Institutional Services and Tools for Content, Metadata and IPR Management

Pierfrancesco Bellini, Ivan Bruno, Paolo Nesi, Michela Paolucci
DISIT, Distributed Systems and Internet Technology Lab,
Dipartimento di Ingegneria dell’Informazione, Università degli Studi di Firenze,
Firenze, Italy
{pbellini,ivanb,paolucci}@dsi.unifi.it, paolo.nesi@unifi.it

Abstract — Multimedia services need to be supported by content, metadata and workflow management systems to efficiently manage huge amount of content items and metadata production as in the case of cultural institutions. Online digital libraries and cultural heritage institutions and the other portals of publishers need an integrated multimedia back office in order to aggregate content and provide them to national and international aggregators. Different solutions need to be defined and adopted in order to efficiently organize and manage large amount of data and processes to cope with them. The aim of this paper is to formalize and discuss about requirements, design and validation of an institutional aggregator for metadata and content, coping with IPR Models for providing content towards Europeana, the European international aggregator. The first contribution presented in this article corresponds to the identification of the Content Aggregator requirements and needs, and thus the definition and realization of a distributed architecture and workflow solution satisfying them. A second contribution consists of the formalization of IPR Model and solution that allow shortening the activities of IPR resolution and avoiding the assignment of conflicting rights/permissions. The proposed solution and models have been validated and results are reported in the paper in the case of the ECLAP service. ECLAP Content Aggregator has been established by the European Commission to serve Europeana for the thematic area of Performing Arts institutions.

Keywords: institutional archive; content aggregator; workflow; IPR Model; IPR management; grid computing; metadata enrichment; metadata validation; semantic computing.

I. INTRODUCTION

With the introduction of web 2.0/3.0, and thus of data mining and semantic computing, including social media and mobile technologies most of the digital libraries and museum services became rapidly obsolete and were constrained to rapidly change. Thus the online digital libraries and cultural heritage institutions such as ACM, PubMed and IEEE and the other portals of publishers need an integrated multimedia back office. In most cases, the cultural institutions see their content ingested, promoted, distributed and exploited by final users via online commercial partners (e.g., Google, Amazon, YouTube), that may take benefits to commercial resell and/or via advertising. In Europe and in US, most of the cultural heritage institutions aggregate content and provide them to national and international aggregators such as Europeana in Europe [Europeana], Library of Congress in US [Congress]. Europeana has more than 24 million of contents coming from more than 2200 different content providers, and about 100 Content Aggregators (such as: thematic aggregators and/or national aggregators). Therefore, in order to cope with these new needs and participating in the aggregation process towards large collectors and archives, the institutional services need to be supported by content, metadata and workflow management systems to efficiently manage huge amount of content items and metadata coming from different sources (local archives, museums) that have to be harmonized and managed for internal purposes and public distribution and aggregation.
In more detail, large Content Aggregators have to cope with the complexity of aggregating heterogeneous collections coming from several archives (different metadata models, different languages, different databases, etc.), by regularizing and augmenting metadata with specific information needed for the international libraries, such as clearance and formalization of intellectual property aspects, regularization of references to standard dictionaries and taxonomies, resolution of references to well-known person names, locations, dates, etc., and finally also adaptation of content format to the new formats for internet distribution and streaming. In most cases, the Content Aggregators have to put at disposal of their personnel and content providers collaborative tools for content management and for supporting massive content processing on their back office. Therefore, specific workflow and metadata processing and enrichment (name resolving, date analysis, linking with open data, creation of relationships, etc.) are needed to cope with the content aggregation and publication duties, at reasonable costs. To this end, specific solutions and architectures are needed to cope with the above mentioned aspects, also taking into account traditional workflow and content management systems.

A. Contributions of this article

In this paper, the main design and architecture aspects and solutions of ECLAP cultural heritage thematic content aggregator are reported together with their validation and usage. ECLAP (European Collected Library of Artistic Performance, http://www.eclap.eu [ECLAP]) has been set up by the European Commission ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, Theme CIP-ICT-PSP.2009.2.2, Grant Agreement N°250481. ECLAP aggregates content and perform metadata semantic enrichment for Europeana [Europeana]. Europeana (namely the European digital library) does not collect content item files, but only collects classification information (i.e., metadata) including the original URLs to the content. The collected data and URLs refer to the original content owner and/or to the Content Aggregator who provided the content metadata. According to this approach, ECLAP plays the role of Content Aggregator for Europeana. At the same time, ECLAP is a live lab in which several new research results and technologies and solutions in the area of semantic computing and social media have been developed and put under trial of the final users and institutions. On this regard, ECLAP has been open for both content and research result experimentations, and presently comprises more than 35 prestigious international institutions coming from 13 European countries plus: Russia, Chile and South Africa (see the full list on ECLAP portal [ECLAP Partners]); 18 of them started since ECLAP early set up in the 2010, others joined ECLAP successively as affiliated and networking partners.

The contributions to advancements of the state of the art reported in this article can be mapped on the following aspects and correspond to:

- The identification of the Content Aggregator requirements and needs, and thus the definition and realization of a distributed architecture and workflow solution satisfying them. The aggregators have to cope with huge amount of content and metadata, and perform their massive processing for semantic enrichment and publication. To this end, a specific Workflow Model for content aggregation and enrichment management is needed to integrate both human and automated back office activities for semantic computing such as: semantic enrichment (annotation, tag, comments, skos-fication, etc.), content adaptations, metadata quality estimation, validation, translation, etc. The Workflow model also implies the identification and the management of roles for the collaborative management of the content activities. These aspects are strongly valuable for many thematic and/or national aggregators that are becoming every day more diffuse.

- The formalization of intellectual property rights, IPR, modeling and management for shortening the activities of IPR resolution and avoiding the assignment of conflicting rights/permissions. Only content with a clear, consistent, coherent and formalized IPR conditions can pass to Europeana to reach the large audience and repositories. If the metadata are frequently requested to be of public domain, the content may be protected. A protection model can be to adopt conditional access model defining permissions and rights. They can be different for different categories of people, and for different kinds of content. Several contribution on standards have been taken into account and the solution proposed is covering aspects and solution not modeled by standards and solutions at the state of the art. Moreover, an additional addressed key factor of complexity is related to the concrete possibilities of guaranteeing the content protection and the respect of rights in a
provided platform and contexts. To this end, a specific IPR Model and IPR Wizard tool model production have been defined to assist the content providers in assigning the licensing and conditional access rules to the content.

The proposed solutions have been developed and validated in ECLAP European Commission project that started the ECLAP infrastructure and service. Until now, ECLAP has ingested and processed about 180,000 object files (images, document, video, audio, 3D, braille, e-books, etc.) with more than 1.1 million items (pages, images, video, audio), and billions of information managed. The workflow validation results and experiences are also reported in this paper.

The paper is organized as follows. In Section II, related work is presented to put in evidence the state of the art of technologies and research in the sectors related to the automated content management and workflow for Content Aggregators, and on IPR management and conditional access for cultural heritage content. Section III reports the general requirements and the ECLAP architecture for institutional services and tools with workflow details. In Section IV, the detailed and formal aspects related to the IPR Model, IPR Logical Model and the corresponding IPR Wizard tool are presented. The exposition puts in evidence the innovation provided with respect to the state of the arts and the related advantages. The main advantage is creation of permission models without inconsistencies (i) among rights/permissions, (ii) among granted permissions and Europeana.rights models; and (iii) in the hierarchy of granted permissions to the users. Section V presents the results obtained in the massive usage of the solution proposed, providing statistical analysis related to the IPR Wizard tool usage on ECLAP portal and content ingestion and distribution services. Conclusions are drawn in Section VI.

