of "web-scale" suggests a general interest in applications that will be able to handle whatever the Internet holds, which we know is a moving target but has numbers of videos in the billions without limitation to particular data sources, data types, production qualities, or content areas. A recent press release from 28th April, 2009 from comScore Inc. estimates there were 14.5 billion Internet videos viewed during March 2009 by US Internet users alone [1]. The official YouTube blog dated 20th May 2009 reports videos are being uploaded to YouTube at a rate of 20 hours of video per minute [6]. If such videos average 3 minutes in duration that means that just YouTube alone is already growing by more than a half a million videos per day.

Given these huge volumes of video data, when considering what "web-scale" means in the context of a test collection of video we must concede that practical considerations will require video test collections to be a sample of the population of interest and to be of a size which is perhaps the smallest sample large enough to reliably test algorithms that will then scale to the full population size. It follows from the size and diversity of a real-world collection of web video that even collection-spanning *descriptions* are possible only at a very high level. The issue of test collection size may never be completely resolved.

More detailed requirements of the test collection should be based on the specific application tasks to be tested and the associated use-cases based on *real* user scenarios. We do not attempted to specify exact use-cases for the collection presented here, as this task is one for the community of users itself. TRECVid remains a user driven experience, and so the tasks which are to be performed come from the community itself. We should thus ask what evidence is available about how real-world web users currently search for and use web video for entertainment, education, and sharing; how do people use web video for more professional reasons such as security or even law enforcement ? A recent press release from comScore, Inc. [1] reports the following about US Internet users based on findings from March 2009. This gives us some raw information about what is happening and how people access web video. It gives no details however on why people are doing things in the ways they are.

* 77.8% of the total U.S. Internet audience viewed online video during March 2009 alone.

* The average online video viewer watched 327 minutes of video, or nearly 5.5 hours, during March 2009.

* 99.7 million viewers watched 5.9 billion videos on YouTube.com (59.1 videos per viewer) during the month of March 2009.

* 47.4 million viewers watched 349 million videos on MySpace.com (7.4 videos per viewer).

* Hulu accounted for 2.6 percent of videos viewed, but 4.9 percent of all minutes spent watching online video.

* The duration of the average online video was 3.4 minutes.

Can we learn anything from what seems to be a mixed message, apart from the fact that both usage and content are growing ? What we do know about video is that it provides a useful core for a *multimedia* collection because video contains images (albeit in motion) and often there is associated human speech, there can be natural environmental sounds in the background, there can be scene and overlay text, for broadcast video there is often closed captioning text, and on top of all that content material there can be metadata about content, production, viewer reactions and ratings, reviews, etc. All this material also forms core content for web video so this should be considered when designing and building web video test collections. Yet test collections should also be designed to accommodate the many types of derived data which we can get from video such as shot segmentation, low-level feature/object/event information, user annotations, transcripts and even language translations of automatically recognized speech, high-level semantic feature/object/event annotations, video summaries, and of course queries with relevance judgments.

Test collections in any domain, including IR, can only promote progress in the field if they are widely used and long-lived. Progress in fields such as IR happens over time and is more likely the greater the number of researchers can make use of the collection. Widespread use requires the collection be generally available in compliance with intellectual property considerations at an affordable cost/effort. While stability over time is needed for comparison of results, provision for expansion should also be made so that the collection continues to represent the population of data of interest, and can accommodate new tasks as they emerge from the population of users, and their testing needs.

We now describe our efforts to create a multimedia test collection which meets many of the needs described above.

3. INTERNET ARCHIVE CREATIVE COMMONS VIDEO (IACC.1)

This section describes the way in which the considerations mentioned above have been addressed so far in the creation of the Internet Archive Creative Commons (IACC.1) Video collection — a snapshot of videos publicly available from the Internet Archive (IA) under Creative Commons (CC) licenses as of May 2009.

