MIRACLE

Machine-readable and interoperable
age classification labels in Europe

Grant agreement no: 621059

Electronic age labels ecosystem

Report on practices of data transfers and web labelling in the field of age classification data

October 24th 2015

Deliverable D3.1

Editors: HBI, JUSPROG, FSM

Contributors: BBFC, NICAM, PEGI, ALLOT OPTENET

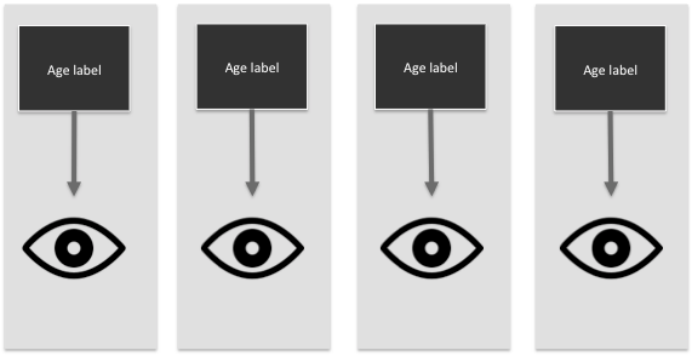
MIRACLE is co-funded by the "ICT Policy Support Programme" of the European Union.

Contents

1. Background: The MIRACLE project

Information about media content is deemed an important aspect for ensuring rational consumer decisions in the digital market. In the field of suitable and unsuitable content for children, many EU member states rely on age classification procedures and visual age labels. By informing children and parents about relevant media content, such age labels empower them to exercise informed choices regarding media usage, literacy and education.

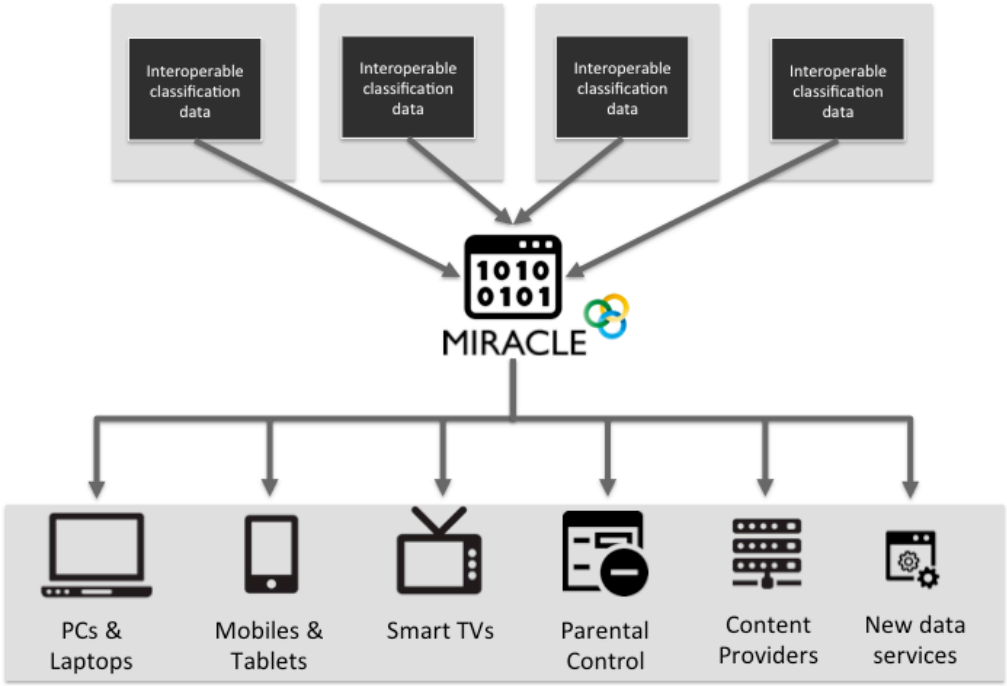
Diag. 1: Status quo - National, media- and scheme-related age classification silos



However, age ratings and classification schemes are highly fragmented among the Member States and – in many cases – rely on visual labels only. In digital contexts, though, *electronic* age labels open the chance for exchanging and processing classification information in digital realms. Given that one common data model is being used across borders and platforms, such an approach could fully utilise the knowledge comprised in these labels – for the benefit of both end users and businesses along the supply chain and across borders, regions and devices.

Against this background the MIRACLE project ("Machine-readable and interoperable age classification labels in Europe") aims at providing one data model for electronic age classification information, implementing it in different classification schemes and showing the added value of interoperable classification data for businesses, educational institutions and end users. The eight project partners spread across five different member states and classification schemes, consisting of classification bodies, Safer Internet nodes, self-regulatory bodies and filter software providers. The 30-month technical pilot started in spring 2014 and is co-funded by the "ICT Policy Support Programme" within the CIP ("Competitiveness and Innovation Framework Programme") of the European Union.

Diag. 2: Advantages of interoperable age classification data



After an initial draft and a public consultation phase, the project members published Version 1.0 of the MIRACLE specification in autumn 2014. With Version 1.0 the data model took into account current classification schemes and practices while considering existing electronic labelling schemes in order to be easily adapted. Its unrestrictive and open approach leaves enough flexibility for existing and future schemes to map their classification data to MIRACLE. More so, it even enables them to extend the specification to deliver more metadata, if desired. With the actual age label still at its core, the data model also provides fields for content descriptors and – for interactive media content – feature descriptors, e.g. location-based services.

By February 2015, four consortium partners had implemented the specification in their very own classification scheme context, offering API endpoints that provide MIRACLE datasets. For the first time, interoperable age labels are available across borders: UK-based BBFC, Netherlands-based NICAM as well as the Pan-European Game Information System PEGI have mapped their existing classification data to the MIRACLE data model and offer MIRACLE-compatible access to (parts of) their classification databases. The German FSM provides an online mapping service that "translates" existing age-de.xml labels into MIRACLE data sets on the fly. Moreover, NCBI as the Safer Internet node in the Czech Republic started to pilot a MIRACLE-based database with an API endpoint from summer 2015.

During the implementation phase, valuable experiences and insights have been made by the consortium regarding the overall strategy of implementing an interoperable data model, relevant context factors for opening up classification data to cross-border provision as well as regarding technical aspects of mapping a new specification to existing databases and schemes. Both the decision-making regarding the implementation strategies and the actual implementation steps have been documented by each of the five consortium partners. These reports can also be of great value for

third parties when it comes to implementing MIRACLE. Based on the experiences made during the implementation as well as on remarks from industry stakeholders regarding additional requirements in electronic age labels the consortium has published an optimized version 2.0 of the MIRACLE specification in September 2015.

Aiming at becoming a de facto standard for electronic age classification information, MIRACLE put specific efforts in disseminating the pilot project's objectives, the MIRACLE specification as well as general information about the added value of interoperable data in the field of digital child safety. Being a pilot project that is open to and interested in third-party uptake, all relevant stakeholders are invited to provide or use MIRACLE-compatible data. The project members are in constant talks with classification bodies, content producers, distributors, platforms and online services on potential surpluses of using MIRACLE data.