II. RELATED WORK

This section reviews the state of the art techniques and research in the sectors related to the automation of content management and workflow for Content Aggregators, and about IPR management and conditional access for cultural heritage content.

A. Automation of Content Management and Workflow for Content Aggregators

The Content Aggregators have to cope with a large variety of content coming from several content providers, with the aim of processing the content in several manners, targeting the final goal of publishing the content towards several different directions and exporting them to other services. The main problems to be solved are related to the massive management of content processing and the automated application of semantic computing algorithms. A category of tools to cope with these problems are typically called media or semantic grid tools [Bellini et al., 20011]. Among the several solutions: AXCP provides an integrated approach to perform both content management and semantic computing [Bellini et al., 20011]; MmGrid supports interactive applications with graphics, rendering, streaming, and tele-immersion [Basu et al., 2003]; GRISINO combines web services with intelligent content and grid computing, where its integration with workflow management is performed via web service [Toma et al., 2006]. GridCast is a service-oriented architecture for broadcasting media via IP [Harmer et al., 2005]. GridCast provides some content management capabilities, mainly for broadcasters as controlling playout servers. In [Volckaert et al., 2008], the usage of grid solutions for media production and distribution have been introduced, focused the application of the technology on the content delivering and management. In [Messina et al., 2011], HMNews system for automated aggregation and consumption of information streams from digital television and the Internet has been presented. These solutions provided support for video segmentation, annotation and indexing, while many other aspects have been not considered, such as text processing, semantic computing, etc. In the management of News, it is quite common to perform natural language processing analysis to extract information. In the usage of news, NewsML as well as other formats provides a limited number of metadata [Bellini, Bruno, Nesi, 2011]. In Google News tool [Dolan and Pedersen, 2010] automatically aggregates information extracted from newspapers and magazines to offer personalization service displaying news articles tailored to the user’s interests. Most of the high-level activities may be passed from grid to a workflow
management system, which also plays the role of integrating other tools via programming [Yu and Buyya, 2005], [VanDerAalst and Van Hee, 2002]. In general, the above mentioned solutions and tools provide limited semantic computing capabilities for processing content metadata for cultural heritage content; where metadata are typically provided in several different metadata formats, standards and dialects. To this end, specific tools called metadata mappers are typically used [http://mint.image.ece.ntua.gr].

On the other hand, the metadata enrichment has to be addressed by specifically algorithms and tools. For example, processing metadata and exploiting data mining and natural language processing for extracting and disambiguate VIP names and linking them on dbPedia, to extract and disambiguate geographical names, to understand and disambiguate dates, assess metadata quality, etc. Also these aspects are specific of cultural heritage content aggregation and have to be integrated into the workflow and semantic computing tools by programming. If the workflow management tool can support the harmonization of the work to be done, by who and when, the media grid has to provide support for parallelizing the process and executing these complex algorithms on multiple computer machine nodes.

B. IPR Management and Conditional Access for Cultural Heritage Content

The management of the intellectual property on a cultural heritage archive and portal that aims also at the content distribution may present a number of problems and phases, such as:

A. rights clearance which corresponds to the activity for the identification of the rights that effectively are owned by the distributors and content provider. This corresponds to the phase in which the CP identify on contracts of the rights/permissions that could be potentially grated/exploited, or clear them creating new contracts with effective right owners for the identified content to be distributed. This activity is typically performed by legal and administrative personnel via desktop work.

B. formalization of the right profile for a given content on the web service and archive, the right profiles has to defined for each user category and device and context the possible provided rights (these rights have to be unambiguously defined and have to be consistent each other, for example it has not sense to provide the possibility of change and cut a video and not that of streaming; in many cases a right implies a set of other less constrained rights). The activity of formalization can be technically performed only according to identified rights extracted from phase (A). The formalization has to be performed before the workflow management and content production since some of the activities performed on content can only acted exploiting specific rights. For example, a file can be adapted only having the legal permission to do that.

C. right enforcement, that is the effective application of the formalized of right profile defined in phase (B) with some tools with the aim of enforcing the respect of the rights/permissions during the content access and usage. This may implies to adopt one or more specific technological protection solutions, such as special players, encryption model, certification and authentication, etc. [Bellini, Nesi, Pazzaglia, 2013]. The activity of right enforcement is typically performed by the tools which are used for content access, i.e., the player, the browser, the decoder, etc.

Moreover, when the content is delivered to third parties, such as the posting of metadata to Europeana, specific guidelines and formulation of related IPR have to be followed; thus requesting additional information that has to be formalized and not conflicting with the legal aspects identified in (A) and the formal definition of the IRP model derived in (B) phase [EuropeanRights].

The formalization of right profile in (B) can be performed by using an Intellectual Property Rights, IPR, modelling, also called licensing. Licensing models aim to regulate the possible uses of the digital content (e.g., use on a specific device, use for a specific time period, use on a specific territory, use for educational purposes, etc.). In order to allow interoperability, several standards have been developed for the formalization of licenses and access control policies like the MPEG-21 REL [Wang, 2004], ODRL [Iannella et al., 2012], OASIS XAMCL [Moses 2005]. For the definition of the actions, possible rights ontologies or vocabularies could be used such as MPEG-21 RDD [Wang et al., 2004], the ODRL vocabulary [Iannella and Guth, 2012] or the Access Management Ontology. In [Chang et al., 2003], a visual tool for defining authorization workflow models for e-commerce
application has been proposed. The above mentioned models allow the formalization of licenses authorizing a given user to exploit certain rights on a specific identified content. On the other hand, a generic IPR Model should to allow its formalization to create several licenses and to formalize the rights that could be exploited for a given content. This approach has been adopted to formalize the Potential Available Rights, PAR of a given content owner, in AXMEDIS [Bellini, Bruno, Nesi, 2011]. Moreover, the Creative Commons licensing framework [CC] allows users to formalize the usage of content that the owner would like to share, applying some legally formalized restrictions on its reuse (e.g., no commercial use, attribution, no derivative, etc.). This means that the CC models address only partially the problem of phase (B) of right formalization, since CC models assume that the content can be distributed.

All these aspects are still far to be used as IPR Model for defining the whole set of rules to be applied to a content according to the different kind of users and devices. These aspects are addressed into the ECLAP IPR Model and tools.

On the other hand, when the content is distributed as in broadcast or via internet for commercial purpose, all the above mentioned 3 phase from (A) to (C) have to be covered with some coherent solution. In those cases, the IPR management systems can be classified in Digital Rights Management (DRM) systems allowing to control ‘all’ the possible different uses of a digital content (e.g. distribution, enhancement, adaptation) OR Conditional Access Systems (CAS) allowing to control only the access to the digital resource that is provided with some encryption. This means that the main difference from DRM and CAS resides mainly on the enforcement of rights/permissions and constraints that increase the complexity of technological protection solutions adopted. For example, to provide the guaranteed that once downloaded a content it will not be accessed by some on else. Currently the media industry is much more focused on CAS and some solutions are present mainly for audio/video content access developed by Microsoft, Apple and Adobe for PCs/mobile/tablets and by Irdeto, Nagravision and others for satellite, IPTV, cable TV digital transmissions. On the other hand, in the case of cultural heritage content, a limited number of rights can be owned by the archive, while the content distribution may be performed towards specific protected user categories such as: impaired, students, researchers, professors, for which some of the right constraints have to be relaxed and these aspects have to be formalized into the right profile identified in phase (B).

