David ZNATY

MASTER IN SCIENCE OF MANAGEMENT
MASSACHUSETS INSTITUTE OF TECHNOLOGY

CHARGE DE COURS A L'ECOLE CENTRALE DE PARIS
PROFESSEUR ASSOCIE, PANTHEON - ASSAS
EXPERT AGREE PAR LA COUR D'APPEL DE PARIS
EN INFORMATIQUE ET TECHNIQUES ASSOCIEES

EXPERT AGREE PAR LA COUR DE CASSATION
PRESIDENT D'HONNEUR DE LA COMPAGNIE DES EXPERTS
AGREES PAR LA COUR DE CASSATION

2 Bis, Avenue de Ségur
75007 PARIS

TEL. : 01 40 65 04 06
FAX. : 01 45 55 84 91
Email : dznaty@alum.mit.edu

Mission d'expertise HADOPI
Décision du 10 juin 2011

RAPPORT D'EXPERTISE

MONSIEUR DAVID ZNATY

A

HADOPI

PARIS, LE 16 FEVRIER 2012

SOMMAIRE

1. PREAMBULE Page 4

2. INTRODUCTION Page 4

3. RAPPEL DE LA MISSION Page 5 - 6

4. REUNIONS D'EXPERTISE Page 7 - 17

5. REPONSE A LA MISSION Page 18 - 27

6. CONCLUSION Page 28 - 29

ANNEXE 1 *

Liste des Documents Reçus

ANNEXE 2 *

Documents TMG – FPM – FPA

ANNEXE 2.1
CEI & Version 1.0 – Présentation du Système CEI

ANNEXE 2.2
CEI & Version 1.2 – Spécification des Nodes de collecte

ANNEXE 2.3
Algorithmes Fingerprints

ANNEXE 2.3.1
FPM

ANNEXE 2.3.2
FPA

ANNEXE 3 *

Documents SACEM-SDRM, SCPP, SPPF

ANNEXE 3.1
Annexe technique au Contrat n° 1200910

ANNEXE 3.2
Document de recettes SACEM-SDRM, SCPP, SPPF
avec KANTAR MEDIA et TMG

ANNEXE 3.3
Copies d'écrans EXTRANET TMG

ANNEXE 3.4
Captures d'écrans agents assermentés

ANNEXE 4 *

Cahier des charges KANTAR MEDIA

ANNEXE 5 *

Contrôle de la base de référence

ANNEXE 6 *

Dépôt des codes TMG

ANNEXE 7 *

Tests à ajouter dans la base de référence

ANNEXE 8 *

Liste des personnes rencontrées

** Documents strictement confidentiels placés sous scellés par mes soins et sont conservés dans un coffre fort par l'Hadopi*

1. PREAMBULE

Cette mission n'a pu être accomplie que sous réserve de la stricte confidentialité des documents communiqués par les entités listées au paragraphe 3.1 du présent rapport.

2. INTRODUCTION

La Haute Autorité pour la diffusion des œuvres et la protection des droits sur internet (Hadopi) m'a consulté aux fins de réaliser une mission d'expertise sur le système de traitement automatisé mis en œuvre par les sociétés et associations de perception et de répartition des droits saisissant l'Hadopi pour rechercher les mises à disposition d'œuvres protégées par un droit d'auteur sur les réseaux pair à pair et collecter les adresses IP concernées.

Les ayants droit ont mis en place un système de traitement ayant pour finalité la constatation de faits de contrefaçon commis sur des réseaux pair à pair (PaP) et la collecte des adresses IP, à partir desquelles ces faits ont été commis, qui fonctionne de la manière suivante :

- Le système de traitement calcule pour chaque œuvre choisie une empreinte unique et identifie les fichiers illicites dont le contenu correspond aux œuvres originales,
- Le système recherche ensuite les fichiers illicites identifiés en faisant des requêtes sur des réseaux pair à pair (PaP) et enregistre les adresses IP des utilisateurs ayant mis le fichier à disposition,
- Les agents assermentés des ayants droit valident ces constatations et signent les saisines transmises à la commission de protection des droits.

Dans ce contexte, l'objet de la présente mission d'expertise consiste à déterminer si le mode opératoire utilisé permet l'identification sans équivoque d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre.

