Document title:	**Report on Storage and Management Policies for Multimedia Files**
Due delivery date:	30/09/2012
Nature:	**Deliverable**
Project Title:	**Collaborative Information, Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime**
Project acronym:	**CAPER**
Instrument:	**Large Scale Collaborative Project**
Thematic Priority:	**FP7-SECURITY-2010-1.2-1**
Grant Agreement:	**261712**

Organisation name of lead contractor for this deliverable:

Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

History			
Version	First name & Name	Modifications	Date
1.0	Jorge González	Multimedia formats	10/04/2012
1.1	Emma Teodoro	Legal issues concerning storage policies	16/04/2012
2.0	Antoni Roig	Legal and privacy issues	29/10/2012

Validation			
	First name & Name	Organisation short name	Visa
Responsible	Roig, Antoni	UAB	X
WP leader	Casanovas, Pompeu	UAB	X
Coordinator	Monreal, Carlos	S21sec	X

INDEX

EXECUTIVE SUMMARY

This deliverable aims at reviewing and analyzing the ethical and legal risks that arise in the application of the legal framework of data protection relevant to CAPER, including the storage of CAPER information.

First, the deliverable covers Data Protection issues with regard to research activities. It highlights that data gathered for the purposes of research activities are exempt from being processed, in accordance with some data protection principles. Specifically, personal information can be processed for purposes other than those for which it was originally obtained and held indefinitely. This is only true if the personal data are not processed to support measures or decisions relating to particular individuals. When research involves individuals, data should be timely and accurate. In most cases, however, research will be based on information representing a particular moment in time, and there will be no reason to keep up to date these data. A second condition is that the data subject may not suffer substantial damage or distress due to the processing. Indeed, data subjects have a right to object to the processing of data based on the possibility of suffering significant damage or distress from it. Nevertheless, the European Directive and the National Data Protection Acts do not provide a blanket exemption from all the data protection principles for research purposes. Most of the data protection principles still apply, and it is certainly the case for the requirement to keep data secure. Processing of any information relating to an identifiable living individual constitutes “personal data processing” and is then subject to the provisions of the European Directive and the National Data Protection Acts.

Secondly, the deliverable considers three risks scenarios. These three risks scenarios are assessed in order to provide concrete solutions or otherwise indicate some useful ethical principles for the technical partners of the consortium. These scenarios are the following ones: (i) The use of *Social Network Analyses (SNA)* for the purpose of research; (ii) The use of *profiling* for the purpose of research and; (iii) *The storage, access and conservation of multimedia file* for the purpose of research.

Finally, the deliverable proposes a set of tables containing the most relevant storage issues and recommendations for CAPER Researchers and LEAs. According to the European Commission, the Ethical Review on storage in FP7 should give proper answer to at least questions such as (among others): (i) where is the data stored? If stored electronically, this must be a machine, or set of machines located in a physically secured as well as technically secured? (ii) On which hardware type is the data stored: paper, disk, removable device? What will the adapted security processes to be followed? How do they guarantee confidentiality of data? (iii) Who has access to the data? Can the data be accessed by any third party? Can the data be copied by any third party? (iv) For how long will the data be stored, accessed? (v) What data backup policies and processes will be implemented?

CAPER researchers jointly with the IDT researchers identified and agreed to include as storage issues the following ones: storage location, access control, media for storing, data retention, deletion and destruction, data security, anonymization data or making backups. For LEAs this deliverable offers a set of precise recommendations with regard to data storage, data information management and CAPER data reuse and transfer. These materials will be discussed in further working sessions according to the CAPER plan activities, and can and must be amended according to LEAs’ answers and the Ethical Committee comments, in further deliverables.

1 INTRODUCTION

1.1 Aim of the deliverable

This deliverable has the main goal of determining ethical and legal risks scenarios and offering, wherever possible, concrete solutions or otherwise indicate some useful ethic principles for the technical partners of the consortium.

We plan to cover the legal and ethical aspects of the partner's research, related to three scenarios of work:

1. The use of *Social Network Analyses (SNA)* for the purpose of research. We plan to cover the legal and ethical aspects of this issue and provide a concrete legal advice to the partners (WP3 to WP6).

2. The use of *profiling* for the purpose of research. We plan again to cover the legal and ethical aspects of this issue and provide a concrete legal advice to the partners (WP3 to WP6).

3. *The storage, access and conservation of multimedia file* for the purpose of research. We plan once more to cover the legal and ethical aspects of this issue and provide a concrete legal advice to the partners (WP3 to WP 6).

1.2 Corresponding planned work in Annex I

Since CAPER includes different countries, jurisdictions, and Law Enforcement Agencies, the simplification of the access to the stored multimedia information through an international standard becomes a critical issue. In addition, the encoding of these multimedia contents that many ISO/IEC standards provide can help to reduce the storage requirements. As an example, the Joint Photographic Experts Group (JPEG) and the Moving Pictures Experts Group (MPEG) committees have developed standards focused on this topic.

This task therefore involves the review of the formats in which the images and videos captured in the net are encoded, stored and handled in the database for their analysis in WP5. The main objective is to ensure the protection of personal information and some related aspects: i.e. data integrity, source authentication, access control, confidentiality to non authorised users, prevention of external attacks aimed to compromise, modify, or even eliminate, the stored information.

As an additional topic, the access level granted by the system to authorised users is also reviewed to ensure the protection of personal information taking into account the different countries legislation.

2 CAPER RESEARCH PARTNERS AND DATA PROTECTION ISSUES

2.1 Data protection and research activity

Article 13(2) of the European Data Directive states that Member States may restrict the data subjects' rights of access to data when the data are processed "solely for purposes of scientific research or(...) for the sole purpose of creating statistics". The Directive places limits on this exemption:

> First, Member States' restrictions are "subject to adequate legal safeguards, in particular that the data are not used for taking measures or decisions regarding any particular individual".
>
> Second, the restrictions must pose "no risk of breaching the privacy of the data subject".
>
> Finally, Member States may only restrict a data subject's right of access for statistical purposes when the data are kept by the processor "for a period which does not exceed the period necessary for the sole purpose of creating statistics".

The exemption is quite broad since it does not enumerate the types of "research purposes" to which it applies. The research department of a public or private institution could fall within the exemption. Furthermore, any business seeking to avoid the restrictions of the Directive could simply establish its own "research" department. The exemption applies when data are processed *"solely* for research purposes". This seems to ban the use of data collected for research purposes outside the field of research. Article 6 (l) (b), however, provides that "further processing" may be allowed "provided that Member States provide appropriate safeguards". Unfortunately, the Directive does not define "appropriate safeguards”. Another provision limits the scope of the exemption: the data will not be used for “taking measures or decisions regarding any particular individual". The National Data Protection Acts should create clear and strict limits on the ability to use research data for purposes other than research. Member States should also define what constitutes "adequate safeguards”.

National Data Protection Acts implement the article 13(2) exemption of the European Data Protection Directive. Even if some differences can be found at this National level of regulation, the general data protection framework for research is similar to the one we summarize now.

Data gathered for the purposes of research activity are exempt from being processed in accordance with some data protection principles. Specifically, personal information can be (i) processed for purposes other than those for which it was originally obtained and (ii) held indefinitely.

> This is only true if the personal data are not processed to support measures or decisions relating to particular individuals. In cases the research that is being conducted will support measures of decision concerning individuals, it may be essential for data to be timely and accurate. In most cases, on the contrary,

research will be based on information representing a particular moment in time, and there will be no reason then to keep up to date these data.

A second condition is that the data subject may not suffer substantial damage or distress due to the processing. Indeed, data subjects have a right to object to the processing of data based on the possibility of suffering significant damage or distress from it.

Researchers are thus allowed to keep records of questionnaires and contacts in order to re-analyse the information in case future research projects in an associated area would require them. Nevertheless, the European Directive and the National Data Protection Acts do not provide a blanket exemption from all the data protection principles for research purposes. Most of the data protection principles still apply, and it is certainly the case for the requirement to keep data secure. Processing of any information relating to an identifiable living individual constitutes "personal data processing" and is then subject to the provisions of the European Directive and the National Data Protection Acts.

Some good practices are welcome. For instance, fully anonymisation of the collected data, in the sense that there is no way to re-identity the data subject has a clear benefit: the Data Protection National Act does not apply in this case because such information is not considered to be "personal data". Personal data relates to a living individual identified or who can be identified from these data. Re-identification is crucial, and that is the reason why true anonymisation of data is difficult to achieve in practice. In any case, if re-identification is possible in some way, the Data Protection Act does still apply.

Researchers are not entitled to make private use of data controlled by their institution unless they have explicit permission. Institutions need a formal policy on such matters. If data are supplied in anonymised form, there are no further data protection issues to consider.

According to data protection principles, researchers should inform research subjects as far as possible of:

- The purpose of the research for which personal data about them will be collected;
- How their personal data will be used; and
- Who will have access to their data.
- The identity of the data controller

If this involves disproportionate effort, data controllers can avoid this obligation, noting the reasons for doing so. These reasons may be subject to legal challenge.

We have already mentioned the requirement to keep data secure. Researches must ensure that personal data are protected from unauthorised access or accidental loss, damage or destruction. These measures should also be communicated to the data subjects. Anonymisation of data should be carried out to as great an extent as possible. In fact, many research projects are only interested on statistics and general patterns, and there is no need then for data to be directly associated with the individuals who provided it.

The data subjects have the right to request access to intelligible copies of personal data about them where they are identified as the data subject. However, this requirement will not apply when collecting for the purpose of research activity. This is only true when the data are managed according to the data protection principles and the data subjects are not identified in the results of the research.

The processing of sensitive data[1] for research requires:

- The explicit consent (ideally in writing) of the data subject.
- A medical research carried out by a professional with a duty of confidentiality.
- If it is an analysis of racial/ethnic origins, it has to be for the purpose of determining equality of opportunity and with a view to enabling the maintenance or the promotion of equality.

International transfer of personal data will require a new consent. Moreover, the institution overseas will have to provide appropriate levels of security for sensitive data and security of data. Anonymisation of data increases the security of data processing.

Researchers will generally be able to disguise the identity of research subjects in their research publications. Some reports, however, will be focussed on an individual subject's circumstances, and the context could allow thus a re-identification. That's the reason why researchers should disclose the minimum possible personal information in reports. If it is not possible, then the individual must give consent before publication proceeds.

2.2 The use of *Social Network Analyses (SNA)* for the purpose of research

2.2.1 Personal data in Social Networks

Personal data is data from which a living individual can be identified. Even if the individual is not immediately recognisable from one single piece of datum, he may become identifiable by combining it with other information. The key is whether the individual can be identified. Consequently, pictures, telephone numbers or email addresses, for instance, can be personal data. Note that by combining data someone might become identifiable even if she/he was not in the photograph.

Opinion 5/2009 on online social networking has an interesting summary of obligations and rights[2]:

Applicability of EC Directives

1. The Data Protection Directive generally applies to the processing of personal data by Social Network Service providers (SNS), even when their headquarters are outside of the EEA.
2. SNS providers are considered data controllers under the Data Protection Directive.

[1] The following types of information fall into the category of sensitive personal data:
- ethnic or racial origin
- political opinions
- religious beliefs
- trade union membership
- physical or mental health
- sexual orientation and behaviour
- criminal offences or alleged criminal offenses

[2] Opinion 5/2009 on online social networking adopted on 12 June 2009.

3. Application providers might be considered data controllers under the Data Protection Directive.
4. Users are considered data subjects vis-à-vis the processing of their data by SNS.
5. Processing of personal data by users in most cases falls within the household exemption. There are instances where the activities of a user are not covered by this exemption.
6. SNS fall outside of the scope of the definition of electronic communication service and therefore the Data Retention Directive does not apply to SNS.

Obligations of Social Network Service providers (SNS)

7. SNS should inform users of their identity, and provide comprehensive and clear information about the purposes and different ways in which they intend to process personal data.
8. SNS should offer privacy-friendly default settings.
9. SNS should provide information and adequate warning to users about privacy risks when they upload data onto the SNS.
11. Users should be advised by SNS that pictures or information about other individuals, should only be uploaded with the individual's consent.
12. At a minimum, the homepage of SNS should contain a link to a complaint facility, covering data protection issues, for both members and non-members.
13. Marketing activity must comply with the rules laid down in the Data Protection and ePrivacy Directives.
14. SNS must set maximum periods to retain data on inactive users. Abandoned accounts must be deleted.
15. With regard to minors, SNS should take appropriate action to limit the risks.

Rights of Users

16. Both members and non-members of SNS have the rights of data subjects if applicable, according to the provisions of Article 10 – 14 of the Data Protection Directive.
17. Both members and non-members should have access to an easy-to-use complaint handling procedure set up by the SNS.
18. Users should, in general, be allowed to adopt a pseudonym.

2.2.2 Social network tracking and profiling

Search engines are services that help their users to find information on the Web. They can be classified according to the different types of data they aim to retrieve, including pictures and/or videos and/or sound or different kinds of formats. In the context of the Directive on Electronic Commerce (2000/31/EC) search engines have been denoted as a type of information society service, namely information location tools.

According to the Article 29 Data Protection Working Group, the purpose for using search engines can be[3]:

. Improving the service

[3] Opinion 1/2008 on data protection issues related to search engines adopted on 4 April 2008.

Many controllers utilise server logs to improve their services and the quality of their search services. In their view, server log analysis is an important tool in refining the quality of searches, results, and advertisements, and also to build new, yet unforeseen, services.

. Securing the system

Server logs are said to contribute to keeping search engine services secure. Some search engine providers have stated that log retention can help protect the system from security attacks, and that they require a sufficient historical sample of server log data to detect patterns and analyse security threats.

. Fraud prevention

Server logs are said to contribute to protecting search engines' systems and users from fraud and abuse. Many search engine providers operate a 'pay per click' mechanism for the advertisements shown. As a drawback, this may lead to a company being unfairly charged if an attacker uses automatic software to click systematically on the advertisements. Search engine providers devote attention to ensure that this type of behaviour is detected and eradicated.

. Accounting requirements

Accounting requirements are claimed as purpose for services such as clicks on sponsored links, where there is a contractual and accounting obligation to retain data, at a minimum until invoices are paid and the period for legal disputes has expired.

. Personalised advertising

Search engine providers seek personalised advertising in order to increase their revenues. Current practices include taking into account history of past queries, user categorisation and geographical criteria. Therefore, based on the behaviour of the user and on his or her IP address, a personalised advertisement can be displayed.

. Statistics

Statistics are collected by some search engines to determine what categories of users access what information online, at what time of the year. These data can be used to improve the service, to target advertisements and also for commercial purposes to determine the cost for a company that wants to advertise its products.

. Law enforcement

Some providers state that logs are an important tool for law enforcement to investigate and prosecute serious crimes, such as child exploitation.
Law enforcement authorities may sometimes request user data from search engines in order to detect or prevent crime. Private parties may also try to obtain a court order addressing a search engine provider to hand over user data. When such requests follow valid legal procedures and result in valid legal orders, of course search engine providers will need to comply with them and supply the information that is necessary. However, this compliance should not be mistaken for a legal obligation or justification for storing such data solely for these purposes.

The Legal framework for search engines is defined by the Article 29 Data Protection Working Group in the Opinion 1/2008:

Applicability of EC Directives

1. The Data Protection Directive (95/46/EC) generally applies to the processing of personal data by search engines, even when their headquarters are outside of the EEA.
2. Non-EEA based search engine providers should inform their users about the conditions in which they must comply with the Data Protection Directive, whether by establishment or by the use of equipment.
3. The Data Retention Directive (2006/24/EC) does not apply to internet search engines.