After the analysis of related work and state of the art, a number of unsolved problems have been identified, such as an integrated solution and framework for Content Aggregator including (a) workflow management and semantic computing, (b) integration of automated computing and user collaboration, and (c) IPR modeling and management. Thus, the architectural solution designed and developed for ECLAP can be of value for many institutional archives and content aggregators.

III. GENERAL REQUIREMENTS, ECLAP ARCHITECTURE AND WORKFLOW

This section reports the general requirements and the ECLAP architecture for institutional services and tools, also providing major workflow details.

A. General Requirements

Content Providers, CP, provide at the ECLAP Content Aggregator, CA, both content files and metadata. CP can be archives, museum, theatres, educational institutions, foundations, etc. According to ECLAP workflow, the obtained metadata have to be sent to Europeana only after that the metadata have been enriched and linked to a reachable digital resource and when the IPR issues have been correctly defined, thus a good quality level is reached. In this context, ECLAP does not act only as an intermediary between the CP and Europeana but aggregates and maintains the content files accessible, via the links to the content that ECLAP provides to Europeana and that lead to the real content file. Moreover, ECLAP supports the CPs in all their work: starting from the ingestion phase,
passing through the definition and management of IPR permissions and licenses on contents, managing the real content and providing to all the ECLAP users the typical services of a Social Network. The main requirements of the back office tools for cultural heritage content aggregators are reported as follows. Moreover, a large number of detailed requirements have been identified as reported in the documents accessible from [ECLAPSSupportPage].

The Content Aggregator infrastructure has to be capable to satisfy the following requirements.
The system has to be capable to ingest a wide range of metadata formats (such as: XML based or Dublin Core, METS, MPEG-21, etc.) coming from different channels (http, ftp, oai-pmh, etc.), map them on single unified model without losing any semantic meaning of the original metadata model. The process of ingestion includes to acquire both metadata and content files, link them together, collect them on a suitable storage, and ingest IPR information if directly available at the ingestion time. The ingestion process may be performed from web pages uploads, and/or from batch processing from databases and/or crawling (for example from some Web Services, and/or LOD services). The ingestion activities can be partially performed by automated tools. For example, the definition of metadata transcoding maps (from the former format to the ingestion model) has to be manually performed by skilled personnel.
The solution has to provide support to allow skilled personnel to perform manual content enrichment, such as metadata language translations and validation; addition of comments/annotations and tags; social media promotion; voting/rating; promoting; publication to other portals and social media; editing and performing corrections on metadata; quality assessment, etc. This implies that each single expert may perform several different activities on content in the workflow and life cycle. On the other hands, they cannot be granted to all users without creating problems, since not all user have the whole set of expertise. Therefore, there is the need of granting different authorizations to different users, for different languages, for different CPs.
The architecture to be capable to perform automated activities, such as: estimation of technical parameters (duration, size, etc.), extraction of content descriptors, indexing, automated translations, searching for VIP names, geonames resolutions, linking with LOD, metadata assessment (completeness and consistency [Jung-Ran, 2009], [EBellini and Nesi, 2013]), IPR (Intellectual Property Rights) verification. Among these activities, the content adaptation and repurposing has to be included to cope with the production of the different file formats to be produced to satisfy a large set of distribution channels and final users’ devices. For example, ingesting video in any format and producing the multiple formats for the distribution towards: PC, iPhone, iPad, smart phone, SmartTV, etc.
The back office architecture has to be scalable to cope with large number of transactions on metadata information and activities in the back office per day. Therefore, the execution of a massive and distributed processing on content resources is needed, and can be provided by using some media grid solution.
The approach has to support more than one workflow life-cycle for content at the same time that can be associated to specific content or content collection. For example, different CPs may need and prefer different workflow model; for example to exploit different or differently the several automated content processing algorithms and the experts. These aspects are part of a flexible workflow management that should also support the definition of different user roles for different activities and CP.
The proposed solution has to harmonize the activities of human and automated processing among the above mentioned. For example, identifying when the human actions are needed, taking trace of the performed manual activities, blocking the humans when the automated elaboration has locked the resource and vice versa. Among these activities of the workflow the CA have also the needs of (i) logging and keeping trace of metadata versioning: to keep trace of the work done and changes performed, (ii) formalizing and managing a number of different roles / capabilities to be assigned to the ECLAP users (i.e., enricher, publisher and validator), (iii) providing user accessible tools for: multilingual metadata editing, IPR managements, massive content editing of some specific object status associated with metadata (workflow status, IPR, public/non-public, tags, groups, etc.).

In addition, the Content Aggregator infrastructure has to cope with the IPR Model formalization, assignment and verification. It also implies that the IPR model may regulate uses/accesses to the digital content, and the exploitation of rights about the content manipulation and reuse according to the owner rights. Many other issues
related to the IPR management in cultural heritage communities playing the role of CA content provider services have to be taken into account. A different IPR Model can be applied on content collection or sub collection by each CP. The final collector of metadata, such as Europeana, has the need to be authorized of providing the metadata for several different purposes, for creativity, for cultural promotion, for education, etc. Moreover, the process of enriching cross media multilingual metadata related to each content is complex and is facilitated if the metadata are of public domain (e.g., CC0 1.0 Universal Public Domain License) and augmented by technical and multilingual metadata, while maintaining the same accessibility.

B. ECLAP Architecture Overview

In this section, an overview of the ECLAP architecture is reported with the aim of presenting and highlighting the components involved and managed by the workflow and by the IPR management. The ECLAP architecture has been designed to satisfy the above presented requirements for Content Aggregators, and presents three main layers: ECLAP back office, distributed storage shared with the several servers and processes, and the ECLAP Front End (see Figure 1).

The ECLAP Back Office is implemented by a media grid on a set of servers to implement a parallel and scalable architecture for scheduling and managing processes to automatically perform back office activities. The back office activities can be grouped in the categories of Content Processing and Semantic Reasoning. Most of them are activated when content is updated (Update), deleted, exported, etc.
As regards the content processing, it includes the activities of:

- **content ingestion & harvesting:** metadata ingestion from the several CPs performed by means processes of content harvesting and acquisition. Metadata are typically retrieved from the CMSs and databases of CPs by using several kind of protocols (e.g., web services, rest, OAI-PMH), while digital resources are downloaded from the provided URL or FTPs. The metadata are provided in different schemas and have to be mapped according to the ECLAP ingestion semantic model. The mapping process is realized by using MINT metadata mapping tool and service of NTUA [MINT], that collect the original metadata set and expose an OAI-PMH server of mapped data. On the other hand only a small part of the content provided reached the Aggregator infrastructure by using this massive model for metadata. When digital resources are big files, they are provided by using physical device and/or ftp. In this case, ECLAP just starts with metadata ingestion (from web or mapping server) and then, when the digital resources are available, the joining of metadata and content is performed.

- **content production and adaptation** consist of converting the acquired digital files (audio, video, images, documents, etc.) into the several formats needed to play them on different devices and services. The content production process mainly consists in simple file format conversion for images, audio files, documents, etc., the extraction of thumbnails, and of technical metadata descriptors. On the other hand, the content adaptation process presents a certain complexity due to the creation of video in different formats and resolutions: from high definition to low bit rate for mobile, in flash and MP4, with suitable formats for download, streaming and progressive download. These kinds of processes are time consuming and thus may be performed in reasonable time only on a media grid infrastructure.

