RAPPORT ET RECOMMANDATIONS

SYNTHESE

Expérimentation
Agrégation de métadonnées VOD

Transparency pour l'Hadopi

1 Objectif et méthodologie

1.1 Rappel des objectifs de l'expérimentation

Dans le cadre de sa mission d'encouragement au développement de l'offre légale, l'Hadopi souhaite décupler l'accès aux plateformes de VOD en agrégeant et uniformisant les métadonnées des catalogues audiovisuels disponibles en ligne et en les ouvrant aux tiers via la Licence Ouverte d'Etalab.

Afin de progresser de façon pragmatique, l'Hadopi s'appuie dans un premier temps sur la coopération de sept plateformes pour mettre en place une expérimentation de l'agrégation de leurs catalogues de métadonnées VOD et SVOD (Carlotta VOD, Imineo, Vodeo, Lovemyvod[1], Jook Video, Arte Boutique et Universciné).

L'expérimentation a été menée autour de deux axes :

- **la production d'un fichier** agrégeant les métadonnées des plateformes participantes et prêt à être réutilisé par des tiers ;
- **la livraison d'un rapport et de recommandations** sur la poursuite voire l'enrichissement de l'expérimentation – dont le présent document livre une synthèse.

1.2 Périmètre des métadonnées

À la suite des entretiens menés avec chacune des plateformes, une liste de 21 champs, listée ci-dessous, inspirée du schéma Dublin Core a été dressée :

- Title
- Original Title
- Director
- Cast
- Keywords
- Plot
- Genre
- Production country
- Publication date
- Duration
- Image quality
- Parental control
- Language
- Original language
- Subtitle
- Plateform
- Id
- Format

[1] Lovemyvod souhaite participer à l'expérimentation et s'est exprimée sur sa collaboration, mais son catalogue n'a pas pu être intégré à cette première version.

- URL
- Photo
- Rights

Le périmètre des 21 champs de métadonnées partagés a été inspiré du Dublin Core et de la richesse des catalogues participants ; il a également tenu compte des réserves éventuelles des plateformes.

Précisions méthodologiques :

- Lors de la production du fichier, le choix a été fait de ne pas intervenir sur le matériau livré (les métadonnées) afin de ne pas induire de choix structurant ni de dégrader les données.[2]
- Les plateformes participantes ne renseignent pas toutes les champs listés. Néanmoins la publication de cette liste permettra sans doute un enrichissement des métadonnées de leur part.
- Chaque plateforme remplit les données de son catalogue sur un tableau avec des libellés propres. C'est pourquoi il a été nécessaire de créer un tableau de correspondance, afin de regrouper l'ensemble des catalogues au sein d'un même fichier.

2 Contraintes de l'expérimentation

2.1 Choix juridique

Le projet actuel, premier pas vers un moteur de recherche : le projet d'ouverture des données est compris par les participants mais néanmoins considéré comme d'un intérêt relatif – au regard notamment d'un moteur de recherche par œuvre qui puisse améliorer le trafic vers les offres.

Le choix d'ouvrir ces données sous la Licence Ouverte d'Etalab autorise tout tiers à les exploiter y compris commercialement. Dès lors, l'ambition du projet en constitue également sa limite : certains champs ou type d'information ne seront pas partagés ou pas de façon unanime (prix, fenêtres d'exploitation, etc.).

2.2 Contraintes techniques

Absence de normalisation des métadonnées : chaque plateforme investit dans la constitution et la structuration propre de ses métadonnées. Il existe des ponts vers certaines sources extérieures mais aucune normalisation ni aucun code pivot d'identification ne sont utilisés.

Production d'un format unique : déduit du souhait de produire un fichier compatible avec la structure du Dublin Core, il a été décidé de produire un fichier en format XML et non CSV comme initialement prévu. Le XML est plus souple et offre une profondeur plus riche dans la structure de la base, il est également plus fluide pour des développeurs et couramment utilisé.

Limites à l'implication des plateformes : chaque plateforme s'est montrée très positive sur le projet et accepte de faire un pas pour y participer (paramétrage d'un fichier spécifique, modification des fichiers déposés pour intégrer les bons champs de données, envoi par email, etc.) mais il ne serait pas réaliste de prédéfinir un cadre auquel elles devraient se conformer.