Obligations on search engine providers

4. Search engines may only process personal data for legitimate purposes and the amount of data has to be relevant and not excessive in respect of the various purposes to be achieved.
5. Search engine providers must delete or anonymise (in an irreversible and efficient way) personal data once they are no longer necessary for the purpose for which they were collected. The Working Party calls for the development of appropriate anonymisation schemes by search engine providers.
6. Retention periods should be minimised and be proportionate to each purpose put forward by search engine providers. In view of the initial explanations given by search engine providers on the possible purposes for collecting personal data, the Working Party does not see a basis for a retention period beyond 6 months. However, national legislation may require earlier deletion of personal data. In case search engine providers retain personal data longer than 6 months, they must demonstrate comprehensively that it is strictly necessary for the service. In any case, the information about the data retention period chosen by search engine providers should be easily accessible from their homepage.
7. While search engine providers inevitably collect some personal data about the users of their services, such as their IP address, resulting from standard HTTP traffic, it is not necessary to collect additional personal data from individual users in order to be able to perform the service of delivering search results and advertisements.
8. If search engine providers use cookies, their lifetime should not be longer than demonstrably necessary. Similarly to web cookies, flash cookies should only be installed if transparent information is provided about the purpose for which they are installed and how to access, edit and delete this information.
9. Search engine providers must give users clear and intelligible information about their identity and location and about the data they intend to collect, store or transmit, as well as the purpose for which they are collected.
10. Enrichment of user profiles with data not provided by the users themselves is to be based on the consent of the users.
11. If search engine providers provide means to retain the individual search history, they should make sure they have the consent of the user.
12. Search engines should respect website editor opt-outs indicating that the website should not be crawled and indexed or included in the search engines' caches.
13. When search engine providers provide a cache, in which personal data are being made available for longer than the original publication, they must respect the right of data subjects to have excessive and inaccurate data removed from their cache. The Working Party recommends a layered model for privacy policy as described in the *WP Opinion on More*

Harmonised Information Provisions (WP 100, http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2004/wp100_en.pdf)
14. Search engine providers that specialise in the creation of value added operations, such as profiles of natural persons (so called 'people search engines') and facial recognition software on images must have a legitimate ground for processing, such as consent, and meet all other requirements of the Data Protection Directive, such as the obligation to guarantee the quality of data and fairness of processing.

Rights of users

15. Users of search engine services have the right to access, inspect and correct if necessary, according to Article 12 of the Data Protection Directive (95/46/EC), all their personal data, including their profiles and search history.
16. Cross-correlation of data originating from different services belonging to the search engine provider may only be performed if consent has been granted by the user for that specific service.

2.3 The use of *profiling* for the purpose of research

Risk profiling could be defined as behavioural analysis to classify entities or individuals based on risk. Several techniques can be used for this general purpose. We have selected a few of them, the ones that may be used by CAPER partners.

2.3.1 Cookie consent exemption

Article 5.3 of Directive 2002/58/EC, as amended by Directive 2009/136/EC has reinforced the protection of users of electronic communication networks and services by requiring informed consent before information is stored or accessed in the user's (or subscriber's) terminal device. The requirement applies to all types of information stored or accessed in the user's terminal device although the majority of discussion has focused on the usage of cookies as understood by the definition in RFC62651[4]. Article 5.3 allows cookies to be exempted from the requirement of informed consent, if they satisfy one of the following criteria:

CRITERION A: the cookie is used "*for the sole purpose of carrying out the transmission of a communication over an electronic communications network*".

CRITERION B: the cookie is "*strictly necessary in order for the provider of an information society service explicitly requested by the subscriber or user to provide the service*".

Some cookie usage scenarios do not fall in the exemption afforded under CRITERION A or B.

a) Social plug-in tracking cookies

[4] http://tools.ietf.org/html/rfc6265

Many social networks propose “social plug-in modules” that website owners can integrate in their platform, to provide some services than can be considered as “explicitly requested” by their members. However these modules can also be used to track individuals, both members and non-members, with third party cookies for additional purposes such as behavioural advertising, analytics or market research, for example.

With such purposes, these cookies cannot be deemed to be “*strictly necessary”* to provide a functionality explicitly requested by the user. Therefore these tracking cookies cannot be exempted under CRITERION B. Without consent, it seems unlikely that there is any legal basis for social networks to collect data through social plug-ins about non-members of their network. By default, social plug-ins should thus not set a third party cookie in pages displayed to non-members. On the other hand, social networks have ample opportunity to collect consent from their members directly on their platform if they wish to conduct such tracking activities, having provided their users with clear and comprehensive information about this activity.

b) Third party advertising

Third party cookies used for behavioural advertising are not exempted from consent, as already highlighted in detail by the Working Party in Opinion 2/2010 and Opinion 16/2011. This requirement for consent naturally extends to all related third party operational cookies used in advertising, including cookies used for the purpose of frequency capping, financial logging, ad affiliation, click fraud detection, research and market analysis, product improvement and debugging, as neither of these purposes can be considered to be related to a service or functionality of an information society service *explicitly requested by the user*, as required by CRITERION B.

In this regard, since December 22, 2011 the Working Party has been actively participating in the work of the World Wide Web Consortium (W3C) to standardize the technology and the meaning of Do Not Track. In view of the fact that cookies often contain unique identifiers, that allow for the tracking of user behaviour over time and across websites and the possible combination of these identifiers with other identifying or identifiable data, the Working Party is concerned about the possible exclusion from Do Not Track of certain cookies that are said to be necessary for operational purposes. Such purposes are: Frequency Capping, Financial Logging, 3rd Party Auditing, Security, Contextual Content, Research and Market Analytics, Product Improvement and Debugging[5]. In order for the Do Not Track standard to bring compliance to companies serving cookies to European citizens, Do Not Track must effectively mean “*Do Not Collect*” without exceptions. Therefore where a user has expressed the preference to not be tracked (DNT=1) no identifier, for the purpose of tracking, must be set or otherwise processed. There are technical solutions available, and many more are currently being developed, to effectively apply privacy by design, both within the web browser and on the server side to achieve the operational purposes described above.

c) First party analytics

Analytics are statistical audience measuring tools for websites, which often rely on cookies. These tools are notably used by website owners to estimate the number of unique visitors, to

[5] http://www.w3.org/TR/tracking-compliance/

detect the most preeminent search engine keywords that lead to a webpage or to track down website navigation issues. Analytics tools available today use a number of different data collection and analysis models each of which present different data protection risks. A first-party analytic system based on “first party” cookies clearly presents different risks compared to a third-party analytics system based on “third party” cookies. There are also tools which use “first party” cookies with the analysis performed by another party. This other party will be considered as a joint controller or as a processor depending on whether it uses the data for its own purposes or if it is prohibited to do so through technical or contractual arrangements.
While they are often considered as a “strictly necessary” tool for website operators, they are not strictly necessary to provide a functionality explicitly requested by the user (or subscriber). In fact, the user can access all the functionalities provided by the website when such cookies are disabled. As a consequence, these cookies do not fall under the exemption defined in CRITERION A or B.

However, the Working Party considers that first party analytics cookies are not likely to create a privacy risk when they are strictly limited to first party aggregated statistical purposes and when they are used by websites that already provide clear information about these cookies in their privacy policy as well as adequate privacy safeguards. Such safeguards are expected to include a user friendly mechanism to opt-out from any data collection and comprehensive anonymization mechanisms that are applied to other collected identifiable information such as IP addresses.

In this regard, should article 5.3 of the Directive 2002/58/EC be re-visited in the future, the European legislator might appropriately add a third exemption criterion to consent for cookies that are strictly limited to first party anonymized and aggregated statistical purposes.

First party analytics should be clearly distinguished from third party analytics, which use a common third party cookie to collect navigation information related to users across distinct websites, and which pose a substantially greater risk to privacy.

2.3.2 Behavioural analysis for the purpose of research

The article 29 Data Protection Working party, the independent European advisory body on data protection and privacy adopted an opinion on the concepts of “controller” and “processor”[6]. In this Opinion, one example is mentioned on “Behavioural advertising”:

Behavioural advertising is advertising that is based on the observation of the behaviour of individuals over time. Behavioural advertising seeks to study the characteristics of this behaviour through their actions (repeated site visits, interactions, keywords, online content production, etc.) in order to develop a specific profile and thus provide data subjects with advertisements tailored to match their inferred interests. So behavioural advertising gives advertisers a detailed picture of a data subject’s online life, websites viewed, how long they have viewed them, and in which order, for instance.

Behavioural advertising involves the following roles: *(a) Advertising networks providers (also referred to as "ad network providers")*, the most important distributors of behavioural advertising since they connect publishers with advertisers*; (b) Advertisers* who want to

[6] Opinion 01/2010 on the concepts of “controller” and “processor” adopted on 16 February 2010.

promote a product or service to a specific audience; and (c) *Publishers* who are the website owners looking for revenues by selling space to display ads on their website(s).

"Behavioural advertising

Behavioural advertising uses information collected on an individual's web-browsing behaviour, such as the pages visited or the searches made, to select which advertisements to display to that individual. Both publishers, which very often rent advertising spaces on their websites, and ad network providers, who fill those spaces with targeted advertising, may collect and exchange information on users, depending on specific contractual arrangements.
From a data protection perspective, the publisher is to be considered as an autonomous controller insofar as it collects personal data from the user (user profile, IP address, location, language of operating system, etc) for its own purposes. The ad network provider will also be controller insofar as it determines the purposes (monitoring users across websites) or the essential means of the processing of data. Depending on the conditions of collaboration between the publisher and the ad network provider, for instance if the publisher enables the transfer of personal data to the ad network provider, including for instance through a re-direction of the user to the webpage of the ad network provider, they could be joint controllers for the set of processing operations leading to behavioural advertising.
In all cases, (joint) controllers shall ensure that the complexity and the technicalities of the behavioural advertising system do not prevent them from finding appropriate ways to comply with controllers' obligations and to ensure data subjects' rights. This would include notably:

• Information to the user on the fact that his/her data are accessible by a third party: this could be done more efficiently by the publisher, who is the main interlocutor of the user,

• and conditions of access to personal data: the ad-network company would have to answer to users' requests on the way they perform targeted advertising on users data, and comply with correction and deletion requests.

In addition, publishers and ad network providers may be subject to other obligations stemming from civil and consumer protection laws, including tort laws and unfair commercial practices."

Publisher and network provider are thus considered controllers in the sense of the data protection Directive.

Another example from the same Opinion is more close to our present subject: Processing for historical, scientific and statistical purposes:

"Processing for historical, scientific and statistical purposes National law may introduce, with regard to processing of personal data for historical, scientific and statistical purposes, the notion of intermediary organization to designate the body in charge of transforming non-encoded data into encoded data, so that the controller of the processing for historical, scientific and statistical purposes would not be able to re-identify the data subjects.
If several controllers of initial processing operations transmit data to one or more third parties for further processing for historical, scientific and statistical purposes, the data are first encoded by an intermediary organization. In this case the intermediary organization may be considered as controller pursuant to specific national regulations, and it is subject to all resulting obligations (relevance of the data, informing the data subject, notification etc.). This is justified by the fact that when data from different sources are brought together, there is a

particular threat to data protection, justifying the intermediary organization's own responsibility. Consequently, it is not simply considered as processor but fully regarded as controller pursuant to national law".

The encoding of data can thus need an intermediary organization, and this new organization will also be considered "controller" in the sense of the Data Protection Directive and not only a "processor"[7].

After the Opinion 1/2010, the Article 29 Data Protection Working Party took the occasion to return to behavioural advertising on Opinion 2/2010 and more recently on Opinion 16/2011[8]. The difference between predictive profiles and explicit profiles is interesting for our purpose:

There are two main approaches to building user profiles: i) ***Predictive profiles*** *are established by inference from observing individual and collective user behaviour over time, particularly by monitoring visited pages and ads viewed or clicked on. ii)* ***Explicit profiles*** *are created from personal data that data subjects themselves provide to a web service, such as by registering. Both approaches can be combined. Additionally, predictive profiles may be made explicit at a later time, when a data subject creates login credentials for a website.*

The legal framework and obligations is also clarified in the first Opinion:

Applicable laws

• The EU legal framework for the use of cookies is primarily laid down in Article 5(3) of the ePrivacy Directive[9].

• Article 5(3) applies whenever "information" such as a cookie is stored or retrieved from the terminal equipment of an internet user. It is not a prerequisite that this information is personal data.

• In addition, Directive 95/46/EC applies to matters not specifically covered by the ePrivacy Directive whenever personal data are processed. Behavioural advertising is based on the use of identifiers that enable the creation of very detailed user's profiles which, in most cases, will be deemed personal data.

Jurisdiction, territorial issue – establishment

[7] Article 2 (d) and (e) of Directive 95/46/EC read as follows:

'Controller' shall mean the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data; where the purposes and means of processing are determined by national or Community laws or regulations the controller or the specific criteria for his nomination may be designated by national or Community law;

'Processor' shall mean a natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller.

[8] Opinion 02/2010 on online behavioural advertising, adopted on 22 June 2010 and Opinion 16/2011 on EASA/IAB Best Practice Recommendation on Online Behavioural Advertising.

[9] Directive 2009/136/EC, that revised the 2002 e-Privacy Directive (2002/58/EC).

• The Directive 95/46/EC applies to the data processing that takes place when publishers and ad network providers engage in behavioural advertising *ex* Article 4.1(a) and (c) of Directive (95/46/EC) and *ex* Art 3 of the ePrivacy Directive. Existing Article 29 Working Party guidance on this issue is fully applicable.

Roles and responsibilities

• **Ad network providers** are bound by the obligations of Article 5(3) of the ePrivacy Directive insofar as they place cookies and/or retrieve information from cookies already stored in the data subjects' terminal equipment. They are also data controllers insofar as they determine the purposes and the essential means of the processing of data.

• ***Publishers*** have certain data controller related responsibilities regarding the processing that takes place in the first phase of the processing, i.e., when by virtue of the way they set up their web sites they trigger the transfer of the IP address to ad network providers (which enable the further processing). Such responsibility entails some, limited data protection obligations (see below). In addition, when/if publishers transfer directly identifiable personal data to ad network providers themselves, they will be deemed joint controllers.

Obligations and rights

Regarding ad network providers:
• Article 5(3) of the ePrivacy Directive which sets up an obligation to obtain prior informed consent applies to ad network providers.

• Browser settings may only deliver consent in very limited circumstances. Notably, if browsers are set up by default to reject all cookies (having the browser set to such an option) and the user has changed the settings to affirmatively accept cookies, for which he has been fully informed about the name of the data controller, the processing its goals and the data that is collected. Therefore, the browser must either alone or in combination with other means effectively convey clear, comprehensive and fully visible information about the processing.

• Ad network providers should encourage and work with browser manufacturers/developers to implement privacy by design in browsers.

• Cookie-based opt-out mechanisms in general do not constitute an adequate mechanism to obtain informed user consent. In most cases user's consent is implied if they do not opt out. However, in practice, very few people exercise the opt-out option, not because they have made an informed decision to accept behavioural advertising, but rather because they do not realise that the processing is taking place, much less how to exercise the opt out.

• Ad network providers should swiftly move away from opt-out mechanisms and create prior opt-in mechanisms. Mechanisms to deliver informed, valid consent should require an affirmative action by the data subject indicating his/her willingness to receive cookies and the subsequent monitoring of their surfing behaviour for the purposes of sending him tailored advertising.

• In accordance with Recital 25 of the ePrivacy Directive, a users' acceptance to receive a cookie could also entail his/her acceptance for the subsequent readings of the cookie, and hence for the monitoring of his/her internet browsing. It would not be necessary to request consent for each reading of the cookie. However, to ensure that data subjects remain aware of the monitoring over time, ad network providers should*:*

i) limit in time the scope of the consent;
ii) offer the possibility to easily revoke their consent to being monitored for the purposes of serving behavioural advertising and
iii) create a symbol or other tools which should be visible in all the web sites where the monitoring takes place (the website partners of the ad network provider). This symbol would not only remind individuals of the monitoring but also help them to control whether they want to continue being monitored or wish to revoke their consent.

• Network providers should ensure compliance with the obligations that arise from Directive 95/46/EC which do not directly overlap with Article 5(3), namely, the purpose limitation principle, and security obligations.

• In addition, the ad network providers should enable individuals to exercise their rights of access and rectification and erasure. The Article 29 Working Party welcomes the practice of some ad network providers to offer data subjects the possibility to access and modify the interest categories in which they have been classified.

• Ad network providers should implement retention policies which ensure that information collected each time that a cookie is read is automatically deleted after a justified period of time (necessary for the purposes of the processing). This also applies for alternative tracking technologies used for behavioural advertising such as JavaScript installed in the user's browser environment.

Ad network providers and publishers:

• Providing highly visible information is a precondition for consent to be valid.
Mentioning the practice of behavioural advertising in general terms and conditions and/or privacy policies can never suffice. In this regard and taking into account the average low level of knowledge about the practice of behavioural advertising, efforts should be applied to change this situation.