- **Multilingual indexing** of cross media content to provide search service on the ECLAP Front End. In this cases, the complexity is mainly due to the heterogeneity of metadata for the different kind of resources, and the optimization indexing model to maximize the precision and recall with respect to the user intention [Bellini, Cenni and Nesi, 2012]. The classical incoming metadata (Dublin Core, Taxonomy, Groups, Performing Arts metadata, workflow type,..) are augmented with technical metadata built by analyzing the digital resource: (i) content format (document, audio, video, image, crossmedia), (ii) content type (file format), (iii) structural information (size, duration, number of pages). The produced enriched metadata and digital resource are aggregated and stored in the corresponding database. In specific cases, the system requests the metadata translation of metadata by using tool or web service for text translation. Then, metadata are indexed to make the content ready for access on the portal.

- **assessment and quality assessment.** This process consists in the general quality assessment of the ingested content according to identified quality parameters to identify the content that does not present the minimum level of quality (completeness and consistency of metadata with respect to the acceptance level imposed). The quality assessment has to be performed by using different quality models and parameters in different phases of the content lifecycle. For example, in order to assess the quality of candidate content to be published towards Europeana; estimation of social network analysis metrics and relationships; estimation of several kinds of metrics and statistics on users, content and groups are adopted (for example, user behavior, query appreciation, content appreciation, group activity, ..). The quality assessment is also requested in the cases in which metadata have to be massively corrected and or verified for a given CP. All the estimated metrics, statistical analysis, and usage data about users are at disposal of the different administrators and for the end users (in different forms and aggregation, and in some cases cleaned by the user ids), according to the Terms of Use of the service.

The activities of **Semantic Computing** are related to the following processes. Semantic Enrichment allows enriching the metadata by using natural language processing algorithms to extract person names, geographical names and dates; with the aim of finding synonyms, disambiguation of names, geographical names and dates; connecting person names with dbPedia; connecting geographical names with geonames, and regularizing dates for temporal navigation. Once identified the geonames the resolution of the GPS coordinates is performed. Another
activity of semantic computing consists in computing similarities among users and content, to produce suggestions via the portal and newsletters. All these activities are time consuming and implies the connection with external services such as dpPedia, Geonames.org, mailing server, etc.

The back office activities regarded as **Metadata Import Export Services** include processes and services for: producing metadata in the EDM (Europeana Data Model) mapping a part of the ECLAP semantic model; exporting of metadata via an OAI-PMH server to allow Europeana harvesting them; producing LOD of the ECLAP semantic model [Bellini and Nesi, 2013]; and finally the export of user performed queries towards Europeana (exploiting their API), with the aim of collecting from Europeana possible results to be shown on the ECLAP Front End portal. This service allows to the final users to see not only the results of ECLAP search engine but also those of Europeana. This approach can be used to integrate Europeana content into ECLAP archive, if the identified is accessible according to the Europeana.right model.

The **Social Network Interoperability** processes include activities for the integration of ECLAP infrastructure with third party social networks and services such as: Facebook, LinkedIn, Google, Digg, MySpace, twitter, etc. The interoperability consists of a set of tools and solutions based on authorization and certification to: (i) migrate contacts from those services towards ECLAP by exploiting the OpenAuth standard and protocol, (ii) the exploitation of the social portals API to facilitate the propagation of ECLAP posts into those services.

In the ECLAP back office, the elements for parallel and distributed computation and scheduling have been implemented by using the AXCP content processing tool [Bellini et al., 2011], thus developing a set of new and specialized processes and plugins with respect to those available on the standard edition. The AXCP scheduler can put in execution periodic and on-demand processes. Any process can verify the conditions for requesting additional distributed processes, for example to cope with the arrival of user generated content, the production of messages and newsletters to be delivered in real time, the arrival of new content to be ingested, etc. The above mentioned processes are synchronized with the workflow status and activity performed by the administration services of the ECLAP Front End.

![Figure 2 - ECLAP Workflow diagram for Europeana content life cycle. The activities which can be performed without looking the content are note reported.](image)

In order to better understand the content and metadata management, it is useful to describe the scenario in terms of content status, workflow roles, and procedures, etc. All the back office processes work concurrently in accessing
and updating information on the shared and distributed storage: back-office services could access and process content in parallel on multiple computational nodes concurrently with the users’ activities performed on the ECLAP Front-End. Several thousands of new content and complex transactions per day have to be processed. Activities of metadata enrichment, content adaptation, validation, IPR definition and assessment have to produce consistent results. To manage the concurrency and to guarantee a safety access to the content mechanisms of lock-unlock on data have been defined. Each content item managed in ECLAP is associated with a specific workflow and lifecycle. For example, the workflow state diagram coded for content that is planned to be published towards Europeana is described in Figure 2. In this case of Europeana-based workflow, content can be: proposed; uploaded/ingested; automatically enriched by the processing rules; manually enriched by humans for example via the metadata editor; manually or automatically associated to an IPR model; accessed and validated by humans; automatically assessed to verify the minimum quality requirements by the automated tools; and finally proposed to be published on Europeana.

The ECLAP Front End provides a number of services which can be classified in Administration, Qualified and End-users Services. It provides a social media style front end service with more than 2500 expert and qualified users; directly linked via service oriented interfaces to the back office capable to really cope with the complexity of managing a complex workflow and content processing and ingestion.

The Administration Services are those related to the management of: content, users, workflow, social networking, statistics and metric analysis rendering tools. They also include tools for managing: configurations, spam, newsletters, blogs, web pages, events, calendar, forum, GUI translations, etc. During the content life-cycle, the application of massive actions on content could be needed: changes in the workflow status, changes in the metadata and IPR, and addition of details in the metadata sets. Specific actions are also needed to maintain and manage the content and work both on single and multiple content elements such as: delete, update metadata, and publish content uploaded by common users.

The Qualified Services are a set of tools at disposal of qualified personnel, such as the group administrators, and the trusted users. Among the most relevant qualified tools we should mention: Metadata Editor (for editing, translating and validating metadata according to the engagement requested by the workflow); IPR Wizard to define IPR Models and associate IPR Model to content; Workflow Editor to change the workflow status of content and move content into different workflow kinds and status if needed; e-Learning Course Editor for the authoring of courses; and the Professional Upload service to upload set of content elements via a professional web page for metadata ingestion, multiple languages, multiple instances for each metadata field, etc.

The End-User Services are oriented to support ECLAP final users’ activities on web and mobile. The most important tools related to content workflow/IPR and users’ activities are reported as follows:

- **Cross Media Player, with CAS (Conditional Access System)** is capable to produce and play on browser and mobiles: documents, video, audio, e-pub, 3D, animations, slides, etc., according to the device and browser used. It also provides a CAS which enforces the IPR Model to avoid the access of un-authorized users, provides video streaming in progressive download selecting the video format and bitrate that can be sustained by the established connection, and presents the links to perform content download and/or embedding according to the IPR Model associated to the content;

- **Metadata Viewer** provides metadata in multiple languages: DCMI metadata consisting of a set of Dublin Core fields; Taxonomy as multiple classification terms selection; Groups assignment (ECLAP Groups); Resource data by selecting one or multiple files; Workflow type associated with the content life-cycle. The Metadata are also enriched by links to VIP names of dpPedia, geonames, and to ECLAP users mentioned in them. The same module, also presents the IPR Model of the selected content with respect to the different kinds of users to clarify the content accessibility;

- **Geo Mapping and QR**: Geo mapping technology is used to provide content access via geographical coordinates and may exploit the Google Map service for rendering and access. The administrators can use the
service to set up GPS coordinates to content which has be to activated in those locations when the user is move there with its mobile. The QR markers can be associated to content URL for shortening the content access, which in turn may have GPS coordinates and thus may also generate positioning on maps;

- **User Generated Content** provides a service to the final users which intend to upload content on the portal. This service is connected to the back office ingestion service. The same service allows the addition of comments and taking votes to the content. Comments as new web pages have to be ingested and indexed in real time to allow rendering them in real time to the other connected users.