2. Relevance of age classifications in child protection policies

Age ratings are one of the most traditional regulatory instruments when it comes to child protection regulations in the field of media[1]; age classification and age labelling are cornerstones within EC child protection policies accordingly[2]. Comparative analyses show that almost all legal or self-regulatory frameworks in child protection have put up a scheme for media-related age classifications and install different consequences based on the respective age rating, e.g. displaying age labels or limiting the access for minors.[3]

Fields of application of age classification and age labels have been motion pictures being shown in public cinemas or on TV and – later on – movies, videos and games being sold in retail stores on video cassettes, DVDs or Blu-ray discs. While age labels ultimately aim at protecting minors and informing parents (and kids), legal frameworks regularly impose obligations to acquire an age classification on content producers or distributors. Where no legal frameworks are in place, e.g. in the field of video games, self-regulatory initiatives have often been established where industry players have pledged to provide age classifications. These circumstances make age classification and age labelling an important aspect of corporate compliance.

From the perspective of child protection policies, age labels are a two-fold instrument within child protection frameworks: On the one side, they are used to inform retail and cashier staff whether a specific product may be handed out to or accessed by minors below the age rating. Regularly, violations against such legal provisions are sanctioned. On the other side, age labels are deemed to enabling parents to take better-informed decisions when it comes to their assessment whether a specific piece of media content is suitable for their child, thus empowering parents to educate their children on a better-informed knowledge base.

---

[1] E. Wauters, E. Lievens, P. Valcke, The use of labels to empower minors, parents and educators in the social media environment, An explanatory report, Leuven, February 2013, p. 8.

[2] F. J. C. Blázquez, M. Cappello, A. Lépinard, S. Valais, The protection of minors in a converged media environment, European Audiovisual Observatory IRIS plus 2015-1, Strasbourg 2015, pp. 18.

[3] W. Schulz, S. Dreyer, K. Dankert, M. Puppis, M. Künzler, Chr. Wassmer, Identifikation von Good Practice im Jugendmedienschutz im internationalen Vergleich [Identification of Good Practice in Youth Media Protection in International Comparison], Bericht im Rahmen des nationalen Programms Jugend und Medien, Berichtnummer 12/15, Bern, pp. 33.

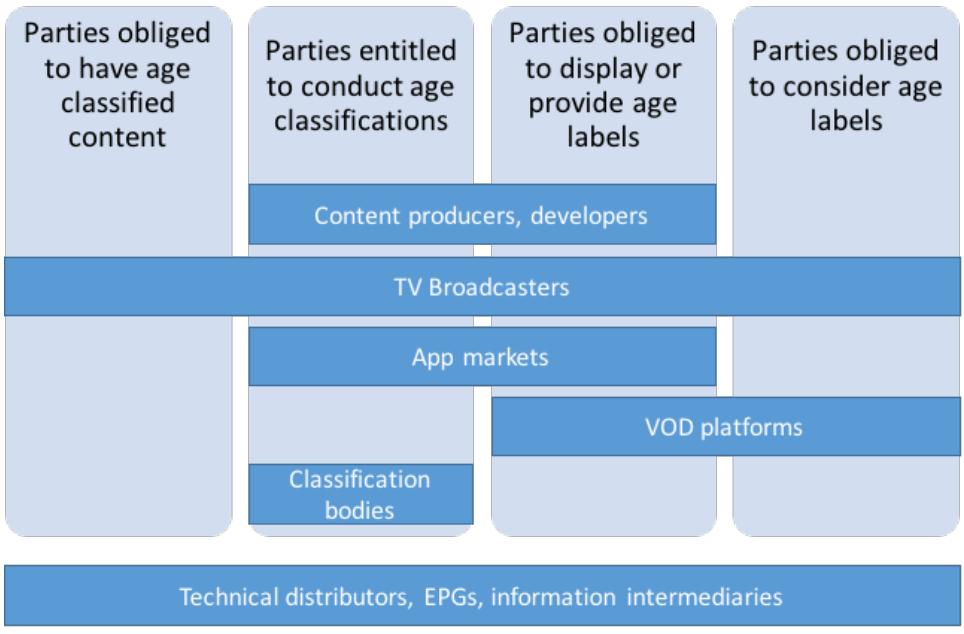
The main actions within child protection policies are (a) the decision on the actual age classification, (b) the labelling of the content with the respective age label and (c) the consideration of the age rating in view of possible protective measures, including the deployment of such measures (e.g. age verification, watersheds, PIN codes etc.).

Diag. 3: Actions in the area of age classification

The classification decisions are being made either by the content producers or distributors themselves or – depending on the national legal framework – by official classification bodies that have been established by law, act under state acceptance or are part of a self- or co-regulation scheme. This does not only show the importance of age classifications and labels in child protection policies, it also explains the main stakeholders in classification-related governance structures.

In many cases, market players fall into more than one category, depending on their activities and business models. For instance, content producers like game developers or production firms will not need to obtain an age classification for their products as long as they do not publish it. In such cases, the age classification usually will be obtained or performed by the publisher, e.g. a movie distributor, game publisher or broadcaster. It is not unusual, though, that content producers publish their products themselves, especially in app market environments. In such cases, the content producer will have to perform the age classification herself or have to obtain an age classification from a classification body and provide the respective age label with the content.

Diag. 4: Classification and labelling activities of relevant stakeholders



The overlapping areas of classification and labelling activities among the different stakeholder groups as well as the need to transfer the classification information from one party to another make age classification data an important aspect of media content-related metadata for all stakeholders within the system of child protection – way beyond the plain age label.

3. Electronic age classification data provision, transfer and exchange

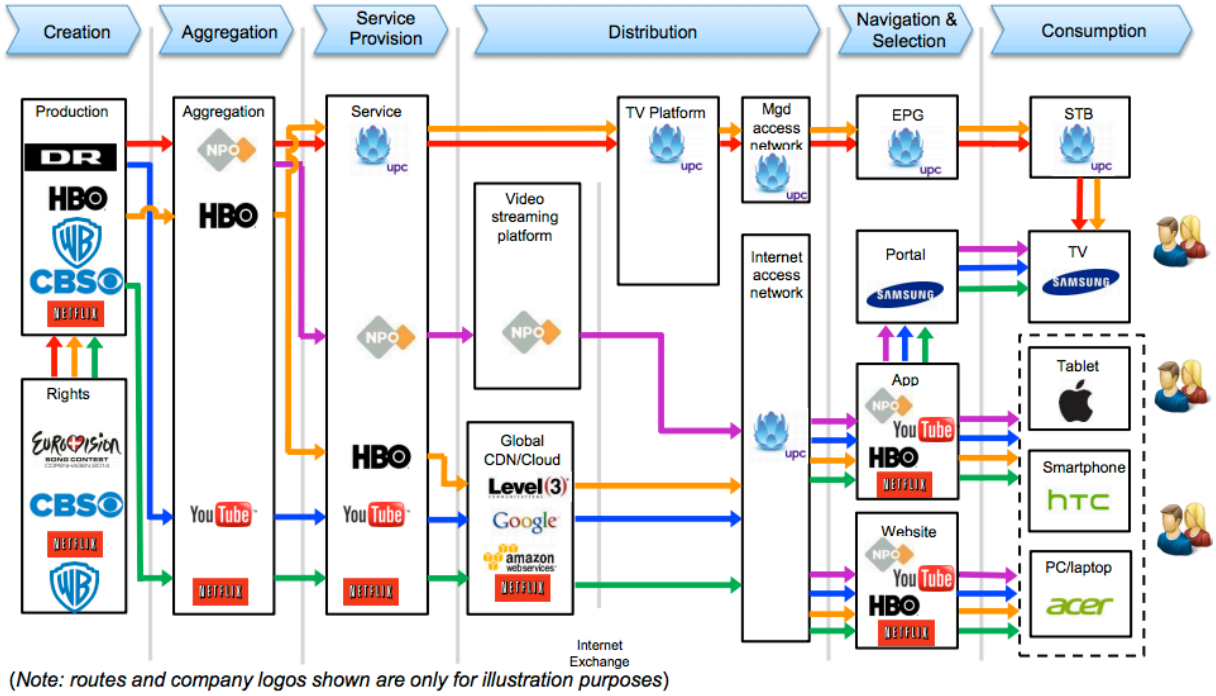
Regularly publishers and providers of age rated products are obliged to inform the consumer about the age classification – they have to label the content. In an analogue world, this usually is done by the main distributors and publishers who provide the retail product for the (national) market, e.g. film distributors or game publishers. Regarding TV content, the respective broadcaster is – due to Art. 27 sec. 3 AVMSD – obliged to inform about an age classification of a broadcast at the beginning of the programme.