The Creative Commons is *"a nonprofit corporation dedicated to making it easier for people to share and build upon the work of others, consistent with the rules of copyright"* [2]. The CC offers 6 pre-defined licenses, which vary in the degree to which they restrict use of the licensed materials but facilitate an owner's desire to make permission for sharing of the material and its reuse explicit. CC licenses are thus very compatible with research-only use.

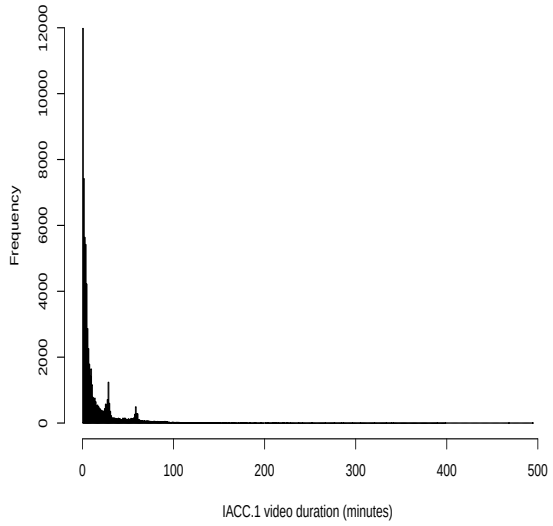
The IACC.1 video is taken from the Internet Archive (IA), which was founded in 1996 as a non-profit organization *"with the purpose of offering permanent access for researchers, his-*

torians, and scholars to historical collections that exist in digital format" [3]. The IA's collections have grown over the years and now include moving images as well as text, audio, etc. The IA grew originally by incorporating whole collections of video (e.g. the Prellinger Collection) but the IA now also accepts individual donated videos.

3.1 IACC.1 - what do we know about it?

The IACC.1 collection was crawled from the Internet Archive, both the video and the associated metadata, and contains about 64 000 color and black&white videos whose durations total about 13 584 hours. At the time of writing we are in the process of dividing the collection into color and non-color. Original video formats vary depending on the video donation and these are preserved by the IA but the IACC.1 includes only the derived versions in MPEG-4, encoded using H.264 and a bit rate of 512 Kb/s. The video occupies about 3.4 TB. It is expected that the full IACC.1 collection would be divided into subsets for use over several years.

Figure 1: Frequency distribution of IACC.1 video durations



The IACC.1 videos have a median duration of 4.343 minutes, a mean duration of 12.720 minutes and a maximum durations of 494.1 minutes. Figure 1 shows the distribution of durations.

The source and content of the collection are diverse, and would be impossible to describe in a detailed and complete way. About 40 % (by duration) of the collection comes from what the IA calls its "opensource_movies" collection. Another 21 % comes from the its "bliptv" collection. The remaining 39 % comes from 125 different collection sources:

Here are the collections contributing 1 % or more of the IACC.1's duration.

```
Percent of IACC.1 total duration
 |       Hours
 |           |        Number of videos
 |           |            | IA collection name
 |           |            | |
40.2     5457.26     26300 opensource_movies
21.4     2908.68     19237 bliptv
 8.8     1194.94      3054 opensource_religionvideo
 3.8      513.79      1007 feature_films
 3.4      460.69      1754 prelinger
 3.3      453.34      1417 george_bush_archive
 2.8      376.76      1028 classic_tv
 1.9      255.64      1029 sports
 1.2      160.16        66 us_congress
 1.2      158.32       538 FedFlix
 1.1      155.76       158 alternative_views
 1.1      150.71       334 thehappinessshow
```

Another 115 collections contribute less than 1 % each of the IACC.1's duration. Table 1 shows the full distribution by collection.