Others services are reported in the following. **TimeLine Navigator** provides a browseable model for navigating among dates recognized and associated with content (date of production, first performance, last update, etc.);

**Social Graph** provides at the users a browseable model for playing and navigating among the several relationships established by implicitly citing users, resources, names, geonames, and open data into metadata (see the back office NLP algorithms);

**MyStoryPlayer** provides access for producing and accessing audiovisual annotations and synchronous playing of multiple videos (for example for multicamera view);

**Aggregation Tools** support the production of content aggregations such as: playlists, collections, courses and also comments and votes;

**Mobile Player** provides support for smart phone e pad/tab to access at the right player and content format;

**Search and Advanced Search** module allows the search and retrieval of content with faceted support, by using full text as well as advanced operators and interface for querying. In addition, it also perform the query expansion by translating and expanding keywords and propagating queries to other service portals such as Europeana; thus exposing results to the users;

**Suggestion** module provides suggestions about other users and content in different contexts of the portal usage; it also provides at the users ready to access content lists, such as: most played content, recently uploaded, most voted, content of colleagues, keywords clouds, query cloud, your content, etc.;

**Networking Multitenacy** provides support for creating and working on groups, pages, events, forums, blogs with a multitenacy approach; thus a custom graphic shape for each group of users is provided;

**Social Media Connection** provides support for social network interoperability exploiting back office processes for: migration of contacts from other portals, migration of promotional and publication posts towards other portals, and social networks, etc.

C. **Authorizing Roles of the Workflow Users**

In order to implement a high quality content enrichment process, each specific activity has to be granted to specific people according to their skill, language and the identification of the institutional Content Provider. In the previous sections, tools and content status have been presented. On the other hand, according to the workflow presented in Figure 2, qualified users are delegated to perform specific activities on the content by using corresponding tools. Metadata Editor is the tool for editing, enriching and validating metadata. It allows editing the full set of metadata. Changes made on metadata are tracked to maintain their history including who made the change and when to have the evidence of the work performed and eventually recover wrong situations. According to the user role and workflow phase, the Metadata Editor is activated for enrichment or for validation. Therefore, for the Workflow management, specific user roles have been defined with their parameters which are languages assigned and CPID. The CPID is the Content Provider ID, who has provided the content, which is supposed to be the content owner of the rights needed for content changes and distribution. Therefore, the identified user roles related to the content management and the above presented workflow are:

- **WFIPR (CPID)**: definition and validation of IPR models, and IPR assignment to the content of the CPID; by using the IPR Wizard and during the Upload for the IPR Model Assignment.
- **WFEnricher (CPID, {Languages})**: metadata enrichment and changes in the specified languages (add, edit metadata), via Metadata Editor.
- **WFValidator (CPID, {Languages})**: validation of the metadata for the identified language. The metadata fields can be singularly validated until the object may pass the whole approval phase, via Metadata Editor.
- **WFPublisher (CPID)**: to take the final decision for publishing content from ECLAP to Europeana. The publishing of single or groups of content can be also performed by using the above mentioned **Content Management** and/or by means of specific procedures on the ECLAP Back Office.
Please note that, the many back-office activities are not associated/authorized with/by any specific user’s role since they are executed as processes running at administrator level in the ECLAP Back Office.

IV. MANAGING THE IPR, IPR MODEL AND WIZARD

The above presented requirements have to be refined to cope with the details and complexity of the IPR management. The final goal is to associate at each content element an IPR Model to define its accessibility with respect to users’ categories and for the publication of complete information towards the general aggregator, e.g., Europeana. Once the CPs have understood, from a legal point of view, their position with respect to their content (the so called phase A of Section II), the phase B (of Section II) has to guide them on defining and assigning to content the IPR Model (with related permissions and conditions). This activity has to be performed in according to the content life cycle and workflow by using ECLAP tools.

The aim of phase B (described in Section II) consists in supporting and guiding the CPs while creating the IPR Model, and in particular to

- avoid to the assignment rights/permissions which are different with respect to those really owned by the CP (those cleared during the phase A of Section II).
- allow the creation of a set of custom IPR Models for each CP.
- avoid the production of an IPR Model which may assign inconsistent rights/permissions on the same object.
- give the possibility to apply and modify an IPR Model to a collection of multiple content without constraining to edit a single IPR model for each content element.
- allow to associate the IPR Model to the contents (single or entire collections) during the content ingestion, or at any phase of the workflow. The assignment of the IPR Model at the ingestion may allow optimizing the activities of content processing, for example, the production of model content formats if the permission for mobile distribution are not intended to be granted.
- allow to insert in the IPR Model additional conditions according to MPEG-21, ODRL, OASIS XAMCL, such as nationality, times of plays, temporal usage (from date to date).

Inconsistency among rights/permissions can be due to the definition of limitations that cannot be enforced in a given context; for example, they may request to avoid images to be copied if they are visualized on a computer (i.e., a snapshot is always possible). Other kind of inconsistencies may be due to the relationships among the rights. For example, according to MPEG-21 RDD, the grant of right of adaptation implies that of copy, the grant of download implies that of play and thus of streaming. These ontological relationships have to be taken into account in defining a tool for IPR Model production to avoid the creation of inconsistent IPR Models.

Before starting on describing IPR model and tools a short introduction on the IPR workflow for the CA is needed (see Figure 3). Content and metadata are typically ingested into the ECLAP infrastructure by following different paths. Metadata may be accessible by taking exports from databases and/or archives, then they are ingested into a metadata mapper for producing data that can be easily ingested in the ECLAP Back Office and storage. Content files/resources are provided via ftp services or uploaded from web, or provided as hard disk.

On the other hand, IPR information is typically not formally associated with content since the beginning. So that the metadata does not contain any reference to the selected IPR Model. Metadata standards, such as DC, METS, MPEG-7, EDM, FRBRoo, etc. do not provide a formal model to represent IPR Models for content access licensing. In the best cases, metadata models provide some string field in which one could store some custom information and reference to eventual IPR Model. On the other hand, Europeana requests to provide content with a certain level of formalization of IPR aspects. In fact, Europeana requests to fill their mandatory EDM metadata field called Europeana:rights by using one of the predefined admitted values [EuropeanRights]. These values cover most of the possible cases, ranging from the public domain content to strongly protected content under some commercial distribution, as better explained in the following.
According to Figure 3, the IPR Managers use the IPR Wizard to formalize the IPR Models (which allows defining the permissions and restrictions, and the coherent identification of the Europeana.rights, as described in the following). Once the IPR Model is defined, it can be associated with content and/or content collection (even if collection contains content of different kind). Thus, the association of the IPR Model to content is considered an integral part of the metadata set, while the IPR Model itself is stored in a separated database. This allows tuning the IPR Model to a large number of content elements limiting the activities on the metadata database.

Once the CP has already produced some IPR Models, the CP can associate the selected IPR Model directly during the ingestion phase. Thus avoiding the manual association of the IPR Model to content via the IPR Wizard and Content Management tool. When content is ingested without any reference to an IPR Model, it is managed by the ECLAP Infrastructure as one that provides the maximum restriction (no permissions to non trusted users). The only permissions are those related to the managed of content at administrator level and by the trusted personnel of the CP owner. This avoids creating situations in which the content cannot be managed for the lacks of rights.