In many countries, age labels are shown for retail products in online shops, too, but also for digitally distributed content that has been age rated as a retail product, e.g. movies or games. In such cases the traditional retail box label will be made available by the shop or content provider either as text information or as an image within the shop or download environment. This way, the analogue visual labels become digitized. To provide digitized age labels to their customers, content providers need information about age classifications. In case of online shops for retail products, the shop owner usually has access to the retail product and hence can provide information on grounds of the age label on the product. However, he might not be entitled to use the official age label visuals, since he needs a licence to depict the age classifier's logo (see below for IP issues in age labels). In all other situations, the content provider needs to know the age rating of the specific piece of content. Electronic data can provide this information. The following chapter gives an overview on regular data sources and data flows between the different groups of stakeholders in digital media environments.

3.1 Increase in routes and players: Age labels and the "value web"

With the current shift away from retail ("offline") distribution as well as linear one-to-many broadcasts of linear content to online distribution and usage of non-linear content, e.g. web TV, IPTV, video on demand, video platforms, digital sales markets, app stores etc.), the so-called "value web" offers users diverse new choices in routes for watching audiovisual content compared to the access options to traditional retail products and broadcasting content. Traditional value chains are being extended to a web of possible routes from production via distribution to consumption.

Diag. 5: Content routes in the value web (example)



Source: P. Nooren, Empowering users in the converged media-internet-telecom value web, OBS-EPRA workshop on Empowering Users, Strasbourg, 15 December 2014, available at http://publi.obs.coe.int/documents/205595/8166225/2_TNO+(NOOREN)_OBS_EPRA_ws_empowering_users_Overview_of_instruments_final.pdf/a394d91c-eaf5-40d9-8b74-0d9fe7f68fad

The diversity in content delivery routes plus the emergence of many new players in these markets lead to the fact that digital content is being handled by a lot of different players on its way from production or development to the end user. With the specific perspective on age classifications and age labels this report gives an overview of the relevant players and their role as either sources and suppliers of age classification information, or as demanders and users of such information. Where possible, specific factors for interests, incentives and strategies in providing or acquiring content classification information are described.

3.2 Self-classification: Publishers, distributors and broadcasters

The general principle of age classification in the area of audiovisual content, especially in TV environments, is that the publisher of content has to assess whether it is suitable for all age groups or whether it might impair or even harm minors. This form of age classification has to be performed ex ante, i.e. prior to publication. Here, many national frameworks have set up specific watersheds for specific age limits. This results in duties of all content providers within the scope of law or self-regulatory provisions to age classify their content according to the applicable scheme, e.g. deeming content suitable for minors under 6, 12, 16 or 18 years. In many national schemes other age brackets

apply; except for the age group of 18+ due to the definition of "minors".[4] The classification procedure is usually being performed in-house by trained and experienced content editors or youth protection officers.

Market players that assess all their content prior to publication this way accumulate classification knowledge and a significant amount of age rating information over the years. They become relevant providers of classifications either for those who (sub)license the content, e.g. other TV broadcasters, VOD and OTT providers or service providers who want to provide content-related information to third parties, e.g. premium EPGs or online shops. While (sub-)licensees usually have to pay extra for specific content-related metadata, content publishers usually have an interest to be listed in EPGs or shops. Therefore, the latter usually do not have to pay for access to the basic data, including age classifications. In Digital TV environments, the broadcasters also provide programme data 14 days ahead in time for basic EPGs via the DVB-SI-EIT standard, including "parental ratings".[5] For additional EPG content premium EPGs and other programme information aggregators have to pay (either directly to the broadcaster or, sometimes, to specific collecting societies).

3.3 Central age classification institutions: Classification bodies

Traditionally, the main players in age classification have been the movie classifiers, mostly entitled by state laws to perform the age classification decisions regarding feature films. Their histories reach back decades in many cases; some have been established in the early 20th century and have conducted age classifications for over a hundred years.

Current classification bodies are either state bodies where the main decision-making organ consists of classification experts, independent bodies conducting age classifications or co- or self-regulatory bodies that have been initiated by industry players to have one central body for age classifications available. No matter how different the formal and organisational configurations or remits of these bodies look like, they all fulfil the core activity of deploying an age classification infrastructure within age rating frameworks: Either they make age rating decisions themselves which the publisher, distributor or broadcaster has to adhere to, or they provide a content classification framework on which grounds (trained) age classification coders within the companies classify their content themselves. In case of the latter the classification bodies usually provide training and support for the content publishers, as well as monitoring and complaint services in case of coders' deviations from the assessment criteria.

The classification body usually has notice of all age classification within its framework, making it *the* central node for age rating information. Regularly, these classification bodies have a central database containing all classification decisions. Access options to these databases look different, though: The content publisher usually receives the classification decision on an individual basis, so the main principle is to use the whole database internally only. Depending on the business model, some classifiers provide all age classification data to third parties for money, e.g. for VOD providers, metadata companies or premium EPG services. In some cases, the classification bodies make all their age ratings available publicly for free, usually providing the data in non-standardised format, e.g. in XML, JSON, CSV format with data elements according to their

[4] Art. 1 UN Convention on the Rights of the Child.

[5] ETSI, Digital Video Broadcasting (DVB) - Specification for Service Information (SI) in DVB systems, ETSI EN 300 468, V1.13.1 (2012-04), p. 70.

internal data structure. Some of the classifiers still provide their data manually and on request only.

In some countries, data exchange between classification bodies becomes necessary, especially where more than one classification body has been established. If there is, for instance, one classification body for feature films and one for TV or VOD content, sometimes they will have to decide on the same content, e.g. if a movie age-rated for cinemas is planned to being broadcasted on TV or published on a VOD platform. In such environments, the classifiers sometimes have agreed on a common procedure for data transfers between the two of them.

3.4 Classifications in walled garden environments: Platforms and their age rating knowledge

Walled gardens (also called closed platforms or closed ecosystem) are software-based environments that are completely controlled by their owners. The provider approves or disapproves applications and content within the whole platform and is able to add internal metadata to the content accordingly.