• Ad network providers/ publishers must provide information to users in compliance with Article 10 of Directive 95/46/EC. In practical terms, they should ensure that individuals are told, at a minimum, who (i.e. which entity) is responsible for serving the cookie and collecting the related information. In addition, they should be informed in simple ways that (a) the cookie will be used to create profiles; (b) what type of information will be collected to build such profiles; (c) the fact that the profiles will be used to deliver targeted advertising and (d) the fact that the cookie will enable the user's identification across multiple web sites.

• Network providers/ publishers should provide the information directly on the screen, interactively, if needed, through layered notices. In any event it should be easily accessible and highly visible.

• Icons placed on the publisher's website, around advertising, with links to additional information, are good examples. The Article 29 Working Party urges the network providers/ publisher industry to be creative in this area.

2.4 Facial recognition

Facial recognition is the automatic processing of digital images which contain the faces of individuals for the purpose of identification, authentication/verification or categorisation of those individuals.

Images can be acquired by the data controller in many ways such as provided by the users of the online or mobile service, their friends and colleagues or from a third party. Images may contain the faces of the users themselves and/or other registered or non-registered users or acquired without the knowledge of the data subject. Regardless of the means by which these images may be acquired a legal basis is required to process them.

We follow here the Recommendations of the Opinion 2/2012 of the Article 29 Data Protection Working Group.

Unlawful processing for the purposes of facial recognition

Recommendation 1: If the data controller is acquiring the image directly then they must ensure they have the valid consent of the data subjects prior to acquisition and provide sufficient information relating to when a camera is operating for the purpose of facial recognition.

Recommendation 2: If individuals are acquiring digital images and uploading them to online and mobile services for the purpose of facial recognition the data controllers must ensure that the image uploaders have consented to the processing of the images which may take place for the purposes of facial recognition.

Recommendation 3: If data controllers are acquiring digital images of individuals from third parties (e.g. copied from a website or purchased from a different data controller) they must carefully consider the source and the context in which the original images are acquired and processed only if the data subjects had consented to such processing.

Recommendation 4: Data controllers must ensure that digital images and templates are only used for the specified purpose for which they have been provided. Data controllers should put technical controls in place in order to reduce the risk that digital images are further processed by third parties for purposes for which the user has not consented to. Data controllers should put in place tools for users to control the visibility of their images that they have uploaded where the default is to restrict access by third parties.

Recommendation 5: Data controllers must ensure that digital images of individuals who are not registered users of the service or have otherwise not consented to such processing are only processed in so far as the data controller has a legitimate interest for such processing.

Security breach during transit

In the case of online and mobile services it is likely that there will be data transit between image acquisition and the remaining processing stages (e.g. uploading an image from a camera to a website for feature extraction and comparison).

Recommendation 6: The data controller must take appropriate steps to ensure the security of data transit. This may include encrypted communication channels or encrypting the acquired image itself. Where possible, and especially in the case of authentication/verification, local processing should be favoured.

Face Detection, Normalisation, Feature Extraction

Data minimisation

Templates generated by a facial recognition system may contain more data than are necessary to perform the specified purpose(s).

Recommendation 7: Data controllers must ensure that data extracted from a digital image to build a template will not be excessive and will only contain the information required for the specified purpose, thereby avoiding any possible further processing. Templates should not be transferrable between facial recognition systems.

Security breach during data storage

Identification and authentication/verification are likely to require the storage of the template for use in a later comparison.

Recommendation 8: The data controller must consider the most appropriate location to store the data. This may include the user's device or the data controller's systems. The data controller must take appropriate steps to ensure the security of the stored data. This may also include encrypting the template. It should not be possible to obtain unauthorised access to the template or storage location. Especially in the case of facial recognition for the purpose of verification, biometric encryption techniques may be used; with such techniques, the cryptographic key is directly bound to the biometric data and is re-created only if the correct live biometric sample is presented on verification, whereas no image or template is stored (thus forming a type of "untraceable biometrics").

Subject access

Recommendation 9: The data controller should provide the data subjects with appropriate mechanisms to exercise their right of access, where appropriate, to both the original images, and the templates generated in the context of facial recognition.

2.5 The storage, conservation and reuse of multimedia file for the purpose of research

2.5.1 Obligation to inform the user

The obligation to inform individuals about the processing of their data is one of the fundamental principles of the Data Protection Directive. Article 10 regulates the provision of this information where data are obtained directly from the data subject. Data controllers are obliged to provide the data subject with following information:

- the identity of the controller and of his representative, if any;
- the purposes of the processing for which the data are intended;
- any further information such as - the recipients or categories of recipients of the data;
- whether replies to the questions are obligatory or voluntary, as well as the possible consequences of failure to reply;
- the existence of the right of access to and the right to rectify the data concerning him.

As controllers of the user data, search engines should make clear to users what information is collected about them and what it is used for. A basic description of the use of personal information should be provided whenever it is collected, even when a more detailed description is provided elsewhere. Users should be similarly informed about software, such as cookies, that might be placed on their computer when they use the website, and how these can be refused or deleted. The Working Party considers that this information is necessary in the case of search engines to guarantee fair processing.

2.5.2 Data minimisation

In *relation to the principle of minimal disclosure and to the duration of the minimum storage of personal data*, the Data Protection Directive stipulates that personal data must be "adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed" and they must be "kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed". In practice these principles implement the concept of the aforementioned *principle of minimal disclosure* in a binding legal text.

A good example of disproportionate information storage is the Passengers Name Record (PNR) Agreement. According to the Annex to the agreement, the following nineteen types of data would have to be collected by the airlines companies and be transferred to the DHS: (1) PNR record locator code, (2) date of reservation/issue of ticket, (3) date(s) of intended travel, (4) name(s), (5) available frequent flier and benefit information (i.e., free tickets, upgrades, etc.), (6) other names on PNR, including number of travellers on PNR, (7) all available contact information (including originator information), (8) all available payment/billing information (not including other transaction details linked to a credit card or account and not connected to the travel transaction), (9) travel itinerary for specific PNR, (10) travel agency/travel agent, (11) code share information, (12) split/divided information, (13) travel status of passenger (including confirmations and check-in status), (14) ticketing information, including ticket number, one way tickets and Automated Ticket Fare Quote, (15) all baggage information, (16) seat information, including seat number, (17) general remarks including OSI, SSI and SSR information, (18) any collected Advance Passenger Information System (APIS) information, (19) all historical changes to the PNR listed in numbers 1 to 18.

The European Data Protection Supervisor (EDPS) noted that the aforementioned types of data would be collected and stored not only for passengers, but also for prospective passengers who may eventually cancel their trip. The list of data was considered as excessive and disproportionate compared to the purposes pursued via the proposed Council decision. The EDPS proposed limiting the data to the following information: “PNR record locator code, date of reservation, date(s) of intended travel, passenger name, other names on PNR, all travel itinerary, identifiers for free tickets, one-way tickets, ticketing field information, ATFQ (Automatic Ticket Fare Quote) data, ticket number, date of ticket issuance, no show history, number of bags, bag tag numbers, go show information, number of bags on each segment, voluntary/involuntary upgrades, historical changes to PNR data with regard to the aforementioned items”. As for the processing of sensitive data, the EDPS recommended that airline carriers should not transfer sensitive data to the American DHS institution.

2.5.3 Anonymisation

If there is no legitimate ground for processing, or for use beyond the well-specified legitimate purposes, search engine providers must delete personal data. Instead of deletion, search engines may also anonymise data, but such anonymisation must be completely irreversible for the Data Protection Directive to no longer apply.

Even where an IP address and cookie are replaced by a unique identifier, the correlation of stored search queries may allow individuals to be identified. For this reason, where anonymisation rather than deletion of data is chosen, the methods used should be considered carefully and performed thoroughly. This might involve the removal of parts of the search history to avoid the possibility of indirect identification of the user who performed those searches.

Anonymisation of data should exclude any possibility of individuals to be identified, even by combining anonymised information held by the search engine company with information held by another stakeholder (for instance, an internet service provider).

Currently, some search engine providers truncate IPv4 addresses by removing the final octet, thus in effect retaining information about the user's ISP or subnet, but not directly identifying the individual. The activity could then originate from any of 254 IP addresses. This may not always be enough to guarantee anonymisation.

Finally, log anonymisation or deletion must also be applied retroactively and encompass all of the relevant search engine’s logs worldwide.

2.5.4 Retention of data

The Working Party does not see a basis for a retention period beyond 6 months[10]. However, the retention of personal data and the corresponding retention period must always be justified (with concrete and relevant arguments) and reduced to a minimum, to improve transparency, to ensure fair processing, and to guarantee proportionality with the purpose that justifies such retention.

[10] Opinion 1/2008 on data protection issues related to search engines adopted on 4 April 2008.

To that effect, the Working Party invites search engine providers to implement the principle of "privacy by design" which will additionally contribute to further reduce the retention period. In addition, the Working Party considers that a reduced retention period will increase users' trust in the service and will thus constitute a significant competitive advantage.

In case search engine providers retain personal data longer than 6 months, they will have to demonstrate comprehensively that it is strictly necessary for the service. In all cases search engine providers must inform users about the applicable retention policies for all kinds of user data they process.

The PNR Agreement between the US and the EU is also interesting to consider in determining the maximum period for data storage of CAPER research data. According to the proposal for the PNR Directive of 02.02.2011, the PNR data would have to be retained for a period of 30 days in a database at the Passenger Information Unit for a period of 30 days after their transfer to the Passenger Information Unit of the first Member State on whose territory the international flight is landing or departing. Upon expiry of the period of 30 days after the transfer of the PNR data to the aforementioned Passenger Information Unit the data shall be retained, masked out, at the Passenger Information Unit for a further period of five years. The Article 29 Data Protection Working Party considers the retention period of five years as disproportionate.

The European Commission proposal for a Council decision of 23.11.2011 on the transfer of PRN data from the EU to the US DHS foresees even longer storage period for the PNR data. In accordance with Article 8 of the proposal, DHS retains PNR data in an active database for up to five years. The data will be depersonalised and masked after the initial six months of this period, but the passenger will still be able to be identified. After this five-year period, the PRN data will be transferred to a dormant database for a period of up to ten years. According to the European Data Protection Supervisor, and similar to the position taken by the Article 29 Data Protection Working Party, the maximum retention period of fifteen years that is foreseen in the Proposal is disproportionate and excessive. Rather a retention period of six months is recommended.148 The position of the EDPS requiring for a retention period of six months instead of the period of fifteen years that is currently proposed illustrates a significant challenge on defining what the appropriate storage and retention period would be for specific types of data. The general data protection principle on the conservation of data stipulating that personal data must be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed" allows room for broad interpretation.

2.5.5 Security and integrity of the data

Articles 30, 31 and 32 of the Data Protection regulation

The EC has proposed to reform the current European data protection framework (Directive 95/46/EC), and has proposed an EU regulation on data protection. The regulation regards organisations that are processing personal data, regardless of the business sector the organisation is in. Security measures and personal data breach notifications are addressed in Articles 30, 31 and 32:

- Organisations processing personal data must take appropriate technical and organisational security measures to ensure security appropriate to the risks presented by the processing.

- For all business sectors the obligation to notify personal data breaches becomes mandatory4.

- Personal data breaches must be notified to a competent national authority without undue delay and, where feasible, within 24 hours, or else a justification should be provided.

- Personal data breaches must be notified to individuals if it is likely there will be an impact on their privacy. If the breached data was unintelligible (data that has either been encrypted (asymmetric or symmetric) or hashed), notification is not required.

2.5.6 Reuse for other purposes

To what extent and how user data are further analysed and whether or not (detailed) user profiles are being created, depends on the search engine provider. The Article 29 Working Party is aware of the possibility that this type of further processing of user data touches on a core field of innovation of search engine technology and can have a high relevance for competition. Full disclosure about the further use and analysis of user data could also result in increased vulnerability of search engine services to abuse of their services.

However, such considerations cannot be pleaded as an excuse for not complying with applicable data protection laws of the Member States. Moreover, search engine providers cannot claim that their purpose in collecting personal data is the development of new services whose nature is as yet undecided. Fairness demands that data subjects are aware of the extent to which their private life might be intruded upon when their data is obtained. This will not be possible unless purposes are more precisely defined.

3 CONCRETE RECOMMENDATIONS ON STORAGE FOR CAPER

3.1 Storage for research

According to the European Commission, the Ethical Review on storage in FP7 should give proper answer to at least these key aspects:

Data storage must be secured so as for the data not to become accessible to unwanted third parties and to be protected against disaster and risk.

To that end, the following topics should be considered:

1. *Where is the data stored? Data must be stored within a secured environment. If stored electronically, this must be a machine, or set of machines located in a physically secured environment – with controlled access - as well as technically secured: proper temperature control, etc.*

2. *On which hardware type is the data stored: paper, disk, removable device? Considering the nature of the data used in the project: what will the adapted security processes to be followed? How do they guarantee confidentiality of data?*

3. *Who has access to the data? Can the data be accessed by any third party? Can the data be copied by any third party?*

4. *For how long will the data be stored, accessed? What will happen to the data after the end of the study: duration of storage should be justified. Destruction at the convenience of the researcher would appear insufficient, unless clearly stated in the consent form and the approval of the local competent authority. Such deletion of data should be defined as irreversible, or reversible.*

5. *If stored on a machine, is the storage machine/server equipped with:*
- *Wifi*
- *Bluetooth*
- *USB drive*
- *On the whole, devices that might ease data duplication of circulation…*

6. *What data backup policies and processes will be implemented?*

7. *Anonymising data*

7. Anonymising data[11]

[11] UK Data Archive, *Managing and sharing data*. *Best Practice for Researchers*, University of Essex, March 2011.

Data may be anonymised by:

- removing direct identifiers, e.g. name or address
- aggregating or reducing the precision of information or a variable, e.g. replacing date of birth by age groups
- using pseudonyms
- restricting the upper or lower ranges of a variable to hide outliers, e.g. top-coding salaries

A person's identity can be disclosed from:

- direct identifiers, e.g. name, address, postcode information or telephone number
- indirect identifiers that, when linked with other publicly available information sources, could identify someone, e.g. information on workplace, occupation or exceptional values of characteristics like salary or age

Special attention may be needed in CAPER for:

- relational data, where relations between variables in related datasets can disclose identities
- geo-referenced data, where identifying spatial references such as point co-ordinates also have a geo-spatial value

Removing spatial references prevents disclosure, but it means that all geographical information is lost. A better option may be to keep spatial references intact and to impose access regulations on the data instead. As an alternative, point co-ordinates may be replaced by larger, non-disclosing geographical areas or by meaningful alternative variables that typify the geographical position.

Managing anonymisation:

- plan anonymisation early in the research at the time of data collection
- retain original unedited versions of data for use within the research team and for preservation
- create an anonymisation log of all replacements, aggregations or removals made
- store the log separately from the anonymised data files
- identify replacements in text in a meaningful way, e.g. in transcribed interviews indicate replaced text with [brackets] or use XML markup tags <anon>…..</anon>

Digital manipulation of audio and image files can be used to remove personal identifiers. However, techniques such as voice alteration and image blurring are labour-intensive and expensive and are likely to damage the research potential of the data. If confidentiality of audio-visual data is an issue, it is better to obtain the participant's consent to use and share the data unaltered, with additional access controls if necessary.

Principal Storage Topics and Recommendations for CAPER Researchers	
Topic	Recommendations
1. Storage location	Storage of CAPER data: • In a locked cabinet • In the institution or university archives • University computer/electronic media that have limited access.

	Storage of CAPER data containing sensible data: - Data should be de-identified. - Identifiers[12] should be store in a separate location. - Electronically data should be storage on a password protected and encrypted hard drive. The device should be in a locked room. - Hard copies of the data should be stored in a locked location, under personal control and supervision of the researcher.
2. Media for storing[13]	- Data should be stored in non-proprietary or open standard formats in order to ensure long-term software readability. - Data should be migrated to a new media device in order to prevent optical and magnetic physical degradation. This operation should be performed within 2-5 years period after its creation. - Check data integrity of stored data files regularly. - Use two different formats of storage (e.g. hard disk/DVD) - Creation of digital versions of paper documentations in PDF/A format for long-term preservation and storage. - Label stored data in order to easing physical accessibility and location.