A. IPR Models Definition

Given the diversity of CPs and of the related needs on their contents, a general and flexible IPR Model formalization has been defined. The IPR Model is grounded on the conceptual relationships among rights defined in MPEG-21, while the defined IPR Model presents relevant innovations that simplify the IPR management of cultural heritage content for the Content Aggregators and archives, satisfying the above presented requirements. Classical and standard licensing models such as MPEG-21 REL and ODRL are unsuitable to satisfy the above mentioned requirements and IPR workflow, since they are focused on producing licenses, as previously discussed. As in MPEG-21, ECLAP IPR Models identify what can be done on the content by each user category, and which are the permissions. All the non-authorized rights are forbidden.

The IPR Model has to allow defining a different set of IPR permissions to different geographical area (localization or domains, for example the University of Florence), for different devices, different content kinds and different types of users. In fact, it is quite common to have different distribution contracts and rights for each country. A distributor could be entitled to distribute a content in a country via streaming and may be is not allowed to make the same in other countries. Another important aspect consists in the specification of different rights for different resolutions of the same content. Typically, CPs allow to download low resolution audiovisual content while high resolution versions are kept for trusted users and/or for commercial purposes.

The set IPRModels of possible IPR Models can be defined as follows:
IPRModels ⊂ ModelID × CPID × CPLicensePageUrl × Description × Europeana.rights × IPRLocalization

Where:

- **ModelID** is the set of possible model IDs, each model ID is the identification of the IPR Model, which can be also used in the ingestion phase;
- **CPID** := \{CP_1, … CP_n\} is the set of possible content provider IDs;
- **CPLicensePageUrl** is the set of possible URLs, each of them links to a web page in which eventual conditions to acquire additional rights;
- **Description** is the set of possible model text descriptions;
- **Europeana.rights** is the set of possible license urls used by Europeana as explained in the following;
- **IPRLocalization** := ϕ(Location × ϕ(Permission)) is the set of possible association of country codes with a set of permissions;
- **Location** := \{other, it, en, de, fr, …, or web domain\} is the set of country codes with the additional ‘other’ country representing any other country, or is an IP domain mask;
- **Permission** := \{ p ∈ ContentKind × Right × UserKind × ϕ(Condition) | right(p) ∈ possibleRights(contentKind(p)) \} is the set of possible permissions considering that not all rights can be associated with all content kinds;
- **ContentKind** := \{Video, Audio, … \} is the set of possible content kinds;
- **Right** := \{ play, playMD, playHD, downloadHD-PC, downloadHD-Mobile, Embed … \} is the set of rights, where HD is for High Definition, MD for Mid Definition, etc.;
- **possibleRights**: ContentKind → Right is a function associating for each kind of content the rights that can be used on it
- **UserKind** := \{Public, CPGroup, EduRes, Trusted\} is the set of possible kinds of users
- **Condition** is the set of possible conditions defined as additional constraints (such as those that can be associated to rights in standard MPEG-21, ODRL, OASIS XAMCL) for example: the expiration date, the duration of the validity, etc.
- \( ϕ(X) \) is the power set of \( X \), \( ϕ(\{x,y,z\}) = \{\},\{x\},\{y\},\{z\},\{x,y\},\{y,z\},\{x,z\},\{x,y,z\}\)

**B. Europeana.rights vs Permissions and Restrictions**

The **Europeana.rights** value can assume a limited number of possible values according to Europeana EDM model and data [EuropeanRights]. The document and guideline for the attribution of Europeana.rights does not report explicitly that not all values can be applied for non-fully accessible content. Therefore, a formal and legal analysis has been performed in order to identify those that can be associated with different sets of Permission we identified. As a conclusion of this analysis, the IPR Logic Model divided Europeana.rights in two categories that can be applied to different conditions of IPR Permission. This approach helps the IPR Managers to take the correct decision and avoid the creation of inconsistencies among the Permissions and the Europeana.rights formalization. These two cases are reported as follows.

For **publicly accessible content** it is intended content which is freely accessible on Internet without access restriction (full access permission for public on the portal and outside). Thus, the associated possible values of Europeana.rights are mainly referred to those of CC (Creative Commons [CC]) and others have been identified from public domain, specifically formalized for Europeana (such as the first and the latter of the following list). The CC is a set of legal models to formalize what the users, accessing to the content, have to legally respect. CC models are very different from DRM and CAS since they are supported by any technological solution to perform the right enforcement (see phase C of Section II). The CC licensing models are formalized at legal level and may be associated to a content file by placing a corresponding mark / logo, for example. On the other hand, the content owner associating a CC license does not has any confidence that those legal licenses will be respected by the final
user. In this case, the CA and thus each CP has to be associated with its public domain content a legal statement, among the possible values ordered by the less restrictive to the most (see guidelines for their meaning [EuropeanRights]):

- Europeana: Unknown copyright status (Unknown)
- Public Domain Mark (Public Domain Mark)
- Public Domain Dedication (CC – Zero universal)
- Attribution-ShareAlike (CC BY-SA)
- Attribution-NonCommercial (CC BY-NC)
- Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)
- Attribution-NoDerivs (CC BY-ND)
- Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)
- Europeana: Rights Reserved - Free Access (Rights Reserved - Free Access)

In the case of **content access with some restriction** applied to its access on the ECLAP portal (for example, defining that the ECLAP public users can play a content, but not download it), the CP has imposed some permissions but not all permissions are granted to public users. In these cases, the above mentioned CC licenses cannot be used, since all of them give for granted the full access to content and impose only restriction on the final user behavior. For example, the assignment of CC BY-NC-ND to a content that is distributed with some restriction (for example only in stream), would violated the CC license that grant to the user the content access and thus free download. Therefore, In this case of content access with an IPR Model with some restriction, the CP may decide to select the Europeana.right only among the possible values:

- Europeana: Unknown copyright status (Unknown)
- Europeana: Rights Reserved - Restricted Access (Rights Reserved - Restricted Access)
- Europeana: Paid Access-Rights Reserved

**C. Classification of Users vs Permissions**

In the IPR Model formalization users’ profile plays an important role since each set of permissions is associated with a specific category of users. The users involved in the IPR management can be users registered to the ECLAP or not (e.g., **public users**). Each registered user may have additional roles: each role can have a set of IPR permissions associated to it. The main categories of users are: Trusted users (a restricted number of users authorized to have a high level of control, they are the administrators), CPGroup Members (registered users on the portal which are also subscribed to a specific group by which a certain content is distributed, associated with), EduRes, Education and Research users of the CPGroup (users that present on their profile and have been recognized to be students, professors, researchers, etc.), Public (users that are not logged into the portal, and may be not registered), etc. These users’ categories provide increasing restrictions in terms of content access. It should be noted that the **IPR Managers are Trusted Users** with the specific task on IPR. For example: if an **IPR Manager** assigns a permission to a Public User, the system has to automatically associate the same permissions to all the registered users (**Group Users and Group and Educational Users**). Note that the **Trusted Users** always have all the permissions.

To complete the view, an example of IPR Model for DISIT CP for distributing Videos and Audio in Italy and Germany is reported:

```plaintext
(Distrib1, DISIT, http://www.eclap.eu/3010, “for distributing in Italy and Germany”,
  http://www.europeana.eu/rights/rr-r/,
  ( Italy, 
    (Video, playMD, EduRes, null), (Video, PlayHD, CPGroup, {}),
    (Audio, play, EduRes, {} ) ),
  ( Germany, 
    (Video, downloadHD-PC, CPGroup, {FROMTO, 2012, 2013}) ,
    (Audio, download-PC, CPGroup, {} ) )
)
```
As a conclusion, a single IPR Model is capable to formalize permissions for different kind of resources, users and countries. It is a change of focus with respect to the licensing models such as MPEG-21 and ODRL, which are based on formalizing the single rights for a specific person. In those models, the main generalization is the concept of domain, by which the specific right for a specific content is associated to a group of users or devices. The proposed model presents a clear advantage for cultural heritage collections and business users, since they have to mimic for each collection the contract they have. In fact, in content collections as well as on artistic works, it is quite frequent to have the need to formalize a unique contract with rights covering a large set of different resources and media, or even treating all the media as works/products of the creative activities. Thus a unique IPR Model is the formalization of the permissions for the content collection, addressing in a single model all content kinds, resolutions, users, etc.; thus opening the space to add more content and data in the same collection as well.