La règle de non intervention sur les valeurs agrégées a plusieurs conséquences qui fragilisent le fichier dans sa première version mais pourront nourrir des choix ultérieurs :

- incohérences (ex : « 16/6 » et « moyen » dans le format)
- doublons (ex : « documentaire documentaire »)
- erreurs probables (dues aux modalités d'export interne)

[2] Les seules exceptions visent à une propreté et à une lisibilité du fichier ; elles concernent la suppression des balises web dans le corps des textes et la fragmentation de certains champs pour isoler les personnes.

- absence de standardisation (ex : « 27/08/2009 » et « 2013 » dans l'année)

3 Limites de cette « version beta » et améliorations

L'expérimentation concrétise un projet d'agrégation de catalogues de métadonnées VOD. Comme toute première version, cette expérimentation met en lumière des fragilités et questionne sur les choix à opérer pour une poursuite pertinente de l'expérimentation.

Le cœur du projet actuel réside dans la prise en compte cumulative des éléments suivants :

3.1 Périmètre des métadonnées concernées

Les 21 champs restent pertinents dans la durée mais pourront être enrichis de nouveaux champs (prix, disponibilité par support, festival, etc.), complétés ou non par les plateformes, nécessitant éventuellement un retraitement selon les exports de catalogues reçus.

3.2 Les modalités d'accès aux fichiers et la fréquence des mises à jour

Les flux de circulation de métadonnées partent des plateformes (fichiers entrants) vers l'agrégation, pour ensuite être exportés sous un unique fichier (fichier sortant). Les prochains développements pourraient prévoir une mise à jour automatisée du fichier agrégé. Dans cette hypothèse, les développements complémentaires à la charge des plateformes (exports spécifiques, adaptation au schéma retenu) se doivent d'être réduits voire nuls, afin de ne pas pénaliser certains participants.

3.3 Formats de fichiers

La priorité reste de s'adapter aux plateformes mais dans la mesure du possible, la simplicité recommanderait de conserver les types de formats entrants courants (CSV et XML). La mise en place d'un format intermédiaire servant de réceptacle aux différents fichiers offrirait une souplesse et un gain de temps évidents au traitement des fichiers entrant. Agissant comme un convertisseur de format, il permettrait également de produire plus d'un format de fichier sortant, notamment les CSV et JSON en plus du XML, les trois types de format ouverts les plus couramment utilisés.

3.4 L'intégration de nouvelles plateformes

Le chemin d'intégration technique d'une nouvelle plateforme équivaut au simple ajout d'un processus de récupération et de traitement des métadonnées, ainsi que leur mise à jour. Les volumes de métadonnées représentant l'offre de VOD restent légers, l'ajout d'un nouveau catalogue ne semble pas avoir d'impact sur la gestion de ces flux.

Néanmoins, si cet enjeu cadre l'expérimentation, sa principale limite réside dans le fait de ne pas apporter de structuration.

La mise à jour du fichier actuel telle que décrite ci-dessus vise à conserver la cohérence du projet et à l'ouvrir au maximum à de nouvelles plateformes participantes. Néanmoins, cette option de poursuite ne lève pas la principale limite du projet : le fichier accumulera les catalogues mais ne les consolidera pas.

4 Autre option de poursuite de l'expérimentation : consolider les catalogues de métadonnées

Une vision plus approfondie de l'expérimentation consisterait à dé-doublonner l'ensemble des catalogues afin de fournir la base d'un moteur de recherche par œuvre et d'éviter que des tiers la conçoivent selon leurs propres règles. Cette option tend vers une structuration et une normalisation des catalogues. Elle nécessiterait d'intervenir sur les points suivants :

4.1 Homogénéiser les données

Cette étape porte sur une standardisation de forme des métadonnées ; à la fois sur le concept attendu comme valeur (afin de lever toute ambigüité - ex : genre vs. thématique vs. catégorie) et

sur la forme des valeurs, afin de se conformer aux vocabulaires contrôlés de dénomination (ex : RFC-4646 pour les langues, ISO 8601 pour les dates, etc.)

4.2 Consolider les données

Consolider les données par titre de film permettrait d'obtenir une ligne par œuvre. En l'absence actuelle de normalisation, cela implique :

- une rigueur dans l'identification des œuvres (titre+réalisateur+acteur+date+pays ne suffisent parfois pas)
- une structuration spécifique de certains champs qui resteront propres à chaque plateforme (ex : Mots clés, Description, Qualité image, Format, etc.)

