"Description of available evaluation resources and their terms of use including data sets, ground truth, usage scenarios"

Deliverable 3.1

Distribution Level: PU

The Chorus+Project Consortium groups the following Organizations:

Partner Name	Short name	Country
JCP-Consult	JCP	FR
The French National Institute for Research in Computer Science and Control	INRIA	FR
Centre for Research and Technology Hellas - Informatics and Telematics Institute	CERTH-ITI	GR
University of Trento	UNITN	IT
Vienna University of Technology	TUWIEN	AT
University of Applied Sciences Western Switzerland	HES-SO	CH
Engineering Ingegneria Informatica SPA	ENG	IT
THOMSON	THOMSON	FR
JRC Institute for Prospective Technological Studies	JRC	EU

Document Identity

Title:	Deliverable 3.1
Subject:	Collaborative platform description and constraints for evaluation resources for audiovisual retrieval
Number:	3.1
File name:	D3_1.doc
Registration Date:	26.11.2010
Last Update:	26.11.2010

Revision History

Version	Edition	Author(s)	Date
0	1	Henning Müller	01.07.2010
Comments:	Creted the basic structure of the document with all sections		
0	2	Henning Müller	11.11.2010
Comments:	Added the various resources and characteristics of resources		
0	3	Henning Müller	19.11.2010
Comments:	Finalized the copyright questions and CreativeCommons		
0	4	Henning Müller	21.11.2010
Comments:	Proofread and finished the document		
0	5	Henning Müller	26.11.2010
Comments:	Final version with the comments of the reviewers		

Abstract

This document describes the created platform for evaluation resources for audiovisual retrieval systems, the constraints for the evaluation resources and the currently available resources. It is based on the collaborative platform for evaluation resources as part of the Chorus+ project[1]. The collaborative platform was actually built based on recommendations described in this report and is one of the goals of the Chorus+ Coordination Action.

Copyright questions of the audiovisual data sources are mentioned but no details on specific collections available via the collaborative Wiki are explained, as there are many different agreements and licenses and for some evaluation sources the copyright situation is not absolutely clear. More details are given on the Creative Common licenses that help non-law specialist to create simple and understandable license agreements.

—————————————————————

Table of Contents

---

[1] http://www.ist-chorus.org/wiki/

1. Introduction

In audiovisual information analysis and retrieval much effort is being put into creating datasets and tools for evaluating a large variety of techniques. Unfortunately, the reuse of these resources is currently not optimal as many researchers recreate new resources for their own research and do not make the created resources available. This means that time is lost for the creation for the new resources and even worse, the results of retrieval systems or tools described in the literature cannot often be reproduced by other researchers and compared to other tools. In some domains where data sharing is much more common, the data sources actually need to be made available together with a publication in a journal. This model gives the data creators the possibility to analyze the data first but also allows other researchers to subsequently work on the exactly same data and potentially improve the analysis comparing it to the original results obtained.

In audiovisual information retrieval evaluation campaigns such as TrecVid or ImageCLEF have had an influence in unifying the evaluation methodologies and particularly TrecVid has shown to have a major influence on research [2] but only on the focused areas and not on generally sharing data sets.

For the audiovisual research domain the Chorus+ project aims at creating an open platform to share various types of evaluation resources including data sets, ground truth and other tools that can be used for evaluation such as trec_eval[2] a script calculating a set of performance measures. Such a platform will help consolidate the audiovisual search domain, reduce the effort of creating new data sets and make published results better comparable to other techniques in the long run.

Within the Chorus+ project thus a collaborative Wiki[3] was created for the multimedia research community to share all types of evaluation resources and create visibility for the resources that are currently available and/or maintained by a large number of research groups. Such a platform can also help to avoid that some data sets disappear and become unavailable.

Many web pages exist with overviews of a certain number of resources and maintained usually by a single group. Very often, outdated information can be found in these pages as the maintenance might have stopped on might not be controlled anymore. The goal of the collaborative Wiki is to have a single access point to the various meta-resources and resource pages that are available. This should lower the amount of work necessary for the maintenance and make new resources visible in a much quicker way. The platform also aims at making all relevant information on evaluation sources easily available, then linking the original sources that detail the resources further.