D. Relationships among right permissions and users’ kinds

According to the MPEG-21 ontology (as well as to other ontologies among rights), to grant a certain right/permission to a user also imply to grant eventual specialized rights of the granted right. For example, providing the rights of video editing also implies the right of creating copy of the same video. This also means that has no sense to allow the WFIPR Manager to formalize all rights in a matrix for each user, for each content type, etc. It would be a heavy and un-useful work with the high risk of creating inconsistencies among rights. To avoid this problem, the IPR Wizard provides the automatic propagation of permissions according to the reference ontology. Moreover, not all the rights reported in the standards can be enforced and thus controlled into a web and mobile distribution. This fact can be clearly stated to the IPR Manager during the creation of a new IPR Model. For example, on web browsers for PC it is almost impossible to protect the copy of an Image. The users may use a screen capture and or simply use the browser facilities to download the image. So that, for images the rights under control can be limited and specialized.

In order to solve the above mentioned problems, a reference IPR Logical Model has been adopted. It is based on the following two main aspects:

- relationships among user kinds;
- relationships among permissions.

For example, the set of relationships among video permissions can be formalized as follows and hold for any kind of users:

- Download-HD-PC → Play-HD-PC
- Download-HD-PC → Download-LD/MD-PC
- Play-HD-PC → Play-LD/MD-PC
- Download-LD/MD-PC → Play-LD/MD-PC
- Download-LD/MD-PC → Download-mobile-browser
- Download-mobile-browser → Download-LD/MD-PC
- Play-LD/MD-PC → Play-mobile-browser
- Play-mobile-browser → Play-LD/MD-PC
- Download-mobile-browser → Play-mobile-browser
- Download-mobile-browser → Download-mobile-app
- Play-mobile-browser → Play-mobile-app
- Download-mobile-app → Play-mobile-app

Similar relationships (with different rights set) have been identified and formalized for Audio, Documents (PDF, Slides, PPT, etc.), Images, crossmedia (MPEG-21, SMIL, HTML, ePub, etc.) and Animations.
The relations among possible permissions are represented as different kinds of relationships, for example:

- **Direct implication**, for example between *Download-HD-PC → Play-HD-PC*: if a CP allows a user to download a content, the CP implicitly allows also to play it (play via streaming and/or progressive download). This because, from a technical point of view: if someone downloads a content (without encryptions or protection) from the web he can play/view it on his PC whenever he wants.

- **Reciprocal implication**, for example between *Download-LD/MD-PC ↔ Download-mobile-browser* if a Content Partner allows an ECLAP user to download a content from PC, implicitly allows him also to download the same content from a mobile device. This because the users can download a content via a browser in their PC, then transfer the content into a mobile device, so that the application of a restriction to avoid the download via mobile can be easily moved around and has no sense to be applied. It is also true the vice-versa.

```
Figure 4 – Examples of how few clicks (Yes) can be used to define the IPR Model exploiting IPR relationships. In theExample1, a single Yes (grant of permission to download on PC in HD) has automatically granted all the other lower values permissions; Example 2: the grant of Play on PC at LD and MD implied also the play on mobile, but not the download of the same digital resource.
```

The IPR Logical Model also holds a set of relationships among permissions granted to the users. In particular, any permission granted to kind of user is also granted to less restrictive users. This model is valid for all the permissions except for the permission of Embed. Therefore,

$$\forall p \in \text{Permission and } p \neq \text{Embed} \mid (p, \text{Trusted user}), (p, \text{CPGroup user}) \rightarrow (p, \text{EduRes user}), (p, \text{Public user}) \rightarrow (p, \text{CPGroup user})$$

The permission of Embed implies the usage of a provided content in a web page of third party. This technique implies that once provided the permission of content Embedding, the content may be accessed by public users on third party portals. Therefore, this permission can be formalized as:

$$\text{Embed} \rightarrow (\text{Play-LD/MD-PC}, \text{Public user})$$

This implies that on granted the permission of embed a number of permissions have also to be granted as described in Figure 5 and according to the above presented logical rule of the IPR Logical Model.
To accelerate the production of the IPR Models, the IPR Wizard tool has been realized. This tool is grounded on the IPR Logical Model that takes into account implications and relationships among user kind and permission, and the relationships among permissions and Europeana.rights modelling. This is a strong simplification with respect to the other solutions provide for the Content Aggregators that have to be very careful in selecting coherent rights and Europeana.rights values.

An IPR Model can be associated with a content manually from the interface of the ECLAP workflow or automatically. If the legal framework adopted by the community of CPs is going to change, this implies a change at level of IPR Logical Model, that also can be changed once for all content.

Please note that the saved IPR Model consists only on the assertions reported in the above examples (the yes for each kind of users as depicted in Figure 5), while all the other permissions are inferred by the IPR Logic Model regarding permissions and users relationships. This activity is performed by both the IPR Wizard and by the Cross Media Player and CAS on the ECLAP Front End, the so called phase C of Section II.

V. REPORT ON TOOLS USAGE

In this section, some results about the ECLAP Back Office activities performed on the content, metadata and IPR until April 2013 are reported. The ECLAP service allowed users and automated workflow processes to interoperate securely and to increase the quality and accessibility of content and metadata, avoiding conflicts and each other. It is currently in use by 31 institutions. The number of state transitions and their distribution in the time period put in evidence the whole activity of the portal on content and metadata and allow analyzing singularly both the back-office and the user activities. Some results are reported in the temporal domain considering the “month” as a time period unit.

Actually, there are 29 workflow qualified users. Each user may have single or multiple user roles (grant authorization). The workflow user roles are distributed as: 24 enrichers (WFENRICHER), 6 validators (WFVALIDATOR), 23 IPR users (WFIPR) and 9 publishers (WFPUBLISHER).

At current date, 1,036,406 workflow transitions have been handled for 173,562 content items with an average of 6 transitions per content and a maximum of 139 transitions for a single content. These transitions were performed in 652 days with an average of 1,589 transitions per day and a maximum of 15,786 transitions in one day, with a maximum of 14 different virtual nodes in the AXCP grid, on DISIT Cloud.
TABLE I. DISTRIBUTION OF WORKFLOW TRANSITIONS

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>Number of Transitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Uploaded'</td>
<td>'Under-AXCP'</td>
<td>259,086</td>
</tr>
<tr>
<td>'Under-AXCP'</td>
<td>'Uploaded'</td>
<td>259,086</td>
</tr>
<tr>
<td>'Proposed (creation)'</td>
<td>'Uploaded'</td>
<td>173,631</td>
</tr>
<tr>
<td>'Uploaded'</td>
<td>'Under-Approval'</td>
<td>170,609</td>
</tr>
<tr>
<td>'Under-Approval'</td>
<td>'Published'</td>
<td>170,317</td>
</tr>
<tr>
<td>'Uploaded'</td>
<td>'Under-IPR'</td>
<td>945</td>
</tr>
<tr>
<td>'Under-IPR'</td>
<td>'Uploaded'</td>
<td>945</td>
</tr>
<tr>
<td>'Uploaded'</td>
<td>'Under-Enrichment'</td>
<td>681</td>
</tr>
<tr>
<td>'Under-Enrichment'</td>
<td>'Uploaded'</td>
<td>681</td>
</tr>
<tr>
<td>'Under-Approval'</td>
<td>'Uploaded'</td>
<td>224</td>
</tr>
<tr>
<td>'Uploaded'</td>
<td>'Under-Validation'</td>
<td>43</td>
</tr>
<tr>
<td>'Under-Validation'</td>
<td>'Uploaded'</td>
<td>43</td>
</tr>
<tr>
<td>'Published'</td>
<td>'Uploaded'</td>
<td>7</td>
</tr>
</tbody>
</table>

A. Back-Office services

As previously described, the back-office services consist of a set of grid processes that run periodically automated workflow processes both on a single and on multiple contents.