[12] Identifiers could be:
1. Names
2. Postal address information, other than town or city, province, and postal code
3. Telephone numbers.
4. Fax numbers.
5. Electronic mail addresses.
6. Social insurance numbers.
7. Medical record numbers.
8. Health plan beneficiary numbers.
9. Account numbers.
10. Certificate/license numbers.
11. Vehicle identifiers and serial numbers, including license plate numbers.
12. Device identifiers and serial numbers.
13. Web universal resource locators (URLs).
14. Internet protocol (IP) address numbers.
15. Biometric identifiers, including fingerprints and voiceprints.
16. Full-face, oblique or full-profile photographic images and any comparable images.
17. Any other unique identifying number, characteristic, or code.
18. University ID numbers or login.

The following can be identifiers in certain situations:
1. Date of Birth
2. Occupation
3. Ethnicity
4. Gender
5. First three digits of the postal code.

[13] UK Data Archive, *Managing and sharing data*. *Best Practice for Researchers*, University of Essex, March 2011.

	• Areas and rooms for storage of digital and non-digital data should fit risk prevention regulations (e.g. flood and fire)
3. Access control[14]	Data centres may impose additional access regulations: • Specific authorisation from the data owner to access data. • Only granted researchers should have access to data. • Enable secure remote access to confidential data but avoiding the possibility to download data. • Data processing should be carried out on a central secure server. Therefore, no data should be exchanged over the network. • Publications regarding to the project must be conducted under the Statistical Disclosure Control carried out by a trained Service Staff.
4. Data retention, deletion and destruction.	• Data usage beyond the life of the project must be closely supervised. • Period of storage/retention recommended for the CAPER Project[15] [16]: two years after the completion of the project. • As regards data destruction for paper and devices, we recommend a minimum standard of DIN4[17], which ensures cross cut particles of a least 2x15mm (UK Data Archive).
5. Data Security[18]	Computer systems: • Locking computer systems with a password and installing a firewall system. • Servers should be protected through line-interactive uninterruptible power supply systems

[14] UK Data Archive, *Managing and sharing data*. *Best Practice for Researchers*, University of Essex, March 2011.

[15] Saint Mary's University Data Storage Guidelines, November 2007.

[16] The Australian Code for the responsible Conduct of Research, available at http://www.nhmrc.gov.au/index.htm suggests longer periods:

In general, the minimum recommended period for retention of research data is 5 years from the date of publication. However, in any particular case, the period for which data should be retained should be determined by the specific type of research. For example:

– for short-term research projects that are for assessment purposes only, such as research projects completed by students, retaining research data for 12 months after the completion of the project may be sufficient

– for most clinical trials, retaining research data for 15 years or more may be necessary

– for areas such as gene therapy, research data must be retained permanently (eg patient records)

– if the work has community or heritage value, research data should be kept permanently at this stage, preferably within a national collection.

[17] The highest security level is known as DIN 6, this is used by the United States federal government for ultra secure shredding of top secret or classified material, cross cutting into 1x5mm particles. UK Data Archive, *Managing and sharing data*. *Best Practice for Researchers*, University of Essex, March 2011.

[18] UK Data Archive, *Managing and sharing data*. *Best Practice for Researchers*, University of Essex, March 2011.

	(UPS). • Implementing password protection and control access to data files (e.g. no access, read only permission, administrator-only permission, etc.) • Controlling access to restricted materials with encryption. • Imposing non-disclosure agreements for managers or users of confidential data. • Data transmitted should be encrypted, avoiding non-encrypted methods as e-mail, FTP protocol and so on. • Data should be destroyed in a proper and consistent manner (see point 4 in this table) Physical data security requirements: • Access control to rooms and buildings where data and computers are placed. • Computers that contain sensitive data should not be shifted (e.g. a knock in a hard disk may provoke a failure causing a breach of security). Network security: • Confidential data must be stored in a server without access to the Internet. • Operating systems and anti-virus software should be updated in order to avoid viruses and malicious codes.
6. Making Backups[19]	• Minimal number of backup versions • Backups can be stored offline (CD/DVD, removable hard-drive, etc.) or on a networked hard disk. • Devices that contain a backup can be move to another place to keep it safe. • Critical data files should be backed-up daily, using an automated back-up process, preferably stored offline. • Master copies of critical files should be made in open formats which facilitate long-term usage. • Back-up files should be validated regularly For best backup procedures, consider: • To carry out the backup only to specific files or to the whole computer system. • To set the frequency of every backup • To determine strategies for all systems including laptops, home-based computers, devices, etc. • To organize and clearly labelling all backup files.
7. Anonymising data[20]	A person's identity can be disclosed from:

[19] UK Data Archive, *Managing and sharing data. Best Practice for Researchers*, University of Essex, March 2011.

[20] UK Data Archive, *Managing and sharing data. Best Practice for Researchers*, University of Essex, March 2011.

- Direct identifiers, e.g. name, address, postcode information or telephone number.
- Indirect identifiers that, when linked with other publicly available information sources, could help to identify someone, e.g. information on workplace, occupation or exceptional values of characteristics like salary or age.

Data may be anonymised by:

- Removing direct identifiers, e.g. name or address
- Aggregating or reducing the precision of information or a variable, e.g. replacing date of birth by age groups.
- Using pseudonyms.
- Restricting the upper or lower ranges of a variable to hide outliers, e.g. top-coding salaries.

Special attention may be needed in CAPER for:

- Relational data, where relations between variables in related datasets are able to disclose identities.
- Geo-referenced data. Removing spatial references prevents disclosure, but it means that all geographical information is lost. A better option should be to keep spatial references and to impose access regulations on the data instead. In addition, coordinates could be replaced by larger, non-disclosing geographical areas or by meaningful alternative variables that typify the geographical position.

Managing anonymisation:

- plan anonymisation early in the research at the time of data collection
- Retain original unedited versions of data for its use within research team and for preservation
- Create an anonymisation log composed by all replacements, aggregations or removals performed to the original data.
- Store the log separately from the anonymised data files
- Identify replacements in text in a meaningful way (e.g. replaced text with [brackets] or use XML markup tags <anon>…..</anon>)

Digital manipulation:

- Media files are able to contain identifiers. However, techniques specifically devoted to alter voice, video and images are very demanding in terms of time and budget. In addition, this process could decrease the quality of the information contained in these media files.
- In this issue, we recommend to obtain the participant's

	consent to use and share the information unaltered with additional access controls if it is necessary.

3.2 Storage for police investigation on organised crime and terrorism

3.2.1. CAPER data storage

LEA recommendations for data storage	
Topic	Recommendation
1. Interaction with LEA closed databases	• CAPER data must only be used for terrorism and serious crime • The storage of the CAPER data should be implemented in a separate repository. No contact with ordinary criminal databases should be allowed • CAPER is not an e-analysing tool; it is only an e-filter. This is important because CAPER data are not police analysed data. No automatic decision should substitute the police analysis of human experts.
2. Retention Periods	• Time limit of retention is not a fixed period. Time period should be linked with the relevancy of the data, "no reason for storing irrelevant data or data that has become irrelevant" should be the rule • Three examples to consider: (1) e-Communication, the Data Retention Directive has fixed a limit from 6 months to 2 years; (2) EURODAC determines 1 year retention period; (3) EU LEAS have adopted a 5 years period as a retention limit; (4) and, the USA in the PNR Agreement established according to EU a 15 years data retention period • We suggest considering the opportunity to transform into dormant data the information gathered after a 6 months period. The dormant data should be a non-directly operational database. This database should use

	anonimisation techniques and access restriction. It could be an effective way to achieve the necessary balance between privacy and the efficiency of serious crime and terrorism investigations. • Data not needed in any investigation should be transferred to a non-operational third database.
3. Special categories of data	• Minors: retention periods should be shorter • Non-suspects: access and reuse about non-suspects should be extraordinarily limited with due justification and explicit reference to the non use in any investigation

3.2.2. CAPER data information management

LEA recommendations for data information management
<u>Recommendations</u>
• A specific plan for upgrading hardware and software should be implemented.
• Deletion of obsolete user accounts.
• Need to maintain a detailed log of actions related to user accounts plus regular auditions regarding their validity, access rights and roles.
• User actions at CAPER database should be logged.
• Introduce changes needed in order to ensure the integrity of log system as a whole.
• Review existing log analysis in order to 1) identify all systems whose logs should be regularly examined; 2) include these logs within the SIEM solution; and c) develop procedures for non-automated review of logs.
• Introduce a specific plan for the upgrade of the software used for back-up and restore.
• CAPER data are raw data. No classification of suspects, victims and witnesses can be automatically inferred from the CAPER tool. Therefore, no indication of “analysed data” should be given to CAPER data.
• Suspects, victims and witnesses are not CAPER data labels. They are the result of a human analysis on CAPER data. Lately, an information management process should transform CAPER data later into labelled data ready to use for investigation.
• Use of removable media (pen drives, etc.) should be activated only when it is absolutely necessary and the activation/deactivation process should be logged and monitored.
• The use of system integrity tools should enable deletion and reporting of changes applied on servers. System alerts need to be generated and sent to specific user roles.
• Establish and document a personal data breach handling procedure. This refers to all the procedures that a data controller has in place in order to detect a personal

data breach in a timely manner and resolve it. Following the definition of Directive 136/2009/EC, personal data breach means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed by a data controller. The personal data breach handling procedure will describe all the steps to be taken in case that a personal data breach occurs, including the possible involvement of national Data Protection Officers. All data breaches should be adequately documented and maintained in a specific register, and national Data Protection Officers should be able to access the register at any time.
• Regular audits of the CAPER system should be performed. The European Data Protection Supervisor should be informed on the results, including the plans for enforcing recommendations.
• A specific procedure for the secure destruction of personal data within CAPER, covering both electronic and paper files, should be established.

3.2.3. CAPER data reuse and transfer

CAPER data reuse and transfer
Recommendations
• CAPER data are raw data. Consequently, any transfer of CAPER data should warn about this critical fact.
• CAPER transforms the former limitation of using data for only ONE CASE, and then erasing it. With tools like CAPER and their corresponding databases, more and more we have to think in ONE UNIQUE CASE, with concrete uses of a general unique network.
• In this context, purpose limitation is a main aspect: for serious crime and terrorism only.
• Other guarantees can be implemented in the sense that specific justification and authorisation are needed to access the CAPER data resulting network. The part of the network analysed will be restricted to what is needed in the case. The whole network will not be viewed and would require complementary and time limited authorisations.
• The reuse will require quality control on the CAPER data. The right to be informed, right to access and erase data is extraordinary limited within the terrorist and serious crime area. However, CAPER data would be used for investigation and prosecution, after been analysed and determined the relevant data. The European Courts have started to held the possibility of a limited information right on the lawfully procedures to collect data used after in trials. Therefore, some legal validation of the procedure will be required to use it afterwards. Internal controls and independent controls should be implemented.
• A planning of the CAPER transfer could ensure proper transfer of the assets and the highest level of availability. The European Data Protection Supervisor should be informed about the development and implementation of the transfer.

4 CONCLUSION

Technical partners within the CAPER project have been developing software and hardware tools that are able to meet all the requirements posed by the CAPER system. Therefore, Social Network Analyses (SNA), profiling and facial recognition are or will be at some stage unavoidable to test or to validate the tools offered to LEAs. In view of this fact, it is important to stress that technical partners will carry out the validation stage jointly with LEAs, taking advantage of the positive synergies established between them during the project but necessarily and clearly distinguishing the difference between a technical partner and a LEA (in terms of the use of personal data in a validation context).

Identified or identifiable entities or individuals will be considered personal data in the sense of the Data Protection Directive. If re-identification is possible, then the European and the National Data Protection Act will still apply. Therefore, in this document, we have indicated the general regulation and recommendations for such data processing activities. In all these situations, the general principles of data protection should be respected.

Researchers should then inform the research subjects (if any) of the purpose of the research, how their personal data will be used, who will have access and the identity of the controller. Other requirements are to keep data secure and to accept requests of access. Moreover, sensitive data will need complementary measures. International transfer of data or research publications will also require the respect for the institution overseas of the adequacy principle, or the minimal disclosure of personal data in the reports.

However, the Directive exemptions data processed “solely for the purpose of scientific research or (…) for the sole purpose of creating statistics”. This is true as far as data are not used for taking measures or decisions regarding any particular individual and there is no risk of breaking the privacy of the data subject.

Concretely, personal information can be processed for research purposes other than those for which it was originally obtained and it can be held indefinitely.

Researchers can also avoid the obligation to inform the research subject if it involves disproportionate effort. The reasons for doing so may be subject to legal challenge.

Significantly, the requirement to keep data secure still applies. Personal data should be protected from unauthorised access or accidental loss. One good practice would be fully anonymisation of collected data, in the sense that there is no way to re-identify the data subject. This has also a clear benefit: the Data Protection National Act does not apply in this case because such information is not considered to be personal data. Yet, this sensitive point must be discussed further by the Ethical Committee, because the effort must be also proportionate to the objectives of creating a usable and better tool to enhance national security.

Finally, tests and validations performed by technical partners in order to improve LEAs criminal investigations should take into account some good practices adopted to properly respect Data Protection regulations. This is a list of general recommendations that could be concretised in future deliverables:

- The collected data should be erased when there is no additional reason to keep them after the validation of a tool.
- If we want to store them for statistical reasons, then anonymisation should be considered (with preventions already mentioned).
- If re-identification of personal data is necessary for the research, then data should be kept secure.
- Data subjects should not be identified in the results of the research. Minimal disclosure of personal information in reports should be the rule.
- Researchers should always remember that data are collected for research purpose. Further processing for other purposes will be the exception and will need complementary safeguards. On the other hand, the data will never be used for taking measures or decisions regarding any particular individual.

5 ANNEXES/REFERENCES

5.1 References

5.1.1 Privacy issues references

. Alessandro ACQUISTI, From the Economics to the Behavioral Economics of Privacy: A Note, ICEB, LNCS 6005, pp. 23-26, 2010.

. Article 29 Working Party, Opinion 4/2007 on the concept of personal data, 2007, available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf

. Claude CASTELLUCCIA, Emiliano DE CRISTOFARO, Daniele PERITO, Private Information Disclosure from Web Searches (or how to reconstruct users' search histories), in Proceedings of the 2010 Privacy Enhancing Technologies Symposium (PETS), 2010, LNCS 6205, pp. 38-55.

. Catherine DWYER. Behavioral targeting: A case study of consumer tracking on levis.com, in Fifteen Americas Conference on Information Systems, 2009, available at: http://www.ftc.gov/os/comments/privacyroundtable/544506-00046.pdf

. Ann Cavoukian, Privacy by Design … Take the Challenge, 2009, available at http://www.privacybydesign.ca/publications/pbd-the-book/

. M. CHEW, D. BALFANZ, and B. LAURIE. (under)mining privacy in social networks. In Web 2.0 Security and Privacy workshop, 2008.

. Pam DIXON, World Privacy Forum, Consumer tips: How to opt-out of cookies that track you, 2009, available at: http://www.worldprivacyforum.org/cookieoptout.html

. The European Commission - DG Justice, Freedom and Security, Study on the economic benefits of privacy enhancing technologies (PETs), July 2010, available at: http://ec.europa.eu/justice/policies/privacy/studies/index_en.htm

. ENISA, Survey of accountability, trust, consent, tracking, security and privacy mechanisms in online environments, 2010, available at: http://www.enisa.europa.eu/act/it/library

. ENISA, *Study on data collection and storage in the EU [Deliverable – 2012-02-08].*

. ENISA, Cyber incident Reporting in the EU. An overview of security articles in the EU Legislation, August 2012, http://www.enisa.europa.eu/activities/Resilience-and-CIIP/Incidents-reporting/cyber-incident-reporting-in-the-eu

. European Data Protection Supervisor, *EURODAC Central UNIT. Inspection Report. June 2012, Case file: 2011-2013.*

. . European Data Protection Supervisor, *Security Audit VIS Central System. Summary Report. 1 June 2012.*

. European Network and Information Security Agency (ENISA), *Privacy, Accountability and Trust – Challenges and Opportunities*, Feb .2011.

. EUROPEAN COUNCIL, 2006. *Developing a Comprehensive and Coherent EU Strategy to Measure Crime and Criminal Justice: An EU Action Plan 2006-2010*.

. EUROPEAN UNION, 2005. *The Hague Programme – Strengthening Freedom, Security and Justice in the European Union*, OJ C53 C.F.R.

. EUROPEAN UNION, 2010. *The Stockholm Programme – An Open and Secure Europe Serving and Protecting Citizens in 2010*, OJ C115/01 C.F.R.