2. Types of evaluation resources included

When starting to make evaluation resources available, several aspects of the resources need to be discussed. What is included in the term resources? We have limited our definition of evaluation resources for audiovisual retrieval to:

- **databases** of resources that are used or can be used for evaluation containing at least one media resource (image, video, music, etc.); these databases should, if possible, include tasks to perform, ground truth and thus a clear description of how to evaluate a particular type of tool with the database to make the results of several systems comparable and really show progress;

---

[2] available from http://trec.nist.giv/trec_eval/
[3] http://www.ist-chorus.org/wiki/

- **articles** on evaluation methodologies or using one of the databases; this should help to compare new algorithms or system with the existing state of the art and not only with a poorly performing baseline as it is currently common practice in publications;
- **events** for evaluation of multimedia retrieval such as TRECvid[4], ImageCLEF[5], MIREX[6] and others allowing researchers to compare their approaches with other researchers at an actual event and not only via publications;
- **evaluation tools** used for the evaluation process of retrieval systems such as scripts or trec_eval[7];
- **retrieval tools** for retrieval of information or the pre-treatment steps thus allowing groups to concentrate on particular components of the retrieval evaluation such as GIFT[8] or Lucene[9];

This definition of resources should include the most important aspect of the evaluation process and should particularly help new researchers and PhD students to quickly get familiar with a solid evaluation methodology without loosing time for creating resources. The impact of such solid evaluation and more particular the TrecVid benchmark has also been evaluated in [2].

3. Criteria for the use of evaluation resources

For each type of resources described above a certain amount of information is necessary to access and classify it on the collaborative Wiki platform. A template was created in the collaborative Wiki to assemble the most important information. With the used template, the supplied information allows various access methods to the data to be created automatically. These access methods were chosen based on the way that we assume researchers to choose or having to choose the data and tools they apply for validating their research. Not all fields are mandatory and thus allow for adding a resource even with incomplete information. This has the goal to facilitate the process of adding a resource but also has the risk of obtaining only incomplete and potentially unusable information. From the first resources added to the platform we already saw that many resources are incomplete and we started completing the incomplete resources. This should also help initiate a community process of completing resources. Table 1 details the information requested by the template. Additional fields can then be added in the description and are also described below.

Field	Description
URL	Web link to the origin of the data set with potentially more information
Domain	Domain of the images (medical, trademark, photo, etc.)
Media	Image, video, audio, etc.
Size	Size in GB
Instances	Number of instances and if application duration (for music and video)
File format	JPG, MPG, PNG, etc.
Creation date	Take where the resource was created
Task	Type of tasks addressed
Copyright	Detailed description of the copyright where available

Table 1: overview of the main fields required for adding a resource to the collaborative Wiki.

---

[4] http://trecvid.nist.gov/

[5] http://www.imageclef.org/

[6] http://www.music-ir.org/?q=node/13

[7] http://trec.nist.gov/trec_eval/

[8] http://www.gnu.org/software/gift/

[9] http://lucene.apache.org/

At the moment most of the fields are free text and not from a restricted vocabulary but it can make sense to make more concrete propositions for terms in the future and once more data sets have been added.

WEB LINK TO MAINTAINER OF THE RESOURCE (URL)

The web link to the resources is essential for resources to allow for an access directly to the data (if possible without a constraining copyright agreement) and also the possibility to obtain additional information on the datasets and where the data have been used. A resource without a link is difficult to interpret, as it is not clear for example how to obtain the data. Potentially an email address can be given as a contact point as well.
If the data is available from several sources there can potentially be several links.

DOMAIN (MEDICAL, TRADEMARK, PERSONAL PHOTOS, OUTDOORS, ETC.)

The domain of images is important for researchers to know whether the database is in the scope of their work. Medical images need to be treated for retrieval in a different way than general stock photography or trademark images.
There is a list with example domain and this list is expected to grow with an increasing number of specialized collections becoming available. The future goals of the Chorus+ project have shown the importance of domain-specific applications for audiovisual information retrieval.

MEDIA (IMAGE, VIDEO, TEXT, ETC.)

The media is equally important as an access point for researchers because images, videos or audio need different analysis methods. For real multimedia databases it is possible to have several entries of this element such as text and image.
Again, it is possible to add new media such as 3D or 4D data sources in the future and based upon need.