3.4.1 Walled gardens and unrated content

Since publication environments have multiplied and the Internet offers manifold ways to provide exclusive content to customers, many new players have emerged which offer content previously unrated. Either because they are legally obliged to, promised to do so, or because offering unsuitable content to their customers is outside their mission statement (and terms of service, accordingly), many platforms make use of (internal) classification schemes. Platforms making use of such classification measures are, amongst others, (user generated) video platforms, app markets, social network sites as well as specialised platforms, e.g. social networks for kids only or child-specific search engines and whitelists.

Internal classification procedures usually aim at age-gating unsuitable content, enabling child protection features within the platform environment or getting hints at potentially impairing content uploaded by users or other third parties. Approaches that platform providers take differ depending on the type of content and its origins: Curated walled gardens for professionally produced content, e.g. app markets, often implement internal or external classification procedures that have to be completed by content producers, resulting in specific age ratings. A renowned approach in this area is the implementation of IARC within Firefox Marketplace, Google Play, Nintendo eShop and Windows Store.

In contrast to that, user generated content platforms regularly deploy flagging, reporting or complaint based procedures for the community of users so they can report potentially unsuitable content after the upload. Few platforms make additional use of algorithm-based classification. Most platforms use this knowledge to classify content in two or three categories: appropriate for all age groups, appropriate for some age groups, not appropriate at all. Regional differences (e.g. when it comes to taste or decency) or reasons for the flags are usually lost when providers make use of the reporting tools in these ways.

The processing of user-based content classification is mostly three-fold: Illegal and inappropriate content or uploads infringing with the Terms of Service will usually be deleted. Content that is deemed unsuitable for younger age groups is age-gated, e.g. by making it available to registered users only, by applying a warning sign or extra layer against accidental display, or by delisting it from search results for users who have

activated some form of safe search feature. Content that is not deemed inappropriate or in breach of the platform's content guidelines will stay online as is. Some platforms actively flag content that is directed at children, making such positive content retrievable within specific children sections or apps.

Some UGC platforms offer both unrated user-generated content and professional media content that has been rated in advance by an external classification body. These platform providers have to process both external and community-based classification information.

By using one of these approaches, or even both, the walled garden provider cumulates content classification alongside its normal business operations. The vast amount of content in some of the walled garden environments make the platform providers become new and significant players in content classification. Only by the numbers of classifications, many of them surpass the amount of classifications of all traditional classifiers altogether. However, when it comes to "their" classification data, these new players most often make only internal use of these age ratings: They apply age gating features, consider the classification in sorting and filtering or put warning signs before potentially harmful content. Due to the exclusively internal use, the walled gardens providers regularly have no incentive to share their classification knowledge or make it available to third parties. On the contrary: In highly competitive markets, opening up a platform's classification data vault might even enable competitors to either scrape this knowledge or to draw conclusions regarding user numbers or flagging counts, which they might use for their own competitive advance. Classification and flagging features, after all, are services the provider deploys for a better user experience on her specific platform. In the end, this circumstance leads to the fact that there is a lot of classification knowledge stored by walled garden platforms, but it is usually locked away. Public electronic labelling among these platforms is non-existent.

3.4.2 Walled gardens and rated content

In contrast to that, platform providers that offer content which is already age rated (e.g. VOD platforms) regularly receive age classification information from the national classifiers or licensing parties and use the age classifications to provide age based features within their products, e.g. children or family sections, secured adult categories etc. Since the age ratings are used internally as well as for end customer information purposes only, there is no incentive for such content providers for sharing the classification with third parties.

3.5 Technical intermediaries and information intermediaries: Service and orientation

Technical intermediaries are those players that transport the content without interfering with it, e.g. backbone providers, technical distributors and content delivery networks. As they usually leave the transported content unchanged, they are not core stakeholders of age classification data. If the transferred data contains age classification information, they will provide the transport for it, too. To change the delivered content, e.g. by adding metadata to a file transfer like a movie, the service provider would not only need the technical capability to look into the transported file (a technology called deep packet inspection or DPI, which might be deemed a breach of net neutrality and privacy), he would also need access to the respective age classification data, for instance a classifiers' database – an unlikely practice that is currently not applied.

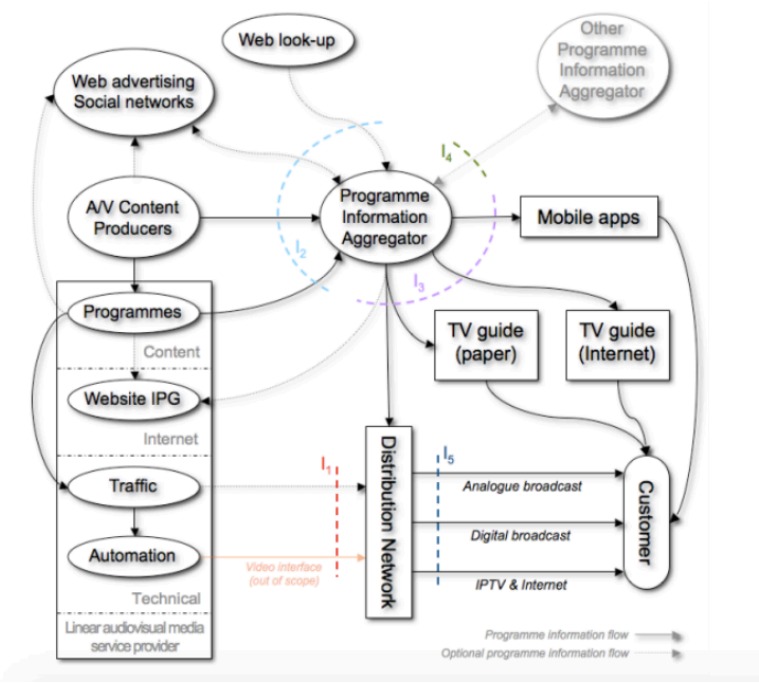
In contrast to technical intermediaries, so-called information intermediaries take an active role in navigating and guiding the user, either by selecting, prioritising or sorting and filtering content or content-related information. Best-known examples for information intermediaries are search engines like Bing and Google, micro-blogging platforms like Twitter and social network sites like Facebook. They all three can give access to content that is already age rated and hence are technologically capable of processing this information, both for labelling and filtering this content within their own services. Google's safe search feature or its capability to display MPAA movie ratings in case you search for a feature film via the "contentRating" information specified on schema.org[6] are two examples where an information intermediary makes use of age rating metadata. In theory, such players could display, apply filters or just provide electronic age classifications where they either have access to the respective content label or to central rating databases. However, all these options come with disadvantages like speed loss, data overhead for all users not requesting such information and the risk of displaying false information, especially in cases where age classification data is being queried by third parties other than the original linked content. A feasible approach can be seen in limiting age classification-related data processing to those forms of content that provide a direct age label. In these cases, the content linked by information intermediaries carries an age rating that can be directly processed and displayed or electronically provided to the end user, enabling him to assess potentially undesired content before accessing it. However, except for specialised search services like kids search engines or searches within a specific content offer, e.g. a movie database, information intermediaries seldom offer age rating-related features.

3.6 Specialised metadata providers and EPGs

Providers of media content-related metadata are usually highly specialised SMBs in the field of TV content. They use several sources and methods to collect metadata, aggregate the information accordingly and provide the aggregated data e.g. to TV programme guides in print or online, for mobile apps or premium electronic EPGs.