According to the metadata for the videos in IACC.1, the following languages are represented within the content: Afrikaans, Albanian, American English, Amuzgo, Arabic, Australian English, Austrian German, Bahasa, Basque, Bavarian German, Bidbidi, Bosanski, Bosnian, Bubi, Bulgarian, Cashubian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, Estonian, Farsi, Finnish, French, Galician, German, Greek, Hebrew, Hindi, Hungarian, Hungarian., Indian English, Indonesian, Irish Gaelic, Italian, Japanese, Khmer, Korean, Kurdish, Latin, Malay, Mandarin Chinese, Mexican Spanish, Ninguno, Norwegian, Osbatansa, Polish, Portuguese, Punjabi, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Swiss German, Tagalog, Tami, Thai, Turkish, UK English, Vietnamese, and some combinations of the foregoing.

Watching even a tiny sample of this large collection of video shows a variety of content e.g., conference talks, school TV, religion, animation, performance art, party videos, amateur sports, documentary, nature, produced news, amateur news, martial arts, infomercial, musical performance, community media, video blog.

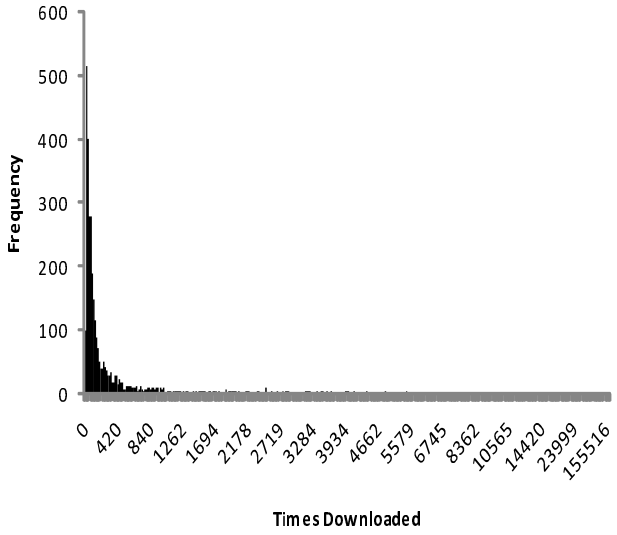
In addition to the video itself, metadata is in many cases available. The nature of the Internet Archive collection is such that the metadata available for each video tends to be quite noisy, because, for example, the content can come from small-scale donations and the donors are then mostly happy enough to just see their content online and "safe" but don't want to get involved in annotating it. Within the collection, each video has optional associated data fields which may remain empty. These fields contain information on the creator, publisher, or director of a movie, as well as information on when the video was uploaded and by whom.

Textual descriptors for the video exist in the form of titles, descriptions, and keywords or subject tags as well as occasional reviews of movies, donated for free use. The mean number of words in the video titles is 4.9 with a median of 4. Approximately 97 % of the videos in the collection also contain a description field. For some videos these descriptions can consist of hundreds of words because the annotators really got into the flow of their task, while others contain only one or two. Overall the mean number of words contained in the description is 52.1 with a median of 20.

The metadata also contains subject and keywords fields which attempt to categorize the videos. These fields, however, are not always present and often one (or both) are missing for a video. There are over 52 000 unique subjects identified in the metadata, and although there are subjects that appear across several videos there does not appear to be a predetermined vocabulary or ontology used across the collection.

From examining the raw HTML associated with each video file we can extract the download count associated with each video as shown in Figure 2. From this figure we can see that most videos in the collection have been downloaded relatively few times. The mean download count for videos is 1311.9 while the median is 95, with a maximum value of 1 137 319.