. FERRANTE, A., 2009. The Use of Data-Linkage Methods in Criminal Justice Research: A Commentary on Progress, Problems and Future Possibilities, *Current Issues in Criminal Justice,* Vol. 20, No. 3, pp. 378-392.

. FORD, D. V., K. H. JONES, J. P. VERPLANCKE, R. A. LYONS, G. JOHN, G. BROWN and K. LEAKE, 2009. “The SAIL Databank: Building a National Architecture for e-health Research and Evaluation”, *BMC Health Services Research*, Vol. 9, p. 157.

. Kate GREENE, Reality mining, in MIT Technology review, 2008, available at: http://www.technologyreview.com/read_article.aspx?id=20247&ch=%20specialsections&sc=emerging08&pg=1&a=f

. Saikat GUHA, Bin CHENG, Alexey REZNICHNKO, Paul FRANCIS. Privad: Practical Privacy in Online Advertising. MaxPlanck Institute for Software Systems Technical report MPI-SWS-2010-001.

. HILDEBRANDT, M., “Profiling: from data to knowledge”, In Datenschutz und Datensicherheit - DuD Volume 30, Number 9, pp.548-552, available at: http://www.springerlink.com/content/k313374344883l21/

. Mireille HILDEBRANDT (Ed.), FIDIS Deliverable D 7.12: Biometric behavioral profiling and transparency enhancing tools, 2009, www.fidis.net

. International Data Protection Authorities, "International Standards on the Protection of Personal Data and Privacy: The Madrid Resolution", Nov. 2009, available at: http://www.gov.im/lib/docs/odps//madridresolutionnov09.pdf

. Paul OHM. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Review, 2010.

. B. KRISHNAMURTHY and C. WILLS. Privacy diffusion on the web: a longitudinal perspective (updated graphs), September 2009, available at: http://www.ftc.gov/os/comments/privacyroundtable/544506-00009.pd

. B. KRISHNAMURTHY and C. WILLS. Privacy leakage in mobile online social networks. In WOSN‘10: Proceedings of the third workshop on Online social networks, 2010.

. John KRUMM, Ubiquitous advertising: The killer application for the 21st century,. IEEE Pervasive Computing, January 2010.

. Teresa LUNT, Paul AOKI, Dirk BALFANZ, Glenn DURFEE, Philippe GOLLE, Diana SMETTERS, Jessica STADDON, Jim THORNTON, Tomas URIBE. Protecting the Privacy of Individuals in Terrorist Tracking Applications. Technical Report AFRL-IF-RS-TR-2005-131, US Air Force Research Laboratory, 2005.

. MINISTRY OF JUSTICE, 2011. *A Guide to Criminal Justice Statistics*, UK Government: Ministry of Justice.

. Prateek MITTAL and Nikita BORISOV. Information leaks in structured peer-to-peer anonymous communication systems. In Paul SYVERSON, Somesh JHA, and Xiaolan ZHANG, editors, CCS'08: Proceedings of the 15th ACM Conference on Computer and Communications Security, pages 267-278. ACM Press, 2008.

. NATIONAL STATISTICS BOARD, 2009. *National Statistics Board Strategy Statement 2009–2014*, Dublin: National Statistics Board.

. OXMAN, Stephen A., *Exemptions to the European Union Personal Data Privacy Directive: Will They Swallow the Directive?*, 24 *B.C. Int'l & Comp. L. Rev*. 191 (2000), http://lawdigitalcommons.bc.edu/iclr/vol24/iss1/8

. Protecting Consumer Privacy in an Era of Rapid Change, Preliminary FTC Staff Report, Federal Trade Commission, December 2010.

. ROGAN, Mary, “Improving Criminal Justice Data and Policy”, *The Economic and Social Review*, Vol. 43, nº 2, *Summer*, 2012, p. 303-323.

. V. TOUBIANA, A. NARAYANAN, D. BONEH, H. NISSENBAUM, and S. BAROCAS, Adnostic: Privacy Preserving Targeted Advertising, ISOC Network and Distributed System Security. Annual Network and Distributed Systems Security Symposium (NDSS), 2010.

. TRUTWEIN, B., D. HOLMAN and D. ROSMAN, 2006. “Health Data Linkage Conserves Privacy in a Research-Rich Environment”, *Annals of Epidemiolog*y, Vol. 16, p. 279.

. VERSCHUUREN, M., G. BADEYAN, J. CARNICERO, M. GISSLER, R. P. ASCIAK, L. SAKKEUS, 2008. “The European Data Protection Legislation and its Consequences for Public Health Monitoring: A Plea for Action”, *The European Journal of Public Health,* Vol. 18, No. 6, pp. 550-551.

. WONDRACEK, G., HOLZ, T., KIRDA, E., and KRUEGEL, K., A Practical Attack to De-anonymize Social Network Users, in Proc. IEEE Symposium on Security and Privacy, 2010, pp.223-238.

. Aydan R. YUMEREFENDI, Jeffrey S. CHASE. Strong Accountability for Network Storage. USENIX Conference of File and Storage Technologies (FAST), 2007.

5.1.2 Multimedia formats references

Adams, M. (2012). *The JasPer Project Home Page*. Retrieved June 2012, from The JasPer Project Home Page: http://www.ece.uvic.ca/~frodo/jasper/

Amin, A., Qureshi, H. A., Junaid, M., & Habib, M. Y. (2011). Modified run length encoding scheme with introduction of bit stuffing for efficient data compression. *International Conference for Internet Technology and Secured Transactions (ICITST)* , 668-672.

Andersen, J. V. (2001, January 5). *Computer Graphics Metafile (CGM)*. Retrieved May 2012, from http://myeasycopy.com/cgm-white-paper

Burnett, I., Van de Walle, R., Hill, K., Bormans, J., & Pereira, F. (2003). MPEG-21: Goals and Achievements. *IEEE Multimedia , 10* (4), 60-70.

Chang, S. F., Sikora, T., & Puri, A. (2001). Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology , 11* (6), 688-695.

Deng-Pan, Y., Yao-Bin, M., Yue-Wei, D., & Zhi-Quan, W. (2005). A Multi-feature Based Invertible Authentication Watermarking for JPEG Images. *Lecture Notes in Computer Science , 3304/2005*, 152-162.

Fridrich, J., Goljan, M., & Du, R. (2001). Invertible Authentication Watermark for JPEG Images. *Proceeding ITCC '01 Proceedings of the International Conference on Information Technology: Coding and Computing* , 223-227.

Graphics Interchange Format (GIF) - Wikipedia. (n.d.). Retrieved 2012, from http://en.wikipedia.org/wiki/Graphics_Interchange_Format

Hannuksela, M. M., Döhla, S., & Murray, K. (2012). The DVB File Format. *IEEE Signal Processing Magazine* , 148-153.

Howard, P. G., Kossentini, F., Martins, B., Forchhammer, S., & Rucklidge, W. J. (1998). The emerging JBIG2 standard. *IEEE Transactions on Circuits and Systems for Video Technology , 8* (7), 838-848.

ISO/IEC10918-1. (1992). ISO/IEC 10918-1: Requirements and guidelines.

ISO/IEC10918-2. (1994). ISO/IEC 10918-2: Compliance testing.

ISO/IEC10918-3. (1996). ISO/IEC 10918-3: Extensions.

ISO/IEC10918-4. (1998). ISO/IEC 10918-4: Registration of JPEG profiles, SPIFF profiles, SPIFF tags, SPIFF colour spaces, APPn markers, SPIFF compression types and Registration Authorities.

ISO/IEC14492. (2001). ISO/IEC 14492: Information technology – Lossy/lossless coding of bi-level images.

ISO/IEC15444-1. (2000, December). ISO/IEC 15444-1: JPEG2000 image coding system. Part 1: Core coding system.

ISO/IEC15444-2. (2004, May). ISO/IEC 15444-2: JPEG2000 image coding system. Part 2: Extensions.

ISO/IEC15444-3. (2002, September). ISO/IEC 15444-3: JPEG2000 image coding system. Part 3: Motion JPEG2000.

ISO/IEC15444-6. (2003, October). ISO/IEC 15444-6: JPEG2000 image coding system. Part 6: Compound image file.

ISO/IEC15444-8. (2006, July). ISO/IEC 15444-8: JPEG2000 image coding system. Part 8: Security.

ISO/IEC15444-9. (2005, December). ISO/IEC 15444-9: JPEG 2000 image coding system - Part 9: Interactivity tools, APIs and protocols.

ISO/IEC-FDIS10918-5. (2009). ISO/IEC FDIS 10918-5: .

JJ2000. (June de 2012). *JJ2000: A pure Java JPEG 2000 image codec*. Recuperado el June de 2012, de http://code.google.com/p/jj2000/

Liang, X. (2008). Reversible Authentication Watermark for Image. *Proceedings of the World Congress on Engineering and Computer Science* , 728-733.

Lin, C.-Y., & Chang, S.-F. (2001). A Robust Image Authentication Method Distinguishing JPEG Compression from Malicious Manipulation. *IEEE Transactions on Circuits and Systems of Video Technology , 11* (2), 153-168.

Portable Network Graphics (PNG) - Wikipedia. (2012). Retrieved from http://en.wikipedia.org/wiki/Portable_Network_Graphics

Rabbani, M., & Joshi, R. (2002). An overview of the JPEG 2000 still image compression standard. *Signal Processing: Image Communication , 17* (1), 3-48.

Raw image format (RAW) - Wikipedia. (2012). Retrieved from http://en.wikipedia.org/wiki/Raw_image_format

Said, A. (2004). Efficient and reliable Dynamic Quality Control for Compression of Compound Document Images. *Proceedings of the International Conference on Image Processing (ICIP)* , 2869-2872.

Sikora, T. (2001). The MPEG-7 Visual Standard for Content Description – An Overview. *IEEE Transactions on Circuits and Systems for Video Technology , 11* (6), 696-702.

Skodras, A., Christopoulos, C., & Ebrahimi, T. (2001). JPEG2000: The upcoming still image compression standard. *ELSEVIER Pattern Recognition Letters , 22* (12), 1337–1345.

Skodras, A., Christopoulos, C., & Ebrahimi, T. (2001). The JPEG 2000 still image compression standard. *IEEE Signal Processing Magazine , 18* (5), 36-58.

Tagged Image File Format (TIFF) - Wikipedia. (2012). Retrieved from http://en.wikipedia.org/wiki/Tagged_Image_File_Format

Taubman, D. S., & Marcellin, M. W. (2002). *JPEG2000: Image Compression Fundamentals, Standards, and Practice.* Springer.

Taubman, D., & Marcellin, M. (2002). JPEG2000: Standard for interactive imaging. *Proceedings of the IEEE , 90* (8), 1336-1357.

The University of New South Wales. (2012, June). *Kakadu Software*. Retrieved June 2012, from http://www.kakadusoftware.com/

Troncy, R., Bailer, W., Hausenblas, M., & Hofmair, P. (2006). Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7 Profiles. *Lecture Notes in Computer Science , 4306*, 41-55.

W3C. (2012). *SVG Overview*. Retrieved 5 2012, from http://www.w3.org/Graphics/SVG/WG/wiki/SVG_Overview

Wallace, G. W. (1992). The JPEG Still Picture Compression Standard. *IEEE Transactions on Consumer Electronics , 38* (1), 18-34.

Weinberger, M. J., Seroussi, G., & Sapiro, G. (1996). LOCO-I: A low complexity, context-based, lossless image compression algorithm. *Proceedings of Data Compression Conference (DCC)* , 140-149.

Weinberger, M. J., Seroussi, G., & Sapiro, G. (2000). The LOCO-I Lossless Image Compression Algorithm: Principles and Standarization into JPEG-LS. *IEEE Transactions on Image Processing , 9* (8), 1309-1324.

Welch, T. (1984). A Technique for High-Performance Data Compression. *Computer , 17* (6), 8-19.

Zhang, W., Chen, B., & Yu, N. (2012). Improving Various Reversible Data Hiding Schemes Via Optimal Codes for Binary Covers. *IEEE Transactions on Image Processing , 21* (6), 2991-3003.

5.2 Privacy issues annexes

. Opinion 02/2012 on facial recognition in online and mobile services, adopted on 22 March 2012.

WP 192 of the Article 29 Data Protection Working Party

http://ec.europe.eu/justice/data-protection/index_en.htm

. Opinion 04/2012 on Cookie Consent Exemption, adopted on 7 June 2012

WP 194 of the Article 29 Data Protection Working Party

http://ec.europe.eu/justice/data-protection/index_en.htm

. Opinion 01/2010 on the concepts of “controller” and “processor”, adopted on 16 February 2010.

WP 169 of the Article 29 Data Protection Working Party

http://ec.europe.eu/justice/data-protection/index_en.htm

. Opinion 02/2010 on online behavioural advertising, adopted on 22 June 2010.

WP 171 of the Article 29 Data Protection Working Party

http://ec.europe.eu/justice/data-protection/index_en.htm

. Opinion 16/2011 on EASA/IAB Best Practice Recommendation on online behavioural advertising, adopted on 08 December 2011.

WP 188 of the Article 29 Data Protection Working Party

http://ec.europe.eu/justice/data-protection/index_en.htm

. Opinion 1/2008 on on data protection issues related to search engines, adopted on 4 April 2008.

WP 148 of the Article 29 Data Protection Working Party

http://ec.europe.eu/justice/data-protection/index_en.htm

. Best Practice Recommendation on online behavioural advertising ("EASA/IAB Code") http://www.easaalliance.org/binarydata.aspx?type=doc/EASA_BPR_OBA_12_APRIL_2011_CLEAN.pdf/download , complemented by the website www.youronlinechoices.eu .

. Treaty of Lisbon, Consolidated versions of The Treaty on European Union and The Treaty on the Functioning of the European Union, Charter of Fundamental Rights of the European Union, 2010, available at: http://europa.eu/lisbon_treaty/full_text/index_en.htm

. Convention for the Protection of Human Rights and Fundamental Freedoms, The Council of Europe, European Court of Human Rights, available at:

http://www.echr.coe.int/NR/rdonlyres/D5CC24A7-DC13-4318-B457-5C9014916D7A/0/ENG_CONV.pdf

. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995, available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

. Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal

data by the Community institutions and bodies and on the free movement of such data, available at:

http://ec.europa.eu/justice/policies/privacy/docs/application/286_en.pdf

. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), available at:

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:NOT

.Council Framework Decision 2008/977/JHA of 27 November 2008 on the protection of personal data processed in the framework of police and judicial cooperation in criminal matters, available at:

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:350:0060:01:EN:HTML

. Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws, available at:

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32009L0136:EN:NOT

5.3 Image and Video standards for the Caper System and Multimedia format annexes

Nowadays, Internet has become the most important channel to share information with the whole world. Most traditional activities including music, film, television, newspapers or books have been reshaped or redefined by the Internet. In addition, the Internet has also enabled new forms of human interactions through instant messages, forums and social networks. It is well-known that these new channels of communication are focused on digital contents such as images and videos.

There are many image and video file formats and representations in the literature which provide different features to final users. These different formats are usually not compatible among them causing an interoperability problem. Consequently, International Standards and Recommendation Organizations make an important effort to develop common formats and representations aimed at providing useful features. The main goal is to improve digital content sharing. In our field, the International Telecommunication Union[21] (ITU), International Standard for Organization[22] (ISO) and International Electrotechnical Commission[23] (IEC) are non-governmental organizations focused on developing and coordinating standards for telecommunications, electronics and related technologies. In this deliverable we are focused on these three organizations and their standards and recommendations to improve the share of digital contents within the CAPER project.

The CAPER system is able to track relevant digital information for its purposes. Therefore, these contents have to be stored in a repository aimed at exchanging information captured by the CAPER system. In this situation, there are two main drawbacks: i) CAPER is composed of a wide consortium where different countries and systems have to interact; and ii) storing all captured information means a high storage requirement. Therefore, the main target of this Section is twofold. The first point is to analyze different standards for digital images and videos to facilitate access to this content for the members of the project. Then, the second point is to study recommendations and standards which provide a certain degree of compression to mitigate the storage requirement that arises within the CAPER platform.

The sections below analyze different recommendations, standards and file formats that are able to manage still images and videos. Particularly, Section 3.1 briefs image representation methods and recommendations and standards that fit in the CAPER platform. Moreover, the Joint Photographic Experts Group[24] (JPEG) committee that is specifically devised on still image management is also considered. On the other hand, Section 3.2 is focused on video representation formats and algorithms that are able to manage this kind of information. In addition, Movie Pictures Experts Group[25] (MPEG) committee that is devised to deal with video and movie files is also summarized in this section. Both mentioned Sections present algorithms and file formats which are able to pose main drawbacks that may arise within the CAPER platform.