4.3 Normalisation et échanges des données

Utiliser les nomenclatures existantes dans le fichier ouvert permet d'utiliser un code pivot international. Les livrer aux plateformes a également pu être considéré comme un atout motivant (ISAN pour les titres, ISNI pour les personnes ou organisations, voire le RPCA du CNC). Une fois les données homogénéisées, des mises à jour hebdomadaires du fichier ouvert seraient souhaitables.

5 Conclusion

La production du fichier XML livré est perçue par les plateformes participantes comme la première pierre d'un projet concret dont les enjeux sont forts et structurants. Ce fichier comporte néanmoins les limites évoquées qui sont intrinsèques au secteur (données non homogénéisées, non normées, non consolidées, 7 plateformes participantes, etc.).

La pertinence de cette expérimentation sera renforcée aux yeux des plateformes par la mise en œuvre rapide des prochaines étapes, souhaitée par l'ensemble des participants.

6 ANNEXE

6.1 Champs de métadonnées intégrés à l'expérimentation

Title (Titre) : Une seule remarque sur ce champ, en forme d'alerte : un même titre peut recouvrir plusieurs films. Si l'on veut consolider les champs de titres, il faudra utiliser un code pivot.

Original Title (Titre VO) : L'importance de ce champ selon être très liée au type de catalogue (Cinéma vs. Documentaire par exemple)

Director (Réalisateur) : Aucune remarque spécifique sur les réalisateurs. Néanmoins certains renseignent aussi d'autres créateurs.

Cast (Acteurs) : Ils sont parfois concaténés dans un champ « cast » avec les réalisateurs, producteurs, etc. Ils font référence aux « intervenants » dans le répertoire du documentaire.

Keywords (Mots clés) : Certaines plateformes n'utilisent pas ce champ. Lorsque c'est le cas, il est créé en interne.

Plot (Description) : Ce champ est très différenciant car il est éditorialisé par chacun pour des raisons de positionnement (artistique, pédagogique…) et de référencement ; son ouverture peu faire débat.

Genre (Genre) : Ce champ couvre plusieurs concepts (« genre », « catégorie », « thème »). Il est également différenciant pour les offres

Production country (Pays de production) : Le pays de production semble être une information interprétée sur la base des principaux (co)producteurs.

Publication date (Date de production) : Ce champ est d'une importance particulière pour certains types de catalogues (patrimoine).

Duration (Durée) : La durée n'aura de sens que dans le pays de sortie.

Image quality (Qualité image) : Ce champ n'est pas toujours renseigné et recouvre plusieurs types de valeurs. C'est une métadonnée de disponibilité qui apportera une différenciation forte.

Parental control (Contrôle parental)

Language, Original Language and Subtitle (Langue, Langue VO et Langue Sous-Titres) : C'est une obligation de sous-titrer en français. Plusieurs parlent de « Piste audio », la plupart ne disposent que de sous-titres en français. Quoiqu'il en soit, la valeur renseignée n'est pas normée. Réintégrer les sous-titres dans les champs pourrait apporter une richesse.

Plateform (Plateforme VOD) : Ce champ fait référence à la dénomination de la plateforme source des données.

Id (Identifiant interne plateforme VOD) : Certaines n'en utilisent pas. Quasiment aucune ne rentre d'ID externe (type ISAN, IMDB…). Lorsque cette information n'est pas publique, certaines plateformes ont émis des réserves à la partager.

Format (Format) : Lorsqu'il est renseigné, ce champ renseigne des valeurs très disparates voire incohérences (16/9, mp4, COURT).

URL (Url fiche vidéo) : Ce champ est indispensable pour drainer vers l'offre.

Photo (Url visuel) : Certaines plateformes ne stockent pas l'url de l'image, d'autres alertent sur le fait qu'elles disposent d'url uniques sur la photo ; toute réutilisation par des développeurs viendrait donc « assommer les serveurs », rendant l'expérience contreproductive, ils préfèrent alors partager des url externes. Plusieurs alertent sur les risques juridiques liés à la création des visuels. Quoiqu'il en soit, il est impensable de mettre en avant un film sans visuel.

Rights (Droits) : L'information n'est pas souvent renseignée.