Figure 2: Frequency distribution for downloads



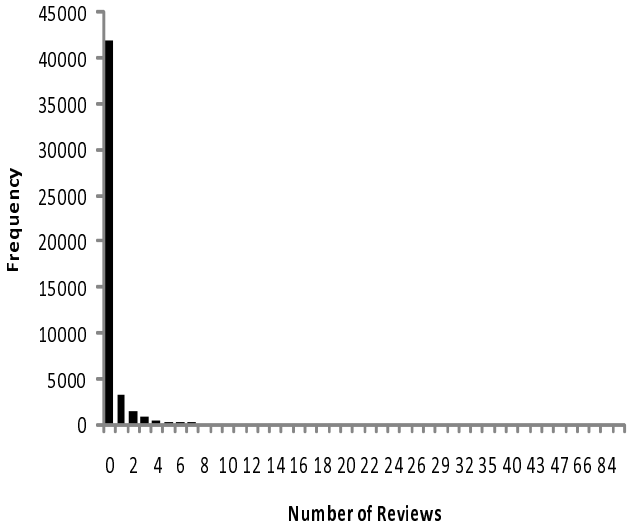
As mentioned earlier, users of the Internet Archive are able to submit reviews for videos and these reviews are stored alongside the videos in the metadata. Again the distribution of reviews shown in Figure 3 reflects a long-tail with the vast majority of videos having no reviews. For those that have reviews the mean number of words in these reviews is 81.5 while the median is 53. There are over 6 000 reviewers identified in the collection.

We have not performed any more statistical tests on the corpus as a whole since the corpus itself will not be used in its entirety. Measurements such as entropy and standard deviations of video length etc. would not be of use as it is the intention of the authors that this corpus be divided up into different portions. These portions themselves may be based on characteristics of the containing video, though as we have already mentioned this is unknown at this time.

3.2 IACC.1 — how could it be used?

Given the base data, it is interesting to speculate about real tasks, abstractions of which systems might be tested on using the IACC.1 in a laboratory setting. Three possibilities come to mind.

Figure 3: Frequency distribution for reviews



It follows from the difficulty of knowing in any detail what a very large and diverse collection contains that a useful task might be a rough categorization of the video. The results of such processing could be used to filter out certain sorts of video, to route videos based on category to different downstream applications, or support browsing or search for an interactive user. The primary difficulty here is understanding the nature of the categories needed for a particular application.

At a lower level of detail, it could be useful for users (searching, browsing) or for downstream applications (categorizing, filtering, routing) to know some of the attributes or features of a video or video segment. A feature detection task can model this need. In this task, for a pre-defined set of features, the system must determine the value of each feature for a given video or video segment. Feature values can be binary (feature present; feature not present) or more complex. Features can vary from low-level (e.g., color, texture, edge) to high-level people, objects, locations, and events. Current work on feature detection suggests that results are largely dependent not just on the number of training examples but on the similarity of training examples and actual instances in the test data[5]. As a result the large variation with which a given type of object, person, or event appears across the IACC.1 videos will likely pose major technical problems in running this task against the IACC.1.

Finally one can imagine wanting to explore the IACC.1 using search or browsing. Trying to find a video or a video segment which one believes to have seen but the name of which one does not recall is often called "known item search". Queries are created based on some knowledge of the collection such that there is a high probability that there is only one video or video segment that satisfies the search. With a very large collection, this is probably a more doable piece of work than the creation of full search queries when an estimate of the frequency of satisfying videos or video segments is needed. We plan to use IACC.1 as the basis for the

data-set to be used as part of a large, annual benchmarking activity TRECVid with over 70 research participants, commencing in 2010.

Use of the video segment, rather than a complete video file, as the unit of retrieval will increase the size of the test collection and offer a more precise answer to an interactive user or a downstream system and this is one of the search parameters which we believe to be important. Segments could be naturally occurring and automatically determined (e.g. shots) or arbitrary pre-defined units, (e.g., sequences of n frames). If alternatively systems are required to return a video and time offset additional complexities in scoring will need to be dealt with. If systems are required to return a ranked list of retrieval units believed to contain the needed known item, then scoring can be based on how close to the top of the system output list the actual known item occurs if at all.