SIZE (IN GB, ETC.)

The size of the database and the number of instances such as the number of images have been separated deliberately to have the two information sources whenever possible. This allows judging download time but also complexity and detail of the data in the instances.
The size in GB or TB gives an indication of how long a download would take or whether the data might need to be obtained via a hard disk, for example.

INSTANCES (NUMBER OF IMAGES/VIDEOS)

The number of instances gives in connection with the size also an indication of the complexity of the information contained in the images. Some retrieval systems have limits in terms of the number of instances to analyze and this can be a good indication in this case.
For temporal data such as video or sound besides the number of actual files it seems necessary to add also the length in terms of time (seconds, minutes, hours) of the video/audio data.

FORMAT (JPG, PNG, MPG, ETC.)

The format of the files is required to give researchers a quick overview of what to expect and whether they can work with the data as it is or whether they need to convert the data. Whereas most often web formats (PNG, JPG) will be used there can be domain-specific datasets that can be in specialized formats such as medical databases in the DICOM (Digital Imaging and Communications in Medicine) format.

CREATION DATE

This field contains the creation date of the source to be able to judge its novelty but also its acceptance and potential use in a community.

Potentially a database can be created and then updated and in this case several dates can be added in the field.

TASK (RETRIEVAL, RECOGNITION, ETC.)

There is a multitude of tasks associated with audiovisual data in terms of information access and the tasks for which a dataset was created and for which it has been used can be entered here. Typical tasks can be the **classification** of instances into a limited number of classes, **detection** of concepts in the instances, including potentially a **localization** of structures (such as face detection). Other tasks can be similarity-based **retrieval**.

The type of task foreseen for the database has then also an impact on the way the evaluation will take place and the measures used for the evaluation.

SOURCE (FLICKR, COREL, ETC.)

Starting from the mid-1990s a large number of groups used varying sources form the Corel data sets [1]. Since the availability of image sharing sites such as FlickR or Picasa this has been replaced by using these sources as the basis for creating a collections of varying sizes, topics and complexity.

Identifying the sources of data gives an idea about the variety and difficulty of the data and can at the same time help a potential data user to interpret the type of images to expect. This allows equally relating databases based on the same sources and their potential difficulty and variety as particularly FlickR has resulted in an extremely large number of different databases.

The entry is free text and the sources are expected to change over time.

ANNOTATION TYPE (FREE TEXT, STRUCTURED, ETC.)

Very often audiovisual data has little in terms of structured data or free text describing the audiovisual instances. This section gives an idea about the details of the structured data that comes with the audiovisual data.

The field is free text and can include a more detailed description of the data. The detailed information can be in free text, it can be structured and it can be in various formats such as pure text or XML for structured data. If particular vocabularies are used this can also be mentioned in this field.

GROUND TRUTH

To be able to use a data set for evaluation there needs to be a standardized setup for an evaluation and a gold standard for this setup. This gold standard in connection with specific tasks can also be called ground truth of the database.

Ground truth can be the classification of instances into classes or it can be relevance of instances for a specific information need or similarity-based retrieval. It is important to mention whether ground truth is available, what type of ground truth and also how it was obtained, for example through a single person or taking into account the subjectivity of several persons (inter-rater disagreement can be important).

This entry is not mandatory but can deliver information important for judging the quality of a data set.

EVENT OR PROJECT THAT CREATED THE RESOURCE

Many audiovisual resources are created as part of an existing project or a benchmarking campaign and this field is meant to give credit to the sources of the data creators such as TREC[10], ImageCLEF[11] or a project such as MUSCLE[12].
The field is not mandatory and many sources might not have been created in the context of a campaign or project.

QUALITY (RESOLUTION)

Particularly for images but also for videos or sound files it can be important to have clear technical indicators about the quality for the data such as the compression rate or the exact resolution of the data to better judge the usefulness of a resource.
The field is free text and is not mandatory. It can include a resolution for images or a quality criterion for compression or a particular codec used.

COPYRIGHT

The copyright is extremely important to know what the data can actually be used for and how. Unfortunately, many datasets even when distributed on the Internet do not detail their licenses and it is then important to know what can effectively be done with the resources. Thus it seems important to add where available the license of the data. A more detailed description on the copyright is given in Section 6. Particularly the CreativeCommons licenses try to make the creation of tailored licenses for media easier.