5.3.1 Image

Digital image has become an important source of information in the Internet Era. Therefore, international organizations, companies and academic institutions worldwide have been working in different recommendations and standards specifically devised for their management. Moreover, over the past years the improvement of sensors aimed at capturing these images increases their weight producing a high storage requirement. Therefore, some of these recommendations and standards also consider compression issue for digital images to facilitate the storage and transmission over the net.

Although compression algorithms and their features are out of the scope of this deliverable, two main concepts are introduced to facilitate the understanding of recommendations and standards proposed for the CAPER system. The first concept is *lossless compression* that allows the exact original image to be reconstructed from the compressed data. On the other hand, the second concept involves *lossy compression* that only allows an approximation of the original image to be reconstructed, in exchange for better compression rates.

[21] International Telecommunication Union (ITU): http://www.itu.int/en/Pages/default.aspx
[22] International Standard for Organization (ISO): http://www.iso.org
[23] International Electrotechnical Commission (IEC): http://www.iec.ch
[24] JPEG committee http://www.jpeg.org
[25] MPEG committee http://mpeg.chiariglione.org

The standard or recommendation performed within the CAPER system to store and manage digital images must meet five main goals: i) to facilitate access to the information acquired and processed by the CAPER system through the adoption of international standards and the recommendations; ii) to provide a certain degree of compression to mitigate storage requirement that arises within the CAPER system; iii) to provide both lossless and lossy strategies; iv) to perform an efficient management of compressed files; and v) to allow the inclusion of metadata to store the result of the analysis performed by the CAPER system in conjunction with compressed images. Furthermore, some other desirable extra features to be considered are: interactive and progressive transmission and security and other related issues.

This Section is organized as follows: Section 3.1.1 introduces the main features provided by two image representations such as vector and raster file formats for still image representation; Section 3.1.2 briefs standards and recommendations proposed by the JPEG committee for still image management and highlights their pros and cons. The JPEG committee has been working on this issue for years and it has become one of the most influential Work Groups. Finally, Section 3.1.3 discusses the suitability of JPEG standards for the CAPER system.

5.3.2 Image representation and file formats

The most common image representation formats are: vector[26] and raster[27] file formats.

Vector graphics representation uses geometrical primitives such as points, lines, curves and shapes or polygons. All of them are based on mathematical expressions to represent images in computers. Vector image formats contain a geometric description which can be rendered smoothly at any desired display size. Shape borders can have associated properties such as thickness and colour. Shapes and regions that are fully enclosed by lines or curves can be filled with a chosen colour. Moreover, arrows, shading effects and shadows can be added to filled areas. A vector representation of an image should be described in such a way that the image can be unambiguously reconstructed from the representation. Figure 1 represents a set of points and its graphical primitive that are able to reconstruct the image.

There are several formats that perform this image representation (see vector formats for further information). However, in this deliverable we are only focused on two common formats within this image representation: Computer Graphics Metafile (CGM) and Scalable Vector Graphics (SVG).

- On the one hand, **CGM** (Andersen, 2001) is a very feature-rich format which attempts to support the graphic needs of many general fields (graphic arts, technical illustration, cartography and so on). It was specifically designed as a common format for the platform-independent interchange of bitmap and vector data, and for the use in conjunction with a variety of input and output devices. Although CGM incorporates extensions designed to support bitmap, files in CGM format are used primarily to store vector information. All graphical elements can be specified in a textual source file that can be compiled into a binary file. Nevertheless, CGM is not widely supported

[26] For further information see: http://en.wikipedia.org/wiki/Vector_graphics
[27] For further information see: http://en.wikipedia.org/wiki/Raster_graphics

for web pages and has been supplanted by other formats. However, it is still prevalent in engineering, aviation, and other technical applications.

- On the other hand, **SVG** (W3C, 2012) is an open standard which is created and developed by the W3C to address the need (and attempts of several corporations) for an all-purpose vector format for the web and otherwise. SVG images and their behaviours are defined in XML text files. This means that they can be searched, indexed, scripted and, if need be, compressed. As XML files, SVG images can be created and edited with any text editor, but it is often more convenient to create them with drawing programs. However, the SVG format does not have a compression scheme of its own.

CGM and SVG images, being text, contain many repeated fragments of text, so they are well suited for lossless compression algorithms. Nevertheless, as we mentioned above, lossless compression is able to achieve lower compression rates than lossy compression. Furthermore, these standards do not consider the efficient management of compressed images and its corresponding metadata fields, interactive and progressive transmission, data integrity and other desirable issues for the CAPER platform.

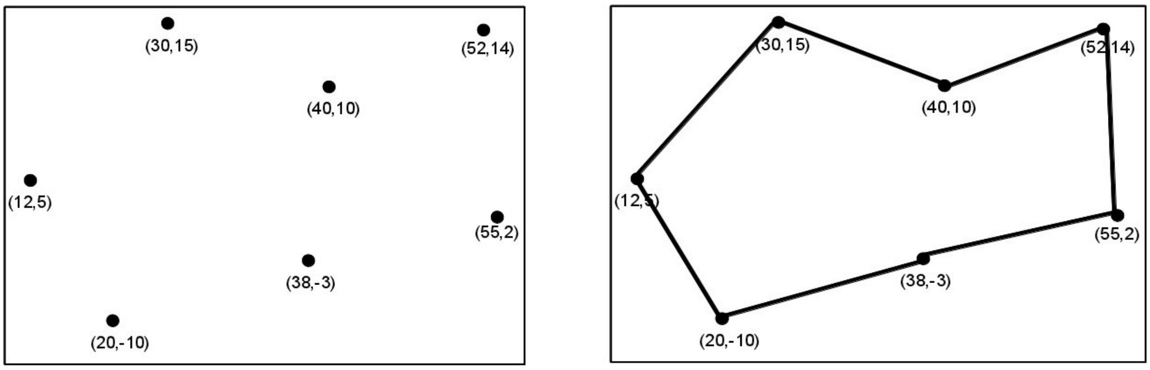


Figure 1: Vector image representation. Left: file store points and geometrical primitives that allow the reconstruction of the original image. Right: original image after the reconstruction.

Raster image representation or bitmap is a dot matrix data structure representing a generally rectangular grid of pixels or points of color. A bitmap is technically characterized by the width and height of the image in pixels and by the number of bits per pixel (color depth), which determines the number of colors that it can represent. Figure 2 shows a *de facto* well-known image standard, lena, and how it is constructed as a grid of different values that represent colors in a specified number of layers. Some of the most common file formats are the following:

- The **BMP format** is used to store bitmap digital images independently of the display device for any operative system. The BMP file format is capable of storing 2D digital images of arbitrary width, height and resolution, both monochrome and colour and optionally with data compression. This file format is the MS-Windows standard format. It holds black & white, 16-color, 256-color and “true colour” images. The palletized 16-color and 256-color images may be compressed via run length encoding (Amin, Qureshi, Junaid, & Habib, 2011).

- A **raw image file (RAW)** (Raw image format (RAW) - Wikipedia, 2012) contains minimally processed data from the image sensor. Raw files are so named because they are not yet processed and therefore are not ready to be printed or edited with a graphics editor. Normally, the image is processed by a raw converter where precise adjustments can be made before conversion to a processed file format such as JPEG or TIFF for further manipulation.

- The **Tagged Image File Format (TIFF)** (Tagged Image File Format (TIFF) - Wikipedia, 2012) is a file format for storing raster images, popular among graphic artist, the publishing industry, and amateur and professional photographers. The TIFF format is widely supported by image-manipulation applications such as scanning, faxing, word processing and so on. Adobe Systems owns the copyright to the TIFF specifications. Although TIFF has not had a major update since 1992, several technical notes have been published with minor extensions to the format.

- **Portable Network Graphics (PNG)** (Portable Network Graphics (PNG) - Wikipedia, 2012) is a bitmapped image format that employs lossless compression. PNG was created to improve upon and replace Graphics Interchange Format (Graphics Interchange Format (GIF) - Wikipedia) as an image-file format not requiring a patent license. The motivation for creating the PNG format was in early 1995, after it became known that the Lempel-Ziv-Welch (LZW) data compression algorithm (Welch, 1984) used in the GIF format was patented by Unisys. There were also other problems with the GIF format that made a replacement desirable, notably its limit of 256 colours at a time when computers able to display far more than 256 colours were growing common. Although GIF allows for animation, it was decided that PNG should be a single-image format.

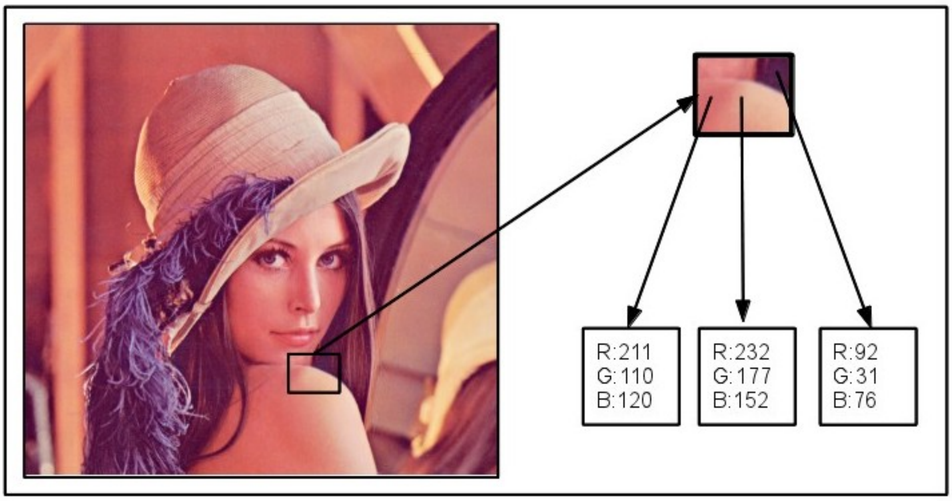


Figure 2: Raster image representation as a grid of values. *Left*: Lena colour image that is a *de facto* standard in the image compression field. *Right*: Each point is represented in three layers (Red, Green and Blue) by a number.

Raster and Vector image representations have some properties that distinguish these data types:

- **Generality.** Raster representation can be used for any two-dimensional image, whereas vector representation makes sense only for images that can be well approximated by geometric lines and shapes.

- **Size** of raster representation tends to be significantly bigger than vector image representations. The size of a vector image depends on its complexity.

- **Editing options**. Vector representation is preferred by professionals who work with line drawings or animators and who want to produce several variations of an initial scene. On the other hand, raster editing tools that use special effect filters can blur or change the overall colour balance of the image.

- **Scalability**. Images in both raster and vector format can be viewed at different scales, for example when zooming in or out of the image. If width and height of a raster format image are doubled, the total area covered by the image increases by a factor of four. Consequently, each of the pixels in the image covers an area which is a factor of four larger than the covered before resulting in a grainy, pixelated look. In contrast, lines, curves and shapes of images in vector format remain sharp even as the image is scaled. As a result, vector representation works well for fonts because characters of an alphabet can be represented using simple lines and curves.

There are situations when working with vector tools is the best practice and situations when working with raster tools is the best practice. There are times when both formats come together. An understanding of the advantages and limitations of each technology and the relationship between them is most likely to result in efficient and effective use of tools.

Plainly, there are many formats and representations able to provide interesting features: i) standard and recommended implementations promoted by international organizations; and ii) a certain degree of compression aimed at ease the store and transmission over the net. Nevertheless, to the best of our knowledge, all of these formats and representations are usually devised to meet requirements in specific fields. In addition, they do not perform suitable properties such as lossless and lossy strategies, interactive and progressive transmission, security issues, efficient management of compressed files and so on.

The CAPER system must be able to share the captured and analysed information from OSI sources. Consequently, how these images are stored becomes an important issue within the system design task. In this deliverable, we propose to take into account the work performed by the JPEG committee because their standards and recommendations usually meet main features that CAPER system faces in this scenario.

5.3.3 JPEG committee

JPEG committee has developed several standards which provide important features required by users and applications in this scenario. The committee meets at least three times a year to discuss and create standards for still image compression. The most important recommendation is the Original JPEG[28] that became a *de facto* standard in Internet.

[28] When most people talk about JPEG, they are referring to this particular standard and its implementation, not JPEG committee.

However, the JPEG committee has developed several standards which provide different features (see Table 1).

Algorithm	Details	Link
JPEG-LS	Lossless and near-lossless compression standard of continuous-tone image. The core is performed by LOCO-I (LOw COmplexity LOsless COmpression for Images).	ITU-T T.87 Recommendation ISO/IEC 14495:2003 Standard
JBIG2	Bi-level images compression lossless and lossy.	ITU-T T.88 Recommendation ISO/IEC 14491:2001 Standard
JPEG	Common still image algorithm of lossy compression. The degree of compression could be adjusted. The standard is composed of 5 parts.	ITU-T T.81 Recommendation ISO/IEC 10918:1994 Standard
JPEG2000	Image compression standard and coding system. The standard considers many desired features and produces a flexible codestream.	ITU-T T.800 Recommendation ISO/IEC 15444:2000 Standard

Table 1: Recommendations and Standards for still images developed by the JPEG committee focused on compression and management.

The main idea of this work is to propose a standard which fulfills five main goals: i) facilitating access to the information acquired and processed by the CAPER system through the adaption of international standards and recommendations; ii) providing a certain degree of compression to mitigate storage requirement that arises within the CAPER system; iii) providing both lossless and lossy strategies; iv) performing an efficient management of compressed files; and v) allowing the inclusion of metadata to store the result of the analysis performed by the CAPER system in conjunction with compressed images. Some extra features are: interactive and progressive transmission and security and other related issues. Table 1 shows five different algorithms belonging to the JPEG committee. Nevertheless, to the best of our knowledge, not all of them are suitable to meet the requirements posed by the CAPER system.

- The first algorithm considered in Table 1 is the **JPEG-LS** standard (Weinberger, Seroussi, & Sapiro, The LOCO-I Lossless Image Compression Algorithm: Principles and Standarization into JPEG-LS, 2000). This algorithm performs the **LOCO-I** algorithm (introduced in (Weinberger, Seroussi, & Sapiro, LOCO-I: A low complexity, context-based, lossless image compression algorithm, 1996)) as the core of the ISO/ITU standard aimed at lossless and near-lossless compression of continuous tone images. It is considered as a low-complexity algorithm which relies in a context-based encoding based on a simple fixed context model which approaches the capability of the more complex universal techniques for capturing high-order dependencies.

- **JPEG-LS** is an ISO/IEC standard specifically devised to perform lossless compression. In spite of providing state-of-the-art lossless compression performance, it does not consider to apply a lossy strategy which achieves better compression ratios. However, in our scenario, after the analysis applied to every captured image, the CAPER system must be able to perform a lossy strategy to mitigate the storage requirements. In addition, the JPEG-LS algorithm has several drawbacks in this scenario: not to implement lossy and lossless strategies; not to consider the management of metadata information in conjunction with the compressed image; not to perform an efficient management of the compressed image; not to provide a progressive reconstruction of the original image.

- The third algorithm proposed in Table 1 is **JBIG2** (Howard, Kossentini, Martins, Forchhammer, & Rucklidge, 1998) that becomes a standard in (ISO/IEC14492, 2001). This algorithm is specifically devised to encode only bi-level images and documents in a lossless and lossy ways. In addition, metadata boxes are also considered in this file format to store some additional information about documents and to provide semantic capabilities.

- **JBIG2** is also a well-known compression algorithm since it is an ISO/IEC standard that achieves state-of-the-art compression performance. Although it considers the management of metadata information, JBIG2 is specifically devised to encode bi-level images. This restriction is its main drawback since the CAPER system is suitable to capture every kind of image present in the Internet (webs, social networks, mass media, and so on). Furthermore, the algorithm has drawbacks related to the CAPER system: not to consider an efficient management, not to perform an efficient management of the compressed image; not to provide a progressive reconstruction of the original image.

- Next algorithm in the table is the **JPEG** which is summarized in (Wallace, 1992). JPEG is a standardization effort known by the acronym JPEG. It is the first international digital image compression standard for continuous-tone still images, both greyscale and colour. The main goal of this work was to develop a general-purpose compression standard to meet the needs of almost all continuous-tone still-images applications. Section 3.1.2.1 provides an overview of this popular standard.