[6] http://schema.org/contentRating

Diag. 6: Data flows in EPG ecosystems



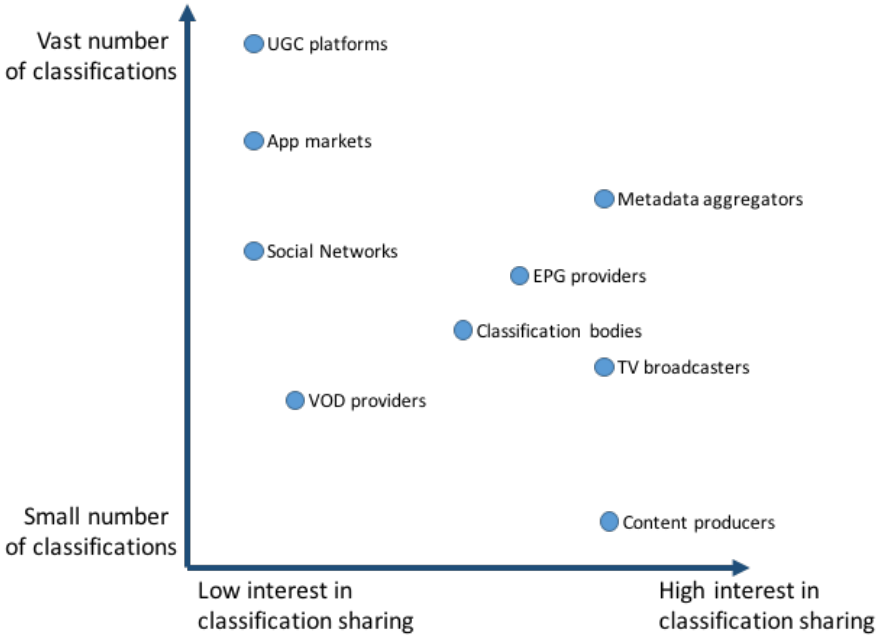
Source: ETSI, Media Content Distribution (MCD); Programme guide information distribution, situation and perspective, ETSI TR 102 988 V1.1.1 (2011-05), p. 11, available at http://www.etsi.org/deliver/etsi_tr/102900_102999/102988/01.01.01_60/tr_102988v010101p.pdf

The aggregated metadata is a valuable resource demanded by TV programme guides, premium EPGs and mobile TV guide apps. Age ratings are always a part of these metadata catalogues. Due to existing watersheds in TV, however, they do not belong to the most important metadata elements. Metadata providers are currently one of the few stakeholders that have started to consider and aggregate age ratings stemming from different schemes as an added value for their services. This is an example for the particular interest of metadata providers to accumulate as much content-related information as possible: The more metadata they have in stock, the more flexible the service might be used by demanders and the bigger the potential group of customers might get.

3.7 Interim conclusion

The overview on current stakeholders in the area of media related electronic age classifications shows two things: Technically, many of the described services and providers either have the knowledge about the age classification of a specific piece of content or could query the respective rating data from a business partner. Yet, the incentives to do so are comparatively low in most cases, due to different but reasonable rationales. The only sectors where age rating provision and transfers are highly endorsed are those where there either is a legal framework in place that obliges affected parties to provide age ratings or where the provision of age-related metadata is part of their core business model.

Diag. 7: Matrix combining amount of classifications with intrinsic incentives to share



4. Age labelling in online environments

The process of applying age classifications to specific media content is called labelling. In contrast to traditional visual age labels, electronic labels are needed in digital environments to enable machines and software to process the information automatically.

4.1 Providing age labels in walled garden environments

As shown above, electronic age labels in walled garden environments are usually restricted to internal use of the respective provider. Since he has no need or incentive to share this data with parties or machines outside the closed ecosystem, the age labelling is regularly based on proprietary code. Thus, the classification data is locked within the walled garden, and external services like parental control filters can only work in these environments in a limited way.

4.2 Providing age labels in the open web: Techniques and potentials

As explained above, obtaining or conducting age classifications and sharing the data through central databases, internal age labels or APIs are all reasonable approaches to implement electronic age labels in digital environments. Another important use case for age classification data in online environments are software and services that make immediate use of the classification data by either (re-)visualising the age rating on screen or by applying parental control filters: Depending on the age set within the parental control software, the requested content is being transmitted/displayed on the end user's device or it is blocked.

The main difference between age rated media content that is distributed online, where one label is applied to one piece of content (usually resulting in one URL), and age ratings for websites is that websites might consist of hundreds or thousands of URLs with a vast amount of *different* content. In such cases one label applies to multiple types

of content and URLs. While one URL can be used as a unique identifier, the scope of websites is potentially encompassing many URLs, often provided hierarchically at one second level domain.

Approaches to cope with this challenge can either be to provide one central label that defines age ratings for a specific scope of URLs by using wildcards, i.e. hierarchical "folders" that a label applies to or to provide single online labels for each unique URL.

Electronic labels can be provided in manifold ways, and many different initiatives have shown potential approaches and currently do so. The following section provides an overview of known approaches including practical examples and a short assessment of their individual advantages and disadvantages from the perspective of third party software that needs to process this data easily, esp. filter software. The latter are one of the most relevant applications when it comes to automatically process age-related metadata of online content. What is needed for such such services to work properly are age labels that can be found and processed quickly and efficiently. While central databases carry all classification-relevant metadata, online labels usually will focus on the age data only. There have been and are several different approaches when it comes to technically label online content.

4.2.1 Information in custom http headers

Age label information can be provided in custom http headers. This way, rating information can be parsed before the full transfer of the requested webpage, enabling an early, time and data saving filtering. Also no additional requests are necessary in cases where a response header contains all necessary information. An example of an age labelling approach that has been using this technique is PICS.[7]

However, the approach also comes with disadvantages, e.g. in many cases content providers or webmasters do not have the rights to manipulate their website's http headers. If they can manipulate their server's http header, wrongly implemented custom headers might break a complete website. Moreover, the user side software must be able to parse custom http headers, too. In general, http headers are not specifically intended for content-related metadata like age labels, but for client-server-communication in general.

4.2.2 Information in html <head> - custom meta tag

Another approach is to provide age rating data within each page's header as a custom <meta> tag. For instance, both PICS and ICRA Version 1 offered this type of online labelling. Meta tags can be implemented in rather flexible ways by almost all content providers and webmasters, and are delivered in easily accessible source code. This way, the information can theoretically be used by search engines. Meta tags are also the technical-semantically right place, as age labels are page-specific metadata. However, before the meta tag can be processed, the whole html file has to be transferred, which results in a loss of time, data overhead and, in the end, an approach that is not scalable for a large amount of data. Some content providers or webmasters might also not be able to implement custom meta tags.

[7] https://www.w3.org/PICS/

4.2.3 Microdata in html source

A recent method to provide age-related metadata is to use microdata within the html source. The most renowned example for this are the schema.org vocabularies that enable content providers to enhance their content by semantic web vocabularies. The learning curve to do so is steep, but the implementation can technically be done by most of the content providers and webmasters. This technique also can provide rating information on more than on one media content within a single webpage and results in accessible source code. However, in this case, too, the whole html file has to be transferred, leading to delayed processing and reduced scalability. The main disadvantage from a perspective of efficient parsing is that the metadata is scattered over the webpage, since different metadata is attached to different source elements. There is no predefined place for the age label, and no central metadata file.