CITATION

This field allows adding a reference to an article that uses the resource, for example in bibtex style so the corresponding article can be found easily. If possible it should be described in a phrase or two, which technique is being used and the references could also be organized based on performance obtained. This should allow new techniques being evaluated based on the latest state of the art.

FEATURES

This field can detail whether visual features are already extracted from the database and are available or whether this is not the case. This could help particularly groups that do not have high-quality visual features but specialize on other system parts such as machine learning or user interaction.

4. Results

This section details some of the results of the collaborative platform in terms of evaluation resources created during the first two month of the platform, thus only in a limited way. The platform is currently being advertised and once a critical mass of data sets is available it is expected to be self-sustainable, requiring only a limited central quality control. The goal is really to create a shared community resource.

PRELIMINARY RESULTS OF THE COLLABORATIVE PLATFORM

In early November 2011, so two months after the official launch, the collaborative platform included 14 user accounts that allow the users to add any kind of evaluation resources to the

---

[10] http://trec.nist.gov/
[11] http://www.imageclef.org/
[12] http://muscle.ercim.eu/

Wiki. In total, 833 accesses on the main page were recorded, highlighting that more publicity is needed to really make it an important evaluation resource in the future.

The users added in total 46 data sets but not all of them in much detail. The Chorus+ project is currently working on controlling the quality of the added resources and also on completing missing information describing the data sets. Within the project lifetime periodic controls of the Wiki quality are foreseen.

SCREENSHOTS OF THE COLLABORATIVE PLATFORM

This section gives some details on the collaborative platform and the design choices in order to collect databases used for the evaluation of multimedia retrieval.

To allow various forms of access and sorting of the results, similar to access towards structure data in a database, a template was created in the Wiki for adding resources.

Figure 1 shows the entry page for the Wiki that allows choosing between datasets, tools, events or publications.
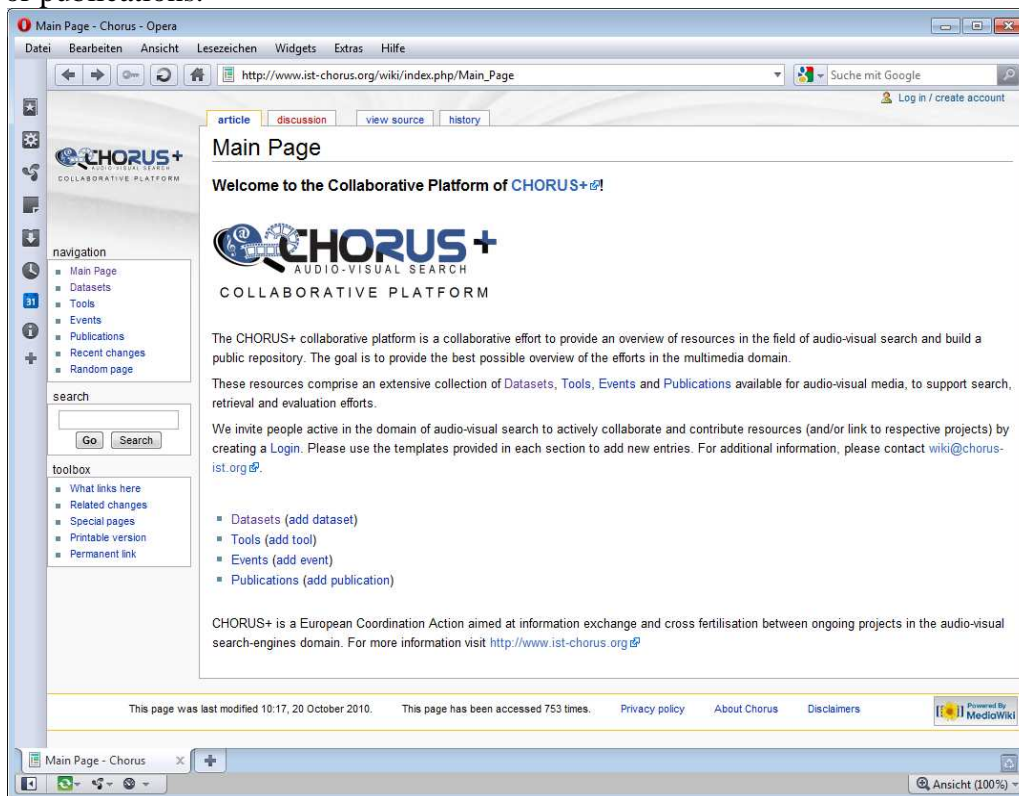

Figure 1: Entry page to the Chorus+ collaborative wiki making available various evaluation resources.