1. Content and Metadata Ingestion. The number of content ingested and processed by the back-office has been 173,631 corresponding to the UPLOADED workflow status of content.

2. Metadata Analysis Metadata analysis for assessment or automated translation performs a transition to the UNDER-AXCP status in order to lock the content and avoid that a user could be access to it for manual editing or validation. In total, 259,086 of these transitions were performed.

3. Metadata Validation Every time content passed the metadata analysis the back-office performs a transition to the UNDER-APPROVAL. In total, 170,609 of these transitions were performed.

4. Content Publication Every time the back-office performs the publication of content in the UNDER-APPROVAL workflow state it performs a new transition to the final state PUBLISHED. In total, 170,317 of these transitions were performed.

B. Front-Office tools

In this section the analysis of the activity performed by users via front-office tools is reported. In the considered period, 11336 content elements have been manually uploaded by users via the upload service. Once uploaded the process has been passed to the back-office for indexing and semantic enrichment. Only the 0.2% of processes failed to conclude due to incompatible file formats. In order to evaluate the usage of Metadata Editor for the enrichment and validation activities, both the number of workflow transitions from UPLOADED to UNDER-ENRICH state and from UPLOADED to UNDER-VALIDATION status have been considered. The former transition gives a measure of manual enrichment activity, while the latter of the manual validation activity. The transitions related to enrichment were 681, and 43 for validation. Figure 6 reports the distribution in time of the enrichment activity.

ECLAP IPR Wizard has been largely used by more than 35 partners in Europe. To keep under control and evaluate the usage of IPR Wizard, the number of workflow transitions from UPLOADED to UNDER-IPR states over time have been tracked. The transitions were 945, and their distribution is reported in the Figure 7. Comparing Figures 6 and 7, it is evident that the IPR tool has been much more used than the metadata editing tool. This is due to the fact that, in most cases the content metadata where ingested by stable institutional databases and archives, while the IPR Model was missing on content; and Europeana constrained them to formalized the IPR aspects before the content submission.
For the IPR aspects, 67 different IPR Models have been used. 40 of them are restrictive not public models, while 27 are public models allowing the full content access. Most of CPs used only 1, 2 or 3 different IPR Models for their content, while a few partners used 4, 8 or 12 IPR Models. Figure 8 reports the number of files used per IPR Model. It is evident the huge advantage of coping with a unique model for unlimited number of content and unlimited number of users. This is a radical cost reduction with respect to DRM computational costs of licensing [Bellini, Nesi, Pazzaglia, 2013]. According to the analysis, it can be seen that the most diffuse IPR Models covered more than the $\frac{1}{2}$ of the whole content collected on ECLAP. On the other hand, the semantic flexibility of the IPR Model proposed allowed to cope with the many different needs of the content owners in the cultural heritage domain that impose specific IPR Models according to legal rights they can really provide, and they identify during phase of right clearance (Section II, phase A).

The 68% of content is associated with a public IPR model. Regarding the 30 restrictive IPR Models defined by CPs, in 11 cases the IPR Model granted the access only to educational group users and in 6 cases only to group
users (both educational and not educational). Moreover, 23 IPR Models have been used to grant play streaming without enabling the download of the digital resource for some kind of resource type (regardless of the user type).

<table>
<thead>
<tr>
<th>Permission</th>
<th>UserKind</th>
<th>Public</th>
<th>CPGroup</th>
<th>EduRes</th>
</tr>
</thead>
<tbody>
<tr>
<td>only play/access</td>
<td>Public</td>
<td>11</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>download &amp; play</td>
<td>CPGroup</td>
<td>3</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>no permission</td>
<td>EduRes</td>
<td>19</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

Table II reports, for three user types (public, group and educational), how many IPR Models allow only play/access of the digital resource or allow the download and play of it or no permission are provided. It can be seen that in most cases the models are used to restrict access from the public users only (19 over 30) and to limit the download of the resource, leaving the permission of play/streaming.

In order to evaluate the usage of Content Management tool for publication activity, a good measure can be the number of workflow transitions from UNDER-APPROVAL to PUBLISHED state made by CPs. The transitions were 6434 and distributed by month as reported in Figure 9.

![Figure 9 – Distribution of publication activities over time.](image)

VI. CONCLUSIONS

The paper discussed about requirements, design and validation of ECLAP which is an institutional aggregator for metadata and content coping with IPR models for providing metadata towards Europeana, the European international aggregator. The proposed solution takes into account issues connected to the cultural heritage cross media content and integrates front office tools and an automated back-office based on a grid. The solution allows to cope with high quality and provides large scale multimedia services to manage huge amount of content and metadata, coming in turn from several national and local institutions: museum, archives, content providers. The first contribution presented in this article corresponds to the identification of the Content Aggregator requirements and needs, and thus the definition and realization of a distributed architecture and workflow solution satisfying them. The aggregators have to cope with huge amount of content and metadata, and perform their massive processing for semantic enrichment and publication. To this end, a specific Workflow Model for content aggregation and enrichment management is needed to integrate both human and automated back office activities for semantic computing such as: semantic enrichment (annotation, tag, comments, skos-fication, etc.), content adaptations, metadata quality estimation, validation, translation, etc. A second contribution of this paper consists of the formalization of IPR Model and solution that allow shortening the activities of IPR resolution and avoiding the assignment of conflicting rights/permissions. The proposed IPR Mode presents computational advantages with respect to classical licensing model of DRM and CAS. In addition, the flexibility of the IPR Model proposed allowed to cope with the many different needs of the content owners in the cultural heritage domain that impose
specific IPR Models according to legal rights they can really provide, and they identify during phase of right clearance.

Finally, the usage analysis puts in evidence the whole activities of ECLAP on content, metadata and IPR until July 2013. It underlines that the activity on content and metadata aggregation, analysis and validation to match the Europeana requirements has been mainly automated and performed by the back-office, thus allowing to keep content processing cheap and sustainable. Regarding the front office side, the most used tools by content providers have been associated with IPR, namely IPR Wizard and the Content Management since they allow users to finalize the rights and to provide a connection of the content versus Europeana. Most of the metadata provided were already in a good shape and less than the 1% of content has been corrected from that point of view. On the other hand, the IPR details requested by Europeana constrained the content provider to associate to the 100% of the content a new IPR model. This huge effort has been kept under control by exploiting the IPR Model, and applying only 67 models to the whole set of more than 170.000 different content coming from more than 35 different collections and institutions.

ACKNOWLEDGMENT

The authors want to thank all the partners involved in ECLAP, and the European Commission for funding ECLAP in the Theme CIP-ICT-PSP.2009.2.2, Grant Agreement No. 250481.

REFERENCES


[CC] Creative Commons, http://creativecommons.org