4. SUMMING UP, NEXT STEPS

Good test collection building begins with careful upfront definition of needs based on various sorts of expected usage, even if practical considerations such as availability turn out to be the main limiting factor in what can actually be achieved. Lots of work still needs to be done to understand how a very large multimedia collection will be used, what data samples (of what size, format, content, interrelatedness, variability, changeability, etc.) are appropriate and efficient for drawing conclusions about what data populations and what system tasks. One could try to fully define the ideal collections and then go looking for them. Efforts to build the IACC.1 collection have instead started from what is available and we will use the benchmarking community as an opportunity to gather feedback on it. Next steps include learning more about the data in hand and exploring the intersection between what is there and what is needed, beginning with a few examples of tasks known to be useful from actual work situations. Only a community effort, such as from a benchmarking activity, can make real progress toward these goals.

5. ACKNOWLEDGMENTS

The authors would like to thank a number of individuals and organizations: the Internet Archive for making its collections available and in particular Cara Binder and Raj Kumar for their help in understanding the structure of the archive; the Creative Commons for support of sharing and collaboration. Alan Smeaton, Colum Foley, and James Lanagan would like to thank Science Foundation Ireland under grant number 07/CE/I1147 (CLARITY CSET).

6. REFERENCES

[1] comScore. Press Release 28. April. URL: `www.comscore.com/Press_Events/Press_Releases/2009/4/Hulu_Breaks_Into_Top_3_Video_Properties`, 2009.

[2] CreativeCommons. About the Creative Commons. URL: `creativecommons.org/about/`, May 2009.

[3] InternetArchive. About the Internet Archive. URL: www.archive.org/about/about.php, May 2009.

[4] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[5] J. Yang and A. G. Hauptmann. (un)Reliability of Video Concept Detection. In *CIVR '08: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, pages 85–94, New York, NY, USA, 2008. ACM.

[6] YouTube. Blog. URL: `www.youtube.com/blog?entry=on4EmafA5MA`, May 2009.

Table 1: IACC.1 Distribution of videos by collection

% Total. IACC.1 Hours.	Coll. Hours	Coll. Videos	Collection Name
40.2	5457.263	26 300	opensource_movies
21.4	2908.683	19 237	bliptv
8.8	1194.940	3054	opensource_religionvideo
3.8	513.795	1007	feature_films
3.4	460.690	1754	prelinger
3.3	453.338	1417	george_bush_archive
2.8	376.757	1028	classic_tv
1.9	255.639	1029	sports
1.2	160.156	66	us_congress
1.2	158.317	538	FedFlix
1.1	155.759	158	alternative_views
1.1	150.710	334	thehappinessshow
0.8	107.564	332	computerchronicles
0.7	93.046	444	iraq_middleeast
0.6	80.573	1153	digitaltippingpoint
0.5	70.367	106	SciFi_Horror
0.4	48.609	188	C64Gamevideoarchive
0.3	45.675	183	virtual_worlds
0.3	44.095	311	thecoffeehouse
0.3	43.424	496	election_2004
0.3	41.410	155	more_animation
0.3	41.309	85	Comedy_Films
0.3	38.570	94	game_replays
0.3	38.229	57	vlog_nancyboys
0.3	35.212	110	netcafe
0.2	32.652	582	universal_newsreels
0.2	32.527	1133	stock_footage
0.2	31.816	59	Tim_Leary_Archive
0.2	27.031	155	prelinger_mashups
0.2	26.315	88	conference_proceedings
0.2	24.888	96	avgeeks
0.2	22.604	189	vj_loops
0.2	22.374	26	groove_tv
0.1	19.340	83	vlog_chrisedwards
0.1	18.629	130	punkcast
0.1	18.409	32	Film_Noir
0.1	16.011	21	bbs_documentary
0.1	14.494	59	collectie_filmcollectief
0.1	14.216	18	CLUGtalks
0.1	12.822	42	vlog_icenrye
0.1	12.667	85	opensource_media
0.1	12.034	73	machinima
0.1	11.809	64	KINOfilm
0.1	11.340	39	governance
0.1	11.313	52	iraq_general
0.1	11.271	25	AlternateFocus
0.1	11.196	32	vlog_bgws
0.1	10.309	24	freespeechtv
0.1	9.028	110	thisorthat
0.1	8.458	35	iraq_911
0.1	7.729	48	mario_ajero_piano