Figure 2 shows the template for adding a new dataset. The main and mandatory elements are shown in the template and the description below and the elements that are less essential.
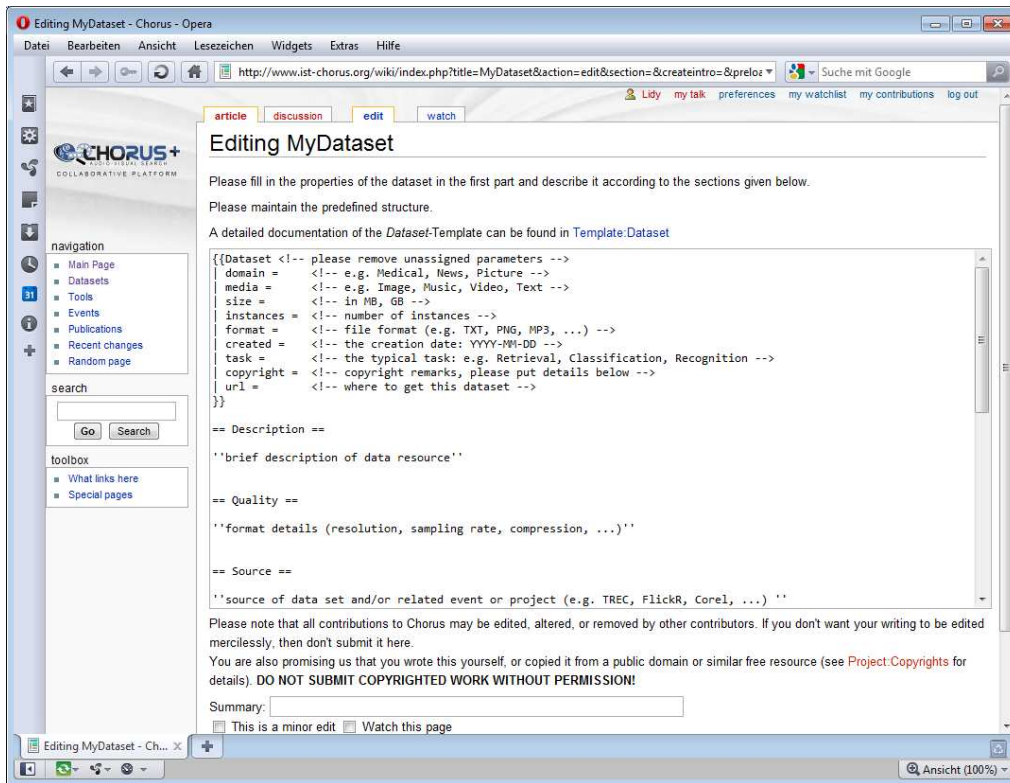
Figure 2: The template for adding a new dataset on the Wiki; such a template has the goal to structure the information and allow access by the information added in the fields.
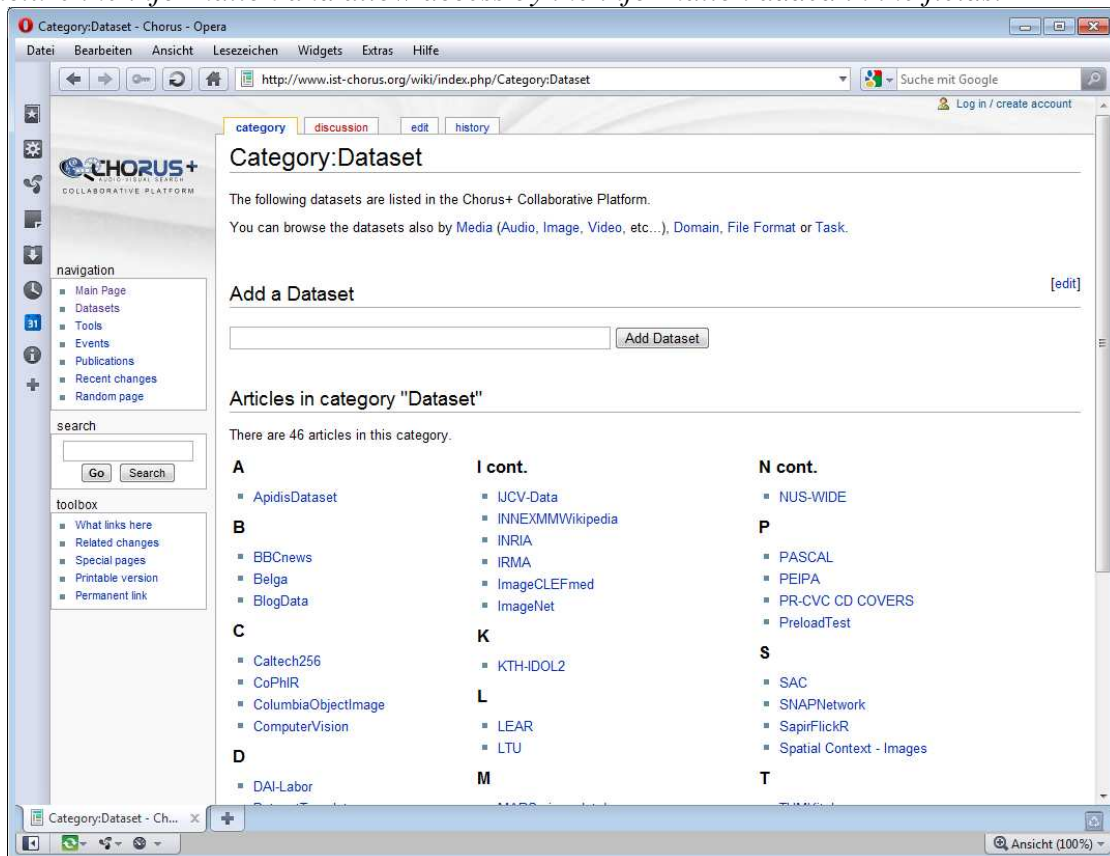