- **JPEG** is a general-purpose standard that meets several requirements posed by CAPER system. However, another standard such as JPEG2000 provides a more efficient management of images and it does not consider the inclusion and management of metadata in conjunction with the compressed image.

- One of the latest algorithms developed by the JPEG committee is the **JPEG2000**. JPEG2000 has become a powerful standard that provides even more features than those initially planned. The technologies supported by the standard have been described in many different works. To mention only some of them, the JPEG2000 features and the most important techniques used in it are reviewed in (Skodras, Christopoulos, & Ebrahimi, The JPEG 2000 still image compression standard, 2001) and (Skodras, Christopoulos, & Ebrahimi, JPEG2000: The upcoming still image compression standard, 2001); an in-depth overview is described in (Rabbani & Joshi, 2002) and (Taubman & Marcellin, 2002). Although the standard was initially structured in six parts, seven additional parts were proposed in 2001. Section 3.1.2.2 provides an overview of this standard.

- **JPEG2000** fulfils the main goals proposed by the CAPER system. In addition, it addresses some extra features such as access control, security tools, interactive transmission protocols, and so on.

The analysis suggests that the most suitable algorithms for the CAPER system are JPEG and JPEG2000. Next sections introduce detailed information about both standards.

5.3.3.1 JPEG

JPEG is a general-purpose standard proposed by the JPEG committee in (ISO/IEC10918-1, 1992) which meets many requirements for a wide set of scenarios. In the last years, the JPEG algorithm has become a *de facto* standard for webs, social networks and other many common applications within Internet.

Regarding technical specifications, JPEG is based on the Discrete Cosine Transform (DCT) which applies the compression taking blocks of 8 by 8 pixels from the original image. Then, a quantizer stage and an entropy coder are applied to obtain the compressed image. Figure 3 depicts an overview of the general scheme proposed for the JPEG algorithm. Although the compression architecture has a low complexity feature, it has drawbacks such as the lack of consideration of metadata information in the standard scheme and the non-progressive recovery of images.

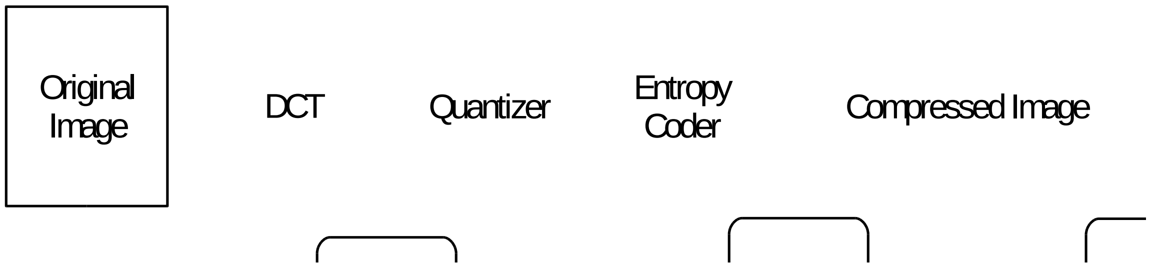


Figure 3: JPEG compression scheme

On the other hand, many works in the literature have proposed new features to those proposed by the JPEG committee. For instance, in (Zhang, Chen, & Yu, 2012) JPEG images are used to add Reversible Data Hiding (RDH) which is able to perform a watermarking process aimed at detecting manipulations within the compressed images; other related works within watermarking field are: (Fridrich, Goljan, & Du, 2001), (Liang, 2008) and (Deng-Pan, Yao-Bin, Yue-Wei, & Zhi-Quan, 2005); and how to differentiate malicious manipulations from acceptable manipulations (*e.g.* compression) of JPEG images is the main idea in (Lin & Chang, 2001).

JPEG Part	Description
Part 1	Requirements and guidelines (ISO/IEC10918-1, 1992)
Part 2	Compliance testing (ISO/IEC10918-2, 1994): rules and checks for software conformance to Part 1.
Part 3	Extensions (ISO/IEC10918-3, 1996): set of extensions to improve Part 1, including the SPIFF file format.

Part 4	Registration of JPEG profiles, SPIFF profiles, SPIFF tags, SPIFF colour spaces, APPn markers, SPIFF compression types and Registration Authorities (REGAUT) (ISO/IEC10918-4, 1998). Methods for registering some of the parameters used to extend JPEG
Part 5	JPEG File Interchange Format (ISO/IEC-FDIS10918-5, 2009). A popular format which has been the *de-facto* file format for images encoded by the JPEG standard. In 2009, the JPEG Committee formally established an Ad Hoc Group to standardize JFIF as JPEG Part 5.

Table 2: Description of the 5 parts of the JPEG standard

Table 2 shows the parts considered by the JPEG committee. Part 1 implements the core coding system and Parts 3 and 4 are related to the Still Picture Interchange File Format (SPIFF) which enables JPEG and other codestreams to be exchanged between a wide variety of platforms and applications. It is devised to transcoding between some other formats such as other JPEG files, JBIG1, JBIG2 (also reviewed in this deliverable), and other formats. For further information about SPIFF file format, please see (ISO/IEC10918-3, 1996). The SPIFF file format could be useful in compound documents compression as in (Said, 2004).

5.3.3.2 JPEG2000

One of the latest algorithms developed by the JPEG committee is the JPEG2000 that is composed of 12 different parts. The main purposes of each part are summarized in Table 3. Since a detailed description for the whole standard is not a goal in this deliverable, we are only focused on 5 parts (Parts 1, 2, 3, 8, 9). For interested readers “Annex I” is focused on JPEG2000 Part 6. Among other features, the JPEG2000 Part 1 provides scalability by quality, spatial location, resolution and component. These scalabilities fulfill most of the requirements of applications and scenarios where images are used.

JPEG2000 Part	**Description**
Part 1	Core coding system (ISO/IEC15444-1, 2000): description of the minimal decoder and a simple file format. This part has a limited number of options in order to facilitate the interchange among applications. It is the basis of the other parts, and allows lossy and lossless compression for still images.
Part 2	Extensions (ISO/IEC15444-2, 2004): extensions of the core coding system, providing advanced coding features which can be used to enhance the coding performance or to manipulate unusual data types. This part also provides an enhanced file format.
Part 3	Motion JPEG2000 (ISO/IEC15444-3, 2002): extensions of the core coding system devised to support the manipulation of image sequences (motion).

Part 4	Conformance testing: information for the compliance and conformance among JPEG2000 implementations.
Part 5	Reference Software: two implementations of the core coding system: JJ2000, developed in Java, and JasPer, developed in C.
Part 6	Compound image file format (ISO/IEC15444-6, 2003): additional file format for tailored and compound documents.
Part 7	This part has been abandoned.
Part 8	Secure JPEG2000 (ISO/IEC15444-8, 2006): description of a file syntax for interpreting secure image data and a normative process for registering security tools.
Part 9	Interactivity tools, APIs and protocols (ISO/IEC15444-9, 2005): description of the transmission protocol JPIP, devised to interactively transmit JPEG2000 images.
Part 10	Volumetric JPEG2000: coding of volumetric data, providing enhanced coding features for floating point data.
Part 11	JPEG2000 Part 11 Wireless JPEG2000: description of error protection techniques for JPEG2000 files aimed at detecting and correcting errors produced during data transmission.
Part 12	JPEG2000 Part 12 ISO base media file format: definition of the ISO file media file, providing an extensible format which facilitates interchange, management and editing.
Part 13	JPEG2000 Part 13 An entry level JPEG2000 encoder: it defines a normative entry level JPEG2000 encoder providing one or more optional complete encoding paths that use various features defined in ISO/IEC 15444.

Table 3: Description of the 13 parts of the JPEG2000 standard.

Part 1: Core Coding System

The basis of the JPEG2000 standard is composed by two main stages: 1) an encoder to compress the image producing a binary file that contains the compressed image (JP2 file format); and 2) a decoder to decompress the binary stream, obtaining a recovered image.

From the JP2 file, the JPEG2000 Decoder is able to obtain either a recovered image with different degrees of quality (scalability by quality), or a reduced version of the original image (scalability by resolution), or a random access to the JP2 binary file aimed at recovering only certain regions within the original image (scalability by spatial location), or a concrete component of the input image (scalability by component). One of the most important features of JPEG2000 is the possibility to decode a selected image taking into account these scalabilities without needing to decompress the whole compressed file *(compress once decompress many ways)*.

The JP2 file format contains the compressed image and the required information to allow the scalabilities mentioned above. However, many applications could find useful to store additional information to enhance the interpretation and classification of the compressed data. In order to address this issue, JPEG2000 framework considers the definition of metadata boxes to store this relevant information in the same file. The JP2 file format provides two mechanisms for embedding metadata into a file:

- **XML (Extensible Markup Language)** boxes allow XML documents to be embedded within the JP2 file. The meaning of the data is specified through the XML Document Type Definition (DTD).

- **UUID (Universal Unique Identifier)** boxes allow binary data to be embedded within the file. The UUID number is generated by the application developer when the format of the data to be contained in the box is determined.

Both methods allow to add metadata to image files. Additionally, XML and UUID boxes may be placed almost anywhere in the file, where it is more appropriate for the target application.

Part 2: Extensions to Part 1

The JP2 file format is devised for applications that are limited to the storage of RGB and greyscale images. The JPX file format defined in Part 2 of the standard expands its capabilities to allow other types of images and transformations devised to improve the compression efficiency, and defines a set of standard metadata fields. The JPX file format also allows the specification of multiple images within a single file, and the combination of those images through composition and animation. While this is similar to the capabilities of the motion JPEG2000 format (JPEG2000 Part 3), it is aimed at simpler applications that do not include synchronized sound or real-time frame rates.

The most important features that the Part 2 of the standard provides are: the possibility to store several compressed images in the same file, a defined metadata structure in XML to enable an Intellectual Property box, and the inclusion of metadata to provide semantic capabilities to each image. Figure 4 depicts a scheme of this file format.

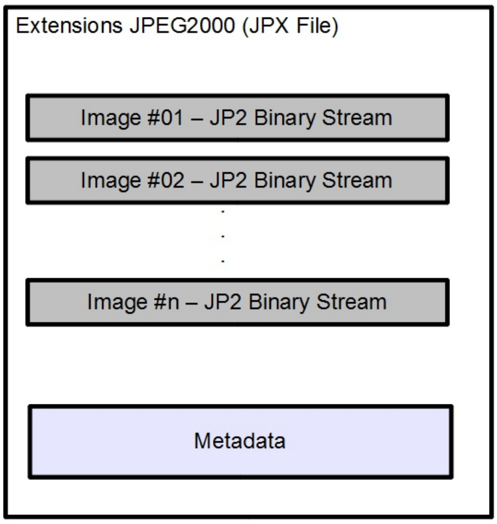

Extensions JPEG2000 (JPX File)
Image #01 – JP2 Binary Stream
Image #02 – JP2 Binary Stream
Image #n – JP2 Binary Stream
Metadata

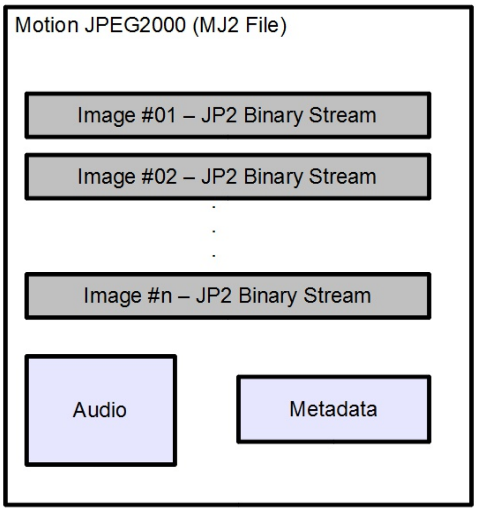
Figure 4: JPEG2000 Part 2 of the standard stores several compressed images and their corresponding metadata information.

Part 3: Motion JPEG2000

Motion JPEG2000 is expected to be used in a variety of applications, particularly where the coder and decoder are already available for other reasons. The application areas include, but are not limited to, digital still cameras, PC-based video capturing, error-prone environments such as wireless and the internet, and high quality digital video recording for professional broadcasting and motion picture production. Motion JPEG2000 is a flexible format, permitting a wide variety of usages such as editing, display, interchange and streaming from streaming servers using a variety of protocols.

MJ2 file format is composed of media-data (eg. video frames and sound samples) and metadata. The metadata can be subdivided into “structural” metadata that is required to allow the file decompression and “descriptive” metadata. While the structural metadata is described in detail in the standard, the descriptive metadata is an open box to store information from different scenarios where the MJ2 file format is used.

A simple MJ2 presentation may be composed of a single video and a single mono or stereo audio track. However, a more complex one might have multiple video tracks overlapping in time, such as in layers. Stored in the disk, the presentation may be in two forms: self-contained in a single file or in multiple files. In the last case, a parent MJ2 presentation file conforms the MJ2 format and contains all the conforming metadata. Figure 5 depicts a schematic description of the MJ2 file format.



Motion JPEG2000 (MJ2 File)
Image #01 – JP2 Binary Stream
Image #02 – JP2 Binary Stream
Image #n – JP2 Binary Stream
Audio
Metadata

Figure 5 JPEG2000 standard file formats: Part 3 of the standard enables the management of motion sequences and the metadata.

Part 8: Security tools

The Internet and the new information technology radically simplify the access of content for the user. Therefore, it is expected that tools and protective methods that are recommended in JPEG2000 must ensure the security of transaction, protection of content, and protection of technologies. Security issues, such as authentication, data integrity, privacy, conditional access, confidentiality, transaction tracing, to mention a few, are among important features in many imaging applications targeted by JPEG2000.

The protection of digital contents is described and can be achieved in many ways such as digital watermarking, digital signature, encryption, metadata, authentication and integrity checking. Part 8 of JPEG2000 standard intends to provide tools and solutions in terms of specifications that allow applications to generate, consume and exchange Secure JPEG2000 binary streams. This is referred as JPSEC. Some interesting features for the legal multimedia management are:

- Access control: The prevention of unauthorized use of a resource, including the prevention of use of a resource in an unauthorized manner.

- Authentication: The process of verifying an identity claimed by or for a system entity.

- Source authentication: Source authentication is the verification that a source entity is in fact the claimed source.

- Confidentiality: Confidentiality is the property that information is not made available or disclosed to unauthorized individuals, entities or processes.

- Encryption: Transformation of data by a cryptographic algorithm to hide the information content of the data.

- Fingerprints: Fingerprints are characteristics of an object that tend to distinguish it from other similar objects. They enable the owner to trace authorized users distributing them illegally.

- Data integrity: Image data integrity denotes the property that data has not been altered or destroyed in an unauthorized manner.

- Watermarking: Watermarking inserts imperceptibly data representing some information into multimedia data.

Part 9: Interactive transmission protocol

Part 9 of JPEG2000 defines tools for supporting image and metadata delivery. The main component of Part 9 is a client-server protocol called JPIP. JPIP may be implemented on top of HTTP, but is designed with a view to other possible transports. To facilitate its deployment in systems with varying degrees of complexity, JPIP handles several different formats for the image data returned by the server: these include ordinary image formats such as complete JPEG or JPEG2000 files. JPIP also supports both stateless and stateful modes of operation, enabling cache-modelling to eliminate the redundant transmission of data.

JPIP provides selective access to the image metadata that may be contained within JPEG2000 files. Although Part 9 is focused on the application of technology from Part 1, including the JP2 file format, it supports some file format extensions from Part 2. A mechanism has also been provided for selection from amongst multiple binary streams in JPX (Part 2), MJ2 (Part 3) and JPM[29] (Part 6) files. Potentially this could be applied to any file format containing images, not just to the JPEG2000 framework. In our scenario, focused on legal multimedia management, the JPIP protocol is useful to transmit these files and their corresponding metadata information over the network.

5.3.4 Image format recommendation for CAPER system

Finally, we point out several conclusions from overviews focused on different image file formats, standards and recommendations. The first point is that the CAPER system must to store the acquired images in a raster format, since vector file representations are usually text. Consequently, the only compression strategy allowed by this kind of representation is a lossless approach. Therefore, taking into account the huge amount of images that CAPER system is foresee to handle, the limitation to one strategy is unwise. In addition, vector file formats are usually devised on a specific field. However, we are looking for a general-purpose file format to facilitate the acquisition of information from different sources within Internet.