4.2.4 Information as comments in html source

There have been proposals to provide age classification data as comments within the html source. While this technique can be implemented by most of the content providers and webmasters and is accessible in the source, again the whole html file has to be transferred before being processed. Technically, comments are also content that should not be processed, but is used for internal documentation purposes. Using comments for actually relevant metadata would neglect this.

4.2.5 HTML5 data-* attributes

With the latest HTML specification (HTML5), html data attributes have been introduced. These attributes theoretically allow for providing age ratings within the html source in a specified manner.

The advantages are the data attributes' easy parsing possibilities, the fact that this is a common web standard and classification information can be viewed in the source code. However, not every webmaster is able to code in html5, plus the current knowledge of how to use html5-data* attributes is yet sparse. And here, too, the whole html page has to be transmitted before parsing. Finally, data-*-attributes are intended for internal features and especially not for interaction with third parties.[8]

4.2.6 Information in (rdf/xml) datasets at domain root

Another take on website labelling is to provide one central file at a defined location of a domain, containing all relevant age rating data for the whole site. Examples for such techniques are ICRA (Version 2, rdf-based) and age-de.xml as an official labelling specification in Germany (xml-based).

By providing information at a determined location age labels are especially easy to find for third parties. As the information is within a file of its own, processing the age classification information takes place independent from or parallel to the html transmission. The information within these files usually stays the same for some time, which makes caching possible. As a central file provides age ratings for more than one URL the approach is highly scalable and reduces data overhead compared to labels within each html file.

[8] http://www.w3.org/html/wg/drafts/html/master/dom.html#embedding-custom-non-visible-data-with-the-data-*-attributes

Disadvantages of this approach are that not all webmasters have access to the domain root. Especially in cases of huge websites with a lot of different content suitable for different age groups central files become bigger, dissolving the advantages in speed and efficiency.

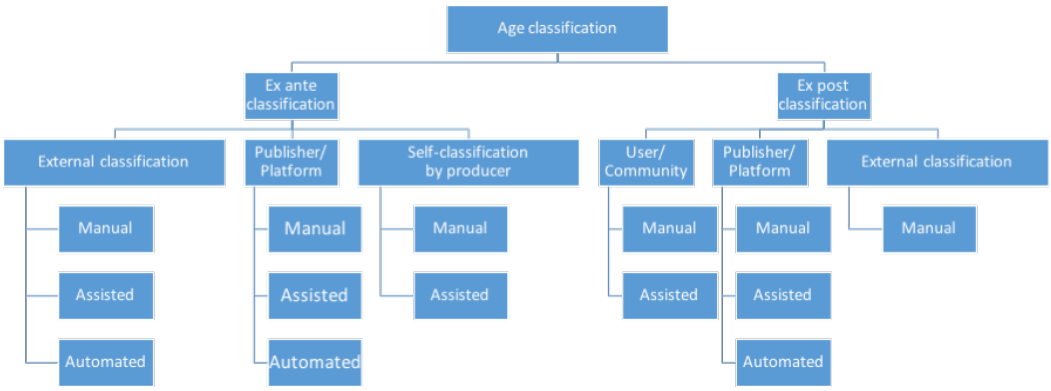
4.3 Differences in filtering between user-side filters and ISP-level filters

Another aspect to consider in web labelling is the way that end user or ISP level filters process the data within electronic labels: Filter software that is running on the user's end device usually digs into the TCP/IP stack of the local computing environment and therefore can analyse the whole web traffic locally. If filtering requires an additional request to the server of the incoming data the local system can easily send this request, enabling the filter service to work in parallel to the actual user request. In contrast to that ISP level filters do not run on the local machine, but within proxies of the transferring ISP. As all requested data is being transmitted through these filter proxies, additional requests, e.g. in case the filter needs to access central XML or RDF files for the requested website, would need to be sent by the proxy. Such requests from within the core of an ISP are uncommon (due to liability issues of the ISP) and might delay filtering significantly. Due to these differences in filtering approaches on these two distinct levels, it is difficult to find a common technical ground for web labels that fulfils the requirements for both systems.

5. Challenges in electronic age classification data and its interoperability

The introduction to the ecosystem of age classification and age labelling shows some diversity in classification procedures and origins as well as regards the approach to process this information internally or to share it with third parties.

Diag. 8: Age classification approaches



The variety in deploying age rating and labelling in digital environments does not come at the price of easier electronic age classifications or simple implementations, nor does it lead to rising amounts of classification data being shared. Besides the lacking incentives shown above, this is due to several legal, strategical and practical hurdles that present

challenges to a greater extent of exchange and interoperability of electronic classification data.

5.1 Low "stickiness" of age classification

Traditional age ratings for media content, e.g. feature films, videos, TV broadcasts usually do not stay with the content when distributed digitally. Except in Digital TV there is no common standard that is being used to transmit age-related metadata together with the media content. Where metadata is being exchanged in parallel, often using proprietary data models, it is not guaranteed that the metadata itself will survive the sometimes longer routes in the value web, as shown above. Nevertheless, if a player within that route needs the age classification yet, he will likely request them by the content produce, main distributor or the competent classification body. Hence, the routes of content differ from the routes of age classification data.

5.2 Many versions of the same professional content

Professional media, especially movies, come in many different cuts to different national markets. Either due to perceiving a specific desired age rating or due to publishing versions with exclusive content it is not uncommon that there are three to five different versions of the same movie available – per country. This results in two major challenges in interoperable data:

First, the data might be interoperable, however, the media content often is not, e.g. there might be different classification decisions from different schemes on the same movie, but they do not apply to the exact same cut version. In a worst-case scenario, this might result in no added value of interoperable data at all for these feature films, since no two versions of the media content are the exact same.

Second, this diversity in versions can make it extremely difficult for someone who needs an age classification for a specific content to determine the right classification information among several existing ones. Usually used criteria are its title, language and length; however, the method to time a movie's length might differ from one classification body to another, yielding in uncertainty as regards valid age ratings.

5.3 Validity and trust of age classification stemming from different sources

The sources for age classification can vary significantly, and so does their levels of validity and trust: While digitally signed electronic age ratings issued by a traditional and well renowned classification body regularly guarantees a trustful valid classification information, an aggregated age rating automatically calculated on grounds of user or community ratings might result in an opposite assessment. Moreover, trusted brands like long-established classification bodies have to keep a close eye on the validity of their age ratings from the moment they are being transferred to third parties.

If interoperable age classifications mix up information from different sources showing different levels of trust and validity, the overall acceptance of such an approach will be determined by the weakest sources. This is a strong argument for including options for digital signatures in interoperable age rating data models, like MIRACLE does. This way, validity and trust come within the classification data.

5.4 Forms of provision of age classification data

The possibilities to offer age ratings electronically are manifold. Players that already provide electronic classification data as a B2B service or within walled ecosystems

usually have opted for proprietary approaches: Depending on the individual IT environment and the shape of the existing classification data the affected parties are using customized data model for exchange. In the area of web labelling quite distinct approaches have emerged over time, following the different requirements of the individual use case.