Figure 3: Automatically generated list of the image databases by name; other ways of sorting include by media type, by domain, by file format or by task.

In Figure 3 we can finally see the alphabetical order of the data sets in the results page and a link to the template for adding new resources. This page also has links to ordering by media, domain, file format or task. Thanks to the used template the views towards the data are created an updated automatically limiting the amount of manual intervention.

5. Copyright questions around multimedia databases

The use or reuse criteria of multimedia databases are often difficult to evaluate as some image databases are made available on the Internet but their terms of use do not allow for a reuse or redistribution even when the data is publicly accessible.

The main criteria analyzed in this section are:

- **Accessibility** of the data (not available, downloadable, downloadable with a password, freely accessible);
- Obligations to **reference the owner** or creator of the data;
- Rights reserved for the use of the data, for example only for scientific use, only for private use or also for **commercial** use;
- Right to show data in **publications** and this only in a specific context such as non-commercial or only on request;
- Right to **modify** the data and/or redistribute the data sets; and if this is possible, is the license to be used in this case the same or can another license be used.

Most early usage models for audiovisual content were rather restrictive, whereas the CreativeCommons model started going towards content sharing offering a toolbox of license models described in the next section.

One of the problems is that for many datasets that have been used in audiovisual data access in the past the license models are not necessarily passed with the data and data are exchanged between researchers, such as the Corel data sets for example, whether the copyright owner is known or not and even without any license agreement attached to it[13].

Often, when downloading data from the Internet, such as images crawled from web pages, the rights for using the images are not at all known. The images can clearly be viewed freely but any other task including the analysis of the images might be prohibited if no clear definition of the terms of use can be found.

For scientists one of the major questions when using data for evaluation is related to the use of the data as a source for the evaluation but also the use in publications to show data as examples when publishing.

CREATIVE COMMONS

Creative commons[14] was created in 2001 in the USA and has since helped to facilitate licensing for content such as audiovisual data. One of the goals was to have and open licensing model (CC-licenses) that helps also the sharing of content and culture and that makes it easier to develop licensing models. This is particularly in contrast with traditional licensing models that are rather restrictive and aim at limiting the reuse of data.