The second point is focused on which JPEG committee standard is suitable to meet the most CAPER system requirements. Table 4 sums up the most desirable features (both lossless and lossy strategies, inclusion of metadata information, efficient management of compressed files and allow of different scalabilities) from the five standards presented in previous sections.

[29] See Annex I for further information about Compound documents in JPEG2000 (Part 6)

	Lossless/lossy	Metadata	Efficient Management	Scalabilities
JPEG-LS	No	No	No	No
JPEG-XR	--	--	No	No
JBIG2	Yes	Yes	No	No
JPEG	Yes	No	No	No
JPEG2000	Yes	Yes	Yes	Yes

Table 4: Main features from the standards proposed in Section 3.

The third point is that we suggest the adoption of JPEG2000 standard for the CAPER system due to it is the most suitable algorithm to meet the posed requirements. Besides the characteristics showed in Table 4, JPEG2000 provides more features, detailed in Section 3.1.2.2, such as access control security tools, interactive transmission protocols and so on. In addition three different implementations introduced in the literature are also presented here: Kakadu Software (The University of New South Wales, 2012), Jasper Software (Adams, 2012) and (JJ2000, 2012). These implementations are developed in different programming languages and offer several particular features.

5.3.5 Video

Digital videos have been used widely in several fields. The content consumption and easier access to information continues to increase rapidly. The result is that people are increasingly creators as well as consumers of digital videos for both in their professional and personal lives. Moreover, over the past years traditional media sources and regular users share a set of concerns: management of content, device capabilities, protection of unauthorized access, modification and so on. Consequently, international standard organizations have been working in different recommendations specifically devised to meet requirements posed by the community composed of professionals and non-professionals users.

The CAPER system has an architecture that is able to acquire and analyse different kind of videos within Internet. Therefore, the recommendation performed to store and manage digital videos must meet five main goals as in the digital images field: i) to facilitate access to the information acquired and processed by the CAPER system through the adoption of international standards and the recommendations; ii) to provide a certain degree of compression to mitigate storage requirement that arises within the CAPER system; iii) to provide both lossless and lossy strategies; iv) to perform an efficient management of compressed files; and v) to allow the inclusion of metadata to store the result of the analysis performed by the CAPER system in conjunction with compressed images. Furthermore, some other desirable extra features that we have to take into account are: interactive and progressive transmission and security and other related issues. In contrast to digital images field, digital videos are able to be drastically compressed due to the intra-frame compression. So, in this situation the lossy strategy achieves better results than the lossless strategy. In consequence the lossy codecs are explained with more detail in this deliverable.

This Section is organized as follows: Section 3.2.1 briefs the most common video file formats and codecs available within literature and industry; Section 3.2.2 sums up standards used in the Digital Video Broadcasting (DVB) field; Section 3.2.3 reviews the standards and recommendations promoted by the Movie Pictures Experts Group (MPEG); and finally, Section 3.2.4 points out a recommendation for the CAPER system.

5.3.6 Video file formats and codecs

Historically, there is a complex balance between the video quality, the quantity of the data needed to represent it, the complexity of the encoding and decoding algorithms, robustness to data losses and errors, random access, state-of-the-art compression and so on. Therefore, professionals have tried different solutions to tackling these issues.

International organizations and academic institutions worldwide have been working in different recommendations and standards specifically devised for digital video management. The MPEG committee has made an important effort to promote general-purpose standards and recommendations which are able to offer some desirable features such as efficient management, compression and so on. However, in digital video field companies and industries usually develop their own algorithm to manage and compress digital video which are incorporated in different frameworks.

In digital video compression field, industries usually adopt a standard or recommendation developed by international standard organizations to meet their own requirements and specifically its customer's proposals. Consequently, several different approaches from a standard could be found in the market. Nevertheless, codecs belonging to different proprietors or industries may not the compatible among them. As an example, in this list we introduce some different approaches from a same standard:

<u>MPEG-4 Part 2 codecs</u>

- DivX Pro Codec: A proprietary MPEG-4 ASP codec made by DivX Inc.
- Xvid: Free/open-source implementation of MPEG-4 ASP, originally based on the OpenDivX project.
- FFmpeg MPEG-4: Included in the open-source libavcodec library, which is used in many open-source video players, frameworks, editors and encoding tools. It is complatible with other standard MPEG-4 codecs like Xvid or DivX Pro Codec.

<u>MPEG-4 H.264 AVC codecs</u>

- x264: A GPL-licensed implementation of the H.264 video standard. X264 is only an encoder.
- QuickTime H.264: H.264 implementation released by Apple.
- DivX Pro Codec: An H.264 decoder and encoder were added in version 7.

<u>Microsoft codecs</u>

- WMV (Windows Media Video): Microsoft's video codec designs including WMV 7, WMV 8 and WMV 9.
- MS MPEG-4v3: A proprietary and not MPEG-4 compliant video codec created by Microsoft. It is released as a part of Windows Media Tools 4.

<u>On2 codecs</u>

- VP series (VP6, VP6-E, VP6-S, VP-7, VP8): Proprietary high definition video compression formats and codecs developed by On2 Technologies used in platforms such as Adobe Flash Player 8 and above, Java FX and other mobile and desktop video platforms.
- Libtheora: A reference implementation of the Theora video compression format developed by the Xiph.org Foundation, based upon On2 Technologies' VP3 codec.

<u>RealNetworks</u>

- RealVideo: Developed by RealNetworks. It is a popular compression format and codec technology a few years ago.

Although industries have been developed their own codecs based on standards and recommendations promoted by international organizations, in this deliverable we are focused on original standards.

Moreover, in the literature and industry there are lossless and lossy codecs. As an example:

Lossless compression

- H264 lossless
- JPEG2000 lossless
- LOCO

Lossy compression

- MPEG-1 Part 2
- MPEG-2 Part 2
- MPEG-4 Part 2
- MPEG-4 Part 10

5.3.7 Digital Video Broadcasting (DVB)

In this section we offer an overview of Digital Video Broadcasting (DVB) project. The interested reader is able to find more details in the DVB project and the standardization within the DVB field in (Hannuksela, Döhla, & Murray, 2012). The DVB project develops standards for television broadcast services. These standards could be interesting for the CAPER system since one of the input sources are Mass Media services. Therefore, the standards performed in this input sources are able to determine the way to acquire contents through the Internet.

The DVB file format is intended to be a recording format for received broadcasts and a file interchange format when recorded broadcast are copied or moved from one DVB-compatible device to another. The DVB file format is an extension of the ISO Base Media File Format (ISOBMFF), which is widely used as a basis format for various container formats such as MPEG-4 and 3GP file formats developed by the MPEG committee and Third Generation Partnership Project (3GPP), respectively.

The use cases can be grouped in the following:

- Efficient recording of broadcast transmissions
- Efficient playback of recorded files
- Ease of movement
- Inclusion of descriptive and segmentation information
- Protection of content

The DVB File Format specification was released by the DVB project office in June 2008 and published by the European Telecommunications Standards Institute in November 2008[30]. Then, in early 2011, a guidelines document was published that provided informative

[30] ETSI Standards and Blue Books are available on-line at: http://www.dvb.org/technology/standards/index.xml

descriptions of the use of the reception hint tracks, branding, fragments, and timing. The document describes expected and preferred modes of operation and describes and explains the preferred usage of the functionality included in the DVB format.

5.3.8 MPEG committee

The Moving Picture Experts Group (MPEG) is a working group of experts that was formed by ISO and IEC standard organizations to set standards and recommendations for video compression, audio and transmission. It was established in 1988 by the initiative of Hiroshi Yasuda (Nippon Telegraph and Telephone) and Leonardo Chiariglione, group Chair since its inception. As of late 2005, MPEG has grown to include approximately 350 members per meeting from various industries, universities, and research institutions. MPEG's official designation is ISO/IEC JTC1/SC29 WG11 - *Coding of moving pictures and audio* (ISO/IEC Joint Technical Committee 1, Subcommittee 29, Working Group 11).

Since every MPEG standard has many different parts, in this deliverable we are focused on the ones aimed at meeting the CAPER system requirements.

Algorithm	Details	Link
MPEG-2 Part 2	Compression codec for interlaced and non-interlaced video signals.	ISO/IEC 13818-2:2000/Cor 1:2002
MPEG-4 Part 2	It is a Discrete Cosine Transform (DCT) compression standard, similar to previous standards such as MPEG-1 and MPEG-2. Several popular codecs including DivX, Xvid and Nero Digital implement this standard.	ISO/IEC 14496-2:2004
MPEG-4 Part 10	MPEG-4 Part 10 Advanced Video Coding (AVC), also known as H.264, is a standard for video compression and is currently one of the most commonly used formats for the recording, compression, and distribution of high definition video.	ISO/IEC 14496-10:2010
MPEG-7	MPEG-7 is a multimedia content description standard. MPEG-7 is formally called Multimedia Content Description Interface.	ISO/IEC 15938-1:2002/Cor 1:2004
MPEG-21	The MPEG-21 standard aims at defining an open framework for multimedia applications.	ISO/IEC 21000-19:2010

Table 5: Recommendations and standards for video developed by the MPEG committee focused on video compression and management.

The main goal of this section is to offer a brief description of the standards promoted by the MPEG committee which are able to fulfil requirements posed by the CAPER system. Next sections sum up some standards presented in Table 5.

5.3.8.1 MPEG-7

Multimedia data has been an important increase of available data in both the scientific and consumer domains and facilities to access the multimedia data. Nevertheless, the extraction of useful information from this data and the application in practical systems such as multimedia search engines are still open problems.

MPEG-7 (Chang, Sikora, & Puri, 2001) aims at standardizing tools for describing multimedia data. The MPEG-7 framework is composed of: Descriptors (Ds); Description Schemes (DSs), a Description Definition Language and a coding scheme (Sikora, 2001). (Troncy, Bailer, Hausenblas, & Hofmair, 2006)

5.3.8.2 MPEG-21

Together, the MPEG-1, -2, -4 and -7 standards provide a complete, powerful and successful set of tools for multimedia representation. However, widespread deployment of multimedia applications requires more than this collection of standards. The problem is that overall multimedia consumption and commerce remain non-transparent and are not happening on a large scale. In the desire to achieve interoperability, the framework may violate the requirement to protect the value of the content and the rights of the rights holders. Digital rights management (DRM) systems can prevent interoperability if they use non-standardized protection mechanism. The MPEG-21 (Burnett, Van de Walle, Hill, Bormans, & Pereira, 2003) aims to guarantee interoperability by focusing on how the elements of a multimedia application infrastructure should relate, integrate and interact, where open standards for elements are missing.

Interoperability is the main goal behind all multimedia standards. It is a necessary requirement for any application that requires guaranteed communication between two or more parties. To achieve this goal we must standardize both the content structure and a minimum set of communication processes. Therefore, effective standardization is to create a minimum (but complete) standard that normatively defines a minimal set of tools aimed at guaranteeing interoperability.

Digital Items is a basic concept that is a combination of resources, metadata and structure. The resources are the individual assets or content. The metadata describes data about or pertaining to the Digital Item as a whole or also to the individual resources in the Digital Item. Finally, the structure relates to the relationships among the parts of the Digital Item. Therefore, the Digital Item is the fundamental unit for distribution and transaction within the MPEG-21 framework.

The MPEG-21 standard is composed of 18 different parts. However, since a complete description of the MPEG-21 specifications is not the main goal of this deliverable, we are only focused on five parts that are able to meet the CAPER system requirements.

MPEG-21 Part	Description
Part 1	Vision, technologies, and strategy: describes the multimedia framework and its architectural elements.
Part 2	Digital Item Declaration (DID): provides a uniform and flexible abstraction and interoperable schema for declaring Digital Items.
Part 4	Intellectual property management and protection (IPMP): provides the management and protection of content across networks and devices.
Part 8	Reference software: includes software that implements the tools specified in the other MPEG-21 parts.
Part 9	File format: determines a file format for storing and distributing Digital Items.

Table 6: Description of the five selected parts of the MPEG-21 standard.

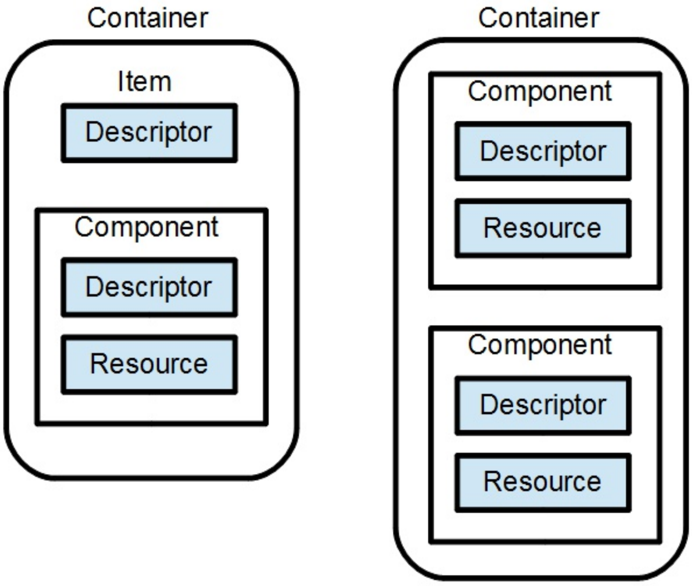


Figure 6: Some Digital Item declaration model elements and their relationships

Figure 6 shows a declaration of model elements and their relationships within the MPEG-21 standard structure.

5.3.9 Video format recommendation for the CAPER system

The main point is that we suggest the adoption of MPEG standard for the CAPER system due to it is the most suitable algorithm to meet the posed requirements. MPEG-7 and MPEG-21 provide some suitable performances as inclusion of metadata and interoperability, respectively. Consequently, both standards should provide useful features for the CAPER system

5.3.10 JPEG2000 – Part 6 Compound Documents

This format, named as JPM, is an extension of the JP2 file format and uses boxes defined for both the JP2 and the JPX file formats. This part of the standard is useful for applications storing multiple pages, images with mixed content and/or images that need more structure than the provided in JP2.

This International Standard is based on the multi-layer Mixed Raster Content (MRC) (ISO/IEC15444-6, 2003) imaging model. It makes provisions for processing, interchange and archiving of these image types in multiple layers, and defines composition models which regenerate the desired image. The efficiency is realized in the segmentation in layers of different types of images, since it allows an image specific compression.

Figure 3 shows an example of how the JPEG2000 Part 6 handles the compound documents. In this figure, the original document is composed by a colour and greyscale images and text. First, a segmentation process is applied to divide the document in three different layers: color image, grayscale image and text.

The JPM file format allows to encode each layer with a different compression algorithm, so if we consider the text as a bi-level image (black and white), we can apply the JBIG2 [24, 25] standard that is specifically devised to encode bi-level images and documents. Metadata boxes are also considered in this file format to store some additional information about documents and to provide semantic capabilities.

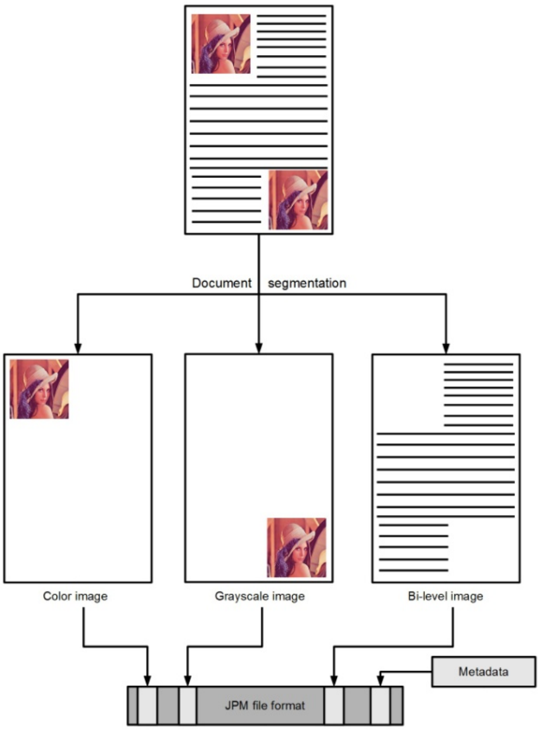


Figure 7 JPEG2000 Part 6 (JPM file format). The original document is composed by color and grayscale images and text. Then, the document is segmented in three different layers and a specific encoding is applied. The compressed binary streams from the three layers and the needed metadata are store in an unique file.