All in all, the different approaches and systems in place are a strong argument for all parties participating to agree to potential new specifications, as any alterations are connected with costs.

5.5 Web labelling requirements are different from traditional age labels

As explained web labelling is aiming at giving filter services and parental control software the fastest possible information about the age rating of a website. Processing age rating information in an efficient manner requires the information (1) to stay at an exact, predefined place, (2) with a focus on the actual age limit and (3) processable before or in parallel to the transmission of the webpage requested. Speed is the main leitmotif.

Compared to that, the requirements of full age classification data is its validity, trustfulness, integrity and completeness of content. The more data a label contains the better. This difference in information objectives makes it very difficult to provide a model that actually encompasses both features.

5.6 Sharing and interoperability as a risk for business models

The examples of walled gardens and user generated content platform show that for these players, sharing (internal) classification data is either not necessary, threatening their business secrets, or even business models. The limited incentives to share age rating data usually results in the situation that the classification knowledge regarding the content is lost for external demanders and consumers. Specifically, the knowledge about potential reasons for a problematic content will not be provided any longer – even if the content is still accessible on the platform. Opening this knowledge to external demanders might suit the individual user's needs much more than one-size-fits-all internal blocking/gating approaches, especially in international accessibility. Interoperable age classification information can provide the opportunity to maximise the value and benefits of user-based age ratings here.

Another example where data sharing threatens businesses are classification bodies that licence their data to third parties for money. In such cases, the fee for content producers and publishers requesting an age classification usually considers the long tail income from such third party licenses; if this income falls apart the operations are in danger of loss-making.

5.7 Motivation dilemma of companies

Where companies cannot estimate the added value of electronic or even interoperable age ratings for their business operations, two arguments are regularly raised when it comes to changing things in the area of providing (better) digital age classification information: The group of companies that already delivered on their promise to make the internet a safer or better place for kids, putting money in the various initiatives and their related tasks, pose the question why they should once again spend money. The other group of companies, mostly not involved in EU-wide child protection initiatives, claim they will not put money in electronic classification as long as this is not necessary

for compliance. The two lines of argument seem reasonable as long as the strategic and policy incentives stay the way they currently are.

5.8 Intellectual property and remit issues

Two further hindrances can be seen in IP rights regarding the brands of classification bodies and their rating their symbols. The age rating from professional classifiers are usually protected under IP law, and so are their brand names, logos and age label visuals. Licenses to use the respective material and ratings are issues to (paying) customers, to ensure the validity of data and to operate the business in a sustainable way. In such cases, the provisions of age ratings can be seen as a service of its own, being provided independent from the media content provision. E.g. a VOD provider has to obtain a licence to use the age ratings form the respective national classification body to be able to provide this data together with the video content offered to her customers. This intensifies the situation of diverging routes of media content and age ratings (see above).

Where content producers or providers, especially in the field of broadcasting, sublicense or sell TV content to third parties, e.g. VOD platforms or other broadcasters, meta-data usually is being sold for extra money. As the content usually is self-rated by the seller, it is deemed an extra item and an additional cost factor within such video package offers. The buyer, however, might be obliged to rate the content herself under the national legal framework. As a complete double check might be necessary, cross-system licensing usually does not include age rating data, as the information does not carry any added value for buyers.

Sometimes, also a geographically limited remit of classification bodies restricts the sharing and/or use of classification data: If a national classification body is only assigned to provide age classification data for its respective country, it might be obliged to restrain from sharing age ratings to third countries or market players from abroad that plan to use the age classification somewhere else. Such limitations as regards its remit pose factual hurdles of classifiers when it comes to internationalisation.

5.9 Culturally related scheme hurdles

Finally, there are hesitations to share classification data across national rating schemes because of the sometimes significant divergences in cultural values, e.g. how national societies judges erotic or violent content, or depictions of alcohol, crimes or drug use. These differences can weaken interoperable electronic classifications' meaningfulness in cases where the originating scheme comes to another age rating then the receiving scheme. Content descriptors explaining the reasons for a specific age rating can lower the risk of "importing" divergent assessments in age ratings, since they might point out content where a second, national look might be indicated.

6. Potential fields of application for MIRACLE-based data

MIRACLE addresses several of the shown challenges already by offering a common data model for age classifications, including options to implement this model for internal and external use cases while providing options to secure underlying business models. This being said, the MIRACLE approach might yield added value for the relevant stakeholders in the following areas.

Processing, transferring and exchanging age classification data in one single format

MIRACLE enables companies that classify their content themselves or with the help of classification bodies or their user communities to use a data specification that is highly adaptable to the specific needs. Using MIRACLE datasets ensures that all age classifications follow the same data scheme, making it easy to process, transfer and exchange classification information within a company, with external partners, or both.

Mapping of existing classification data without touching the underlying scheme

If content providers and classification bodies do not want to change existing classification schemes or IT infrastructures, MIRACLE enables such companies to simply map the categorization to the MIRACLE specification without altering the underlying structures. This way, classification workflows and technical environments can stay the way they are, making MIRACLE an additional/alternative channel to provide age rating data.

Provision of classification data in a comprehensible, machine-readable format

By providing or transferring age classification information in a MIRACLE-based format, additional documentation and transaction costs due to proprietary formats are unnecessary. The more nodes of a network and along the value chain use MIRACLE, the fewer additional APIs or mappings have to be produced.

Streamlined Age Rating Storage and Processing

No matter how many different types of content a company provides, and regardless of the markets it is operating in: By using MIRACLE-based age classification data all the different age rating procedures used will always result in the same data format – without losing the original meaning and information of the underlying rating schemes.

Mapping and aggregating ratings from different sources and national schemes

If a content or platform provider handles media assets that have been classified by a variety of different regional or national legal frameworks, the different age ratings can become impossible to store and process in an efficient way; even more so in cases where external classifications are added due to internal user-based age classifications. By using MIRACLE, such providers are able to map the different schemes on one vocabulary and aggregate the data in one dataset. For media asset management software this is an opportunity with a significant added value: One file, containing all the different age classifications, and machine-readable on basis of only one common vocabulary.

One data format for backend and frontend

The data platform providers make use of within their backend systems offer great opportunities for frontend needs and features. In such cases the data format doesn't change, resulting in one data model fitting all needs – be it internally or demand-side driven.

Frontend labels for provided content regardless of the platform

Providing the classification information to third parties and end users can be a great way of sharing classification expertise that would otherwise stay locked away, enabling the end-user to use third-party software or plugins to process the provided classification data locally. Implementing MIRACLE as a data format for electronic labels enables content providers to provide classification information in an interoperable format, enabling third parties to use the data without much hassle. Browser plugins, parental

control software, embedding partners, intermediaries – they all can make use of interoperable labels provided by the content providers.

Region-specific playout of age classifications

In many countries, content providers are legally or informally obliged to classify and label their relevant content based on the national classification scheme. For internationally active content and platform providers this results in the necessity to provide age labels depending on the user's country of residence. With aggregated, region-country-specific MIRACLE datasets such companies are able to provide exactly that age classification information which is relevant for the individual user.