Creative commons particularly create a sort of toolbox for licenses that allows several levels in terms of whether it is mandatory to cite the creator, whether commercial use is permitted and whether modifications are allowed. Based on these criteria 12 main licensing models were developed ranging from "no rights reserved" to "some rights reserved" or even "all rights reserved".

Creative commons has mainly started with American law but by now several other national initiatives of CreativeCommens exist and allow for applicability in several countries.

---

[13] http://wang.ist.psu.edu/docs/related/
[14] http://www.creativecommons.org/

Besides the standard license models based on the principal questions CreativeCommons also worked on specific licenses, for example on sampling music and also for developing countries.

Many photo sharing web pages such as FlickR now allow the users to make their images available under specific licenses of CreativeCommons. This makes it possible when crawling data to only take images that have rather liberal license agreements.

To make the Licenses accessible for non-law specialists there are sometimes a short and a full version of the licenses, the short version allowing a layperson to understand the main points for choosing a license model. This accessibility for a large public was one of the important ideas of CreativeCommons.

RISKS WHEN DISTRIBUTING DATA SETS FOR EVALUATION

When organizing an evaluation and distributing data sets often-simple license agreements are used such as the one of ImageCLEF[15]. Such agreements are hard to enforce when participants come from many different countries as national laws differ significantly. Having too restrictive permissions might prohibit some companies or Universities to participate. Very often, the license prohibits certain uses but does not propose any specific punishment in case of an abuse, making it hard to be dissuasive of misusing the data.

When creating data sets from crawled data with attached licenses, for example from the FlickR platform one has to keep in mind that sometimes copyrighted material is illegally updated by persons that are not the copyright owners and with a wrong license model. If re-distributing this data, the re-distributor is potentially responsible for removing the illegal parts from the data sets. FlickR is thus not favoring to create external subsets of FlickR for redistribution as the online access via an API has the advantage of making it quick and simple to remove illegally labeled documents. On the other hand the same query terms will deliver differing datasets over time making comparisons of retrieval performance hard to evaluate.

If the distribution of data sets is absolutely not possible due to copyright constraints then there are possible solutions as taken by the MIREX[16] initiative where copyrighted data is used on a central closed server. Particulars of the benchmark can then send in their tools and an explication for its use. Then, a central team is executing the tools on the copyrighted data for evaluation. In the music industry where the data are strongly protected this can be the case but for many researchers this is a rather restrictive mechanism that does not allow tuning of systems on the evaluation data sets and also the possibility to work on the same data for a longer period in a comparative version.

6. Conclusions

This document does not have as a goal to recreate the resources that will be collected in the Wiki but to describe the methodology of resource creation, maintenance and access for persons searching for evaluation resources. The evaluation resources will be maintained in the Wiki as any list would be outdated within a few weeks. Goal is to create a sustainable infrastructure that can be reused by many projects from the audiovisual field. This platform should help give the created datasets the visibility they merit and it allows validating algorithms and comparing them on the same data. This has the objective to improve evaluation practice in the field in the long run.

It is foreseen to collect a larger number of data sets within the project and adding them to the collaborative Wiki and further describing them in a scientific article addressing the audiovisual retrieval community with a set of best practices.

---

[15] http://medgift.unige.ch/ImageCLEF2010EndUserAgreement.doc
[16] http://www.music-ir.org/

The importance of copyright is stressed in this text and the basic principals and risks when creating resources are mentioned, as most researchers are not aware of these risks and the basics of copyright law.

7. References

[1] Henning Müller, Stéphane Marchand-Maillet and Thierry Pun, The Truth About Corel - Evaluation in Image Retrieval, In Proceedings of The Challenge of Image and Video Retrieval (CIVR2002), Springer Lecture Notes in Computer Science LNCS2383, pages 38-49, London, UK, July 2002.

[2] Claire V. Thornley, Andrea C. Johnson, Alan F. Smeaton, Hyowon Lee, The Scholarly Impact of TRECVid (2003-2009), Journal of the American Sciety for Information Science (JASIS), 2011 – to appear.

---End of Document ---