Table 1: IACC.1 Distribution of videos by collection (continued)

% Total. IACC.1 Hours.	Coll. Hours	Coll. Videos	Collection Name
0.1	7.487	190	p2p_politics
0.1	7.348	20	election_2008
0.1	7.319	19	fadimandocumentaries
0.0	6.696	24	dumb_bunny
0.0	6.396	92	bavcYouth
0.0	6.388	52	DriveTime
0.0	6.253	14	hantslug
0.0	5.882	37	home_movies
0.0	5.453	39	classic_cartoons
0.0	5.225	5	gamefootage
0.0	5.155	18	videomisc
0.0	5.153	12	globiansfilmfestival
0.0	5.136	12	Shivbaba
0.0	5.129	55	listenup
0.0	4.922	49	Unknown
0.0	4.760	18	public_library_of_science
0.0	4.412	20	iraq_peace
0.0	4.364	45	brick_films
0.0	4.303	11	GenderVision
0.0	4.073	14	german_cinema
0.0	4.063	32	disembody
0.0	2.984	28	newsandpublicaffairs
0.0	2.899	15	TheBaySchoolofSanFrancisco
0.0	2.836	14	opensource_tv
0.0	2.576	6	vlog_fantasybedtimehour
0.0	2.486	9	cinemocracy
0.0	2.081	8	speed_runs
0.0	1.924	7	pbs...newsandpublicaffairs
0.0	1.850	21	media_burn
0.0	1.805	18	thedeadreport
0.0	1.795	86	shaping_sf
0.0	1.534	20	opensource_youthmedia
0.0	1.457	4	gamevideos
0.0	1.369	6	keepSpace4PeaceOmahaOct07...
0.0	1.219	8	pbs...artsandmedia
0.0	1.202	11	ourmedia
0.0	1.030	1	opensource_religivideo
0.0	0.981	1	talk_to_us
0.0	0.954	2	wgbhforumnetwork
0.0	0.952	7	vintage_cartoons
0.0	0.928	9	iraq_eyewitness
0.0	0.677	7	mosaic
0.0	0.618	3	opensource_pets
0.0	0.614	3	TecnologiaHechaPalabra
0.0	0.548	7	wikimedia
0.0	0.503	1	CommunityChristianChurch
0.0	0.497	3	pbs...history
0.0	0.496	3	zekesgallery
0.0	0.483	1	classic-tv
0.0	0.453	4	videogameprev
0.0	0.443	7	iraq_war

Table 1: IACC.1 Distribution of videos by collection (continued)

% Total. IACC.1 Hours.	Coll. Hours	Coll. Videos	Collection Name
0.0	0.396	2	opensource_newsvideos
0.0	0.344	1	lunarflower
0.0	0.342	9	vlog_meredith
0.0	0.318	3	pbs...spiritualityandreligion
0.0	0.289	5	iraq_tribute
0.0	0.201	1	DriveInMovieAds
0.0	0.182	1	pbs_npr_forumnetwork
0.0	0.178	1	animationandcartoons
0.0	0.174	3	pbs...healthandscience
0.0	0.152	1	test_collection
0.0	0.101	1	headphonica
0.0	0.089	3	feature_comedy
0.0	0.087	1	opensource_homevideo
0.0	0.078	1	game-art
0.0	0.066	1	academic_films
0.0	0.053	1	vlogs
0.0	0.047	2	opensource_machinima
0.0	0.031	1	foreignlanguagevideos
0.0	0.028	1	unversal_newsreels
0.0	0.021	2	opensource_audio
0.0	0.010	1	vlog_poserunning
0.0	0.004	1	guerrilla_news
0.0	0.004	1	opendemocracy
0.0	0.004	2	vlog_newwrinkle
0.0	0.000	1	opensource_films