API-based provision of classification information for third party software and external demanders

In cases of user generated content platforms, companies offer their users a large choice of different content, resulting in highly different age classifications. In such situations, MIRACLE offers ways to provide MIRACLE datasets via an API endpoint, making the data specification both extremely slim and versatile. Demanders for API-based data provisions might be third party software or parental control software providers. The main advantage of using an API in these cases are (a) slim datasets, since it only carries information regarding a specified content and (b) traffic-saving, since the API will only be queried by demanders who request such information. Non-users of electronic labels will not experience any difference in content provision.

Annex: Guidelines and best practices for applying MIRACLE in open web labelling

Taking into account the potential approaches in web labelling and the requirements of both the MIRACLE specification and parental control software, web labelling with MIRACLE data runs into two major challenges:

- Problem 1: A full MIRACLE dataset sometimes is too big for the http response header, or even the html body.
- Problem 2: A vast majority of users will not need the MIRACLE data. Full datasets in each and every page transfer create a significant data overhead.

Hence, full MIRACLE datasets in http headers or html sources are not a feasible option. To maximise interoperability of age classification information, MIRACLE-based data should remain in one block, file or dataset. Otherwise, any form of aggregation and exchange will become unpractical. Moreover, MIRACLE-data will be provided on request only, avoiding data transfer overheads. A central file is, in addition, a solution that is much more scalable. This approach favours file-based approaches (rdf, xml, json) respectively. Moreover, using a full MIRACLE dataset file sticks to the central objective of MIRACLE: interoperability. This way a website's MIRACLE file is interoperable with MIRACLE datasets provided by classification bodies, broadcasters, video on demand providers and app markets, inter alia. Moreover, the requested html site is decoupled from its label, making the parsing independent from the transfer speed of the html page. This is a strong argument for a full MIRACLE dataset either at domain root (miracle.xml) or at a location specified by the content provider. For such approaches, however, an additional request has to be sent to retrieve the file with MIRACLE data. For high-scale filter solutions on ISP level, this might slow down the process of web analysis, resulting in suboptimal service provision. In addition, this request has to emanate from the ISP's core network, which usually is bad practice, as the ISP becomes originator of such requests, not the user (see above).

Web labels based on MIRACLE data thus need to serve two contradicting needs: They need to be easily processable, highly scalable age ratings on the one side (for filtering, age gating), and to provide the most added value possible they will need to provide full sets of classification information on the other (for better consumer information, trust-based analysis, aggregation of classifications). A compromise is to provide a minimal age-related MIRACLE tag, accompanied by a link to the full MIRACLE dataset, either in http header or html source:

From the perspective of filter software providers MIRACLE's <age>-element is the central information for the filtering decision. Using only this information and referencing the full MIRACLE dataset in the http header or somewhere within the html source enables filter software to process this information efficiently. The problem remains that on websites that classify their content in different age groups for different categories or sections of the site, an infrastructure would have to be implemented that is capable of providing different custom headers or custom html tags for the webpages rated differently.

Moreover, websites with huge quantities of differently rated content would have to use huge miracle.xml files, making the files very difficult to parse in a short amount of time. The custom header can pose a workaround when it comes to filter software requests, but if a third party needs the full MIRACLE dataset, a single MIRACLE file is not scalable endlessly.

For these occasions, MIRACLE has to offer such websites the possibility to provide a MIRACLE API endpoint to their service. This way MIRACLE files can be generated on the fly if requested by a third party. The advantages of such an approach are that

- highly flexible MIRACLE datasets are being pulled from a database or are being created on the fly,
- links to internal or external databases are possible, e.g. community ratings, classification bodies etc.,
- API endpoints can be queried without requesting the actual webpage, for instance by filter software crawlers,
- it can an be deployed on multi-user platforms like wordpress.com, blogger.com etc. or on infrastructure level of host providers offering their clients miracle-based labelling.

However, linking from miracle.xml to an API endpoint results in a second request, which is a loss of time. If a respective automatisation process is implemented, the custom MIRACLE header/body tag could already refer to the API endpoint including the specific request for the file. The latter would result in a complex implementation which is probably only feasible for large websites and big companies.

Before the background of these considerations, the MIRACLE consortium suggests the following implementations of open web labelling:

Provision of a central miracle.xml file at the domain web root

If a site is providing a full MIRACLE dataset, a file "miracle.xml" should be provided at the domain's web root. Within the miracle.xml file, the scope-url elements are able to provide simple age differentiated URI-based scopes of website parts in case a content provider hosts content for different age groups. However, to identify the age classification of a specific URL the miracle.xml file will have to be read and parsed until the scope-url value fits the requested URL pattern, which makes it subpar to hierarchically shaped, URL-centred approaches like age.xml.

From the webmaster's perspective, labelling the full second level domain within one single file is the most feasible way to control the web labelling process. Small, medium and big websites that provide content for different age groups within simple different categories will also be able to us full MIRACLE files, however with two or more datasets of <age-declaration>s.

Provision of a custom MIRACLE http header (with <age> and link to full dataset)

To label a website efficiently, especially for ISP-level filtering or in cases a website admin does not have the rights or abilities to provide a miracle.xml file at web root, one should provide the URL's age rating and a link to the full dataset by using a custom http header, e.g.:

```
X-content-age:12
```

```
X-content-age-url:
http://www.example.org/pathtofullMIRACLEfile.xml
```

The custom http header allows the admin to provide the <age> element within the header already, enabling third party filters to easily process the webpage.

Provision of a MIRACLE API endpoint for large scale sites with various ratings

MIRACLE v2.0 features, beside other amendments an element within the scope block: <scope-api>. For large websites with many different URL-based classifications, this element flexibly links to an API endpoint that delivers single MIRACLE datasets on request, after queried with the URL. This way, websites with large age classification metadata repositories do not have to put all information in one xml file, but likely refer to the respective API that can be queried accordingly. Its implementation depends on the specific technical possibilities: A webmaster might either provide a central miracle.xml featuring the API endpoint that can be queried with the URL of the requested website, delivering a full MIRACLE dataset for that specific URL. Or the webmaster implements a feature that generates the custom http header in a way that the link to the full dataset already consists of a respective API request including the URL (or other parameter).

Providing age.xml to optimally fulfil filter software requirements

The suggested MIRACLE implementations all have the drawback that their structure is focussing on <age>, and not on <scope-url>. In cases of websites with content for several different age classification the shown approaches would always deliver less than perfect solutions from the perspective of filter software. Since age.xml has been developed for URL-based scopes and filter-optimised analysis specifically, the MIRACLE consortium endorses the age.xml specification in general. Since JusProg, creator of the age.xml specification, is one of MIRACLE's consortium partners, there are several junctions between the two specifications to enable content providers and webmasters to use both approaches, to provide age classification information in the most interoperable and processable way:

- The age.xml label generator provides a valid miracle.xml file in parallel to the age.xml file. This way MIRACLE- and age.xml-based filter software will be able to process the information.
- The custom header X-content-rating is being used by MIRACLE and age.xml.
- The age.xml specification offers the option to reference a full MIRACLE dataset.

MIRACLE's age labelling best practice recommendation

1. Use a miracle.xml file at web root.
2. Use a age.xml file at web root.
3. Use two custom headers on your webpages (X-content-age with the age rating of that webpage and X-content-age-url with a reference to the MIRACLE file).

By making use of the advantages of both MIRACLE-based and age.xml-based labelling all existing and future approaches to use the age rating provided are being served in an optimal way, no matter what route the software developer has chosen to analyse the provided age classification data.