Le 23 mai 2011, j'ai signé un engagement de confidentialité avec l'Hadopi.

Le 10 juin 2011 j'ai été confirmé pour traiter de la mission.

3. RAPPEL DE LA MISSION

3.1 Analyser avec précision la méthode utilisée pour créer l'empreinte numérique d'une œuvre :

- Vérifier que cette méthode permet la création d'une empreinte unique de l'œuvre,
- Calculer la probabilité pour deux œuvres différentes de donner lieu à la création d'une même empreinte.

3.2 Analyser la méthode utilisée pour comparer les œuvres mises à disposition et les empreintes :

- Déterminer si les critères de comparaison utilisés entre deux empreintes sont suffisants pour authentifier une même œuvre,
- Evaluer dans quelle mesure le processus de comparaison des œuvres ne génère pas de faux-positifs.

3.3 Analyser le processus de collecte des adresses IP :

- Déterminer si le processus de collecte des adresses IP permet d'attester que les adresses IP enregistrées mettent effectivement à disposition les œuvres visées dans le procès verbal de constatation,
- Evaluer dans quelle mesure le processus de collecte protège contre les usurpations d'IP.

3.4 Analyser les processus automatiques et/ou semi automatiques et/ou manuels entre les différents acteurs (ayant droit, collecte des données IP, agents assermentés des ayants droit) qui entrent dans le mode opératoire d'identification d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre.

L'Expert pourra se faire assister d'un sachant ayant au préalable signé un accord de confidentialité et pourra utiliser tous moyens matériels pour accomplir sa mission ci-dessus,

Le délai d'exécution de la mission n'a pas été respecté du fait des agendas des différents acteurs.

4. REUNIONS D'EXPERTISE

4.1 Réunion du 23 mai 2011 chez HADOPI

Cette réunion avait pour objet de communiquer à l'Expert le projet de mission et la liste des contacts :

- Les Ayants- droit qui saisissent l'Hadopi :

SCPP : Société Civile des Producteurs Phonographiques

SPPF : Société Civile des Producteurs Phonographiques de France

ALPA : Association de Lutte contre la Piraterie Audiovisuelle.

SACEM / SDRM

- Leurs prestataires :

  TMG, Trident Media gard, qui gère la plateforme informatique faisant objet de la présente expertise

KANTAR MEDIA, qui transmet les masters des œuvres phonographiques, il s'agit des supports qui contiennent les métadonnées et le fichier de l'œuvre intégrale

FPM et FPA : Pour des raisons de confidentialité nous ne révèlerons pas quels sont les fournisseurs de solution de calcul d'empreintes. Dans la suite du document nous appellerons le fournisseur de la solution pour le calcul d'empreinte d'œuvres musicales FPM et nous appellerons le fournisseur de la solution pour le calcul d'empreinte d'œuvres audiovisuelles FPA.

4.2 Réunion du 10 juin 2011 au Cabinet de l'Expert

Lors de cette réunion, il m'a été exposé la perception technique d'Hadopi du système (approche en 3 étapes) permettant d'aboutir au PV de saisine auquel est joint un « Chunk » (une partie du fichier représentatif de l'œuvre, dont la mise à disposition a été constatée sur un réseau PaP).

4.3 Réunion du 15 juin 2011 chez Hadopi

Lors de cette réunion nous avons poursuivi les discussions techniques ; un certain nombre de documents m'ont été remis (cf. Annexe 1) ainsi qu'une clé USB contenant 3 saisines à titre d'exemple (ALPA, SACEM, SPPF) (format XML) ; une saisine est composée de 3 fichiers XML.
A l'annexe 3.4 figure un exemple de saisine.

4.4 Réunion du 27 juin 2011 chez ALPA

Alpa n'est pas titulaire des droits, mais dispose d'une délégation de pouvoir des ayants droit pour constater les faits de contrefaçon en matière audiovisuelle et pour saisir l'Hadopi.

Lors de cette réunion nous avons visualisé les écrans utilisés par l'ALPA et développés par TMG pour les Agents assermentés qui établissent et signent les constats ainsi que les process (cf. supra). La cinématique de ces écrans a fait l'objet d'une édition avec commentaires et figure à l'annexe 3.3 et 3.4 du présent rapport.

4.5 Réunion avec TMG du 28 juin 2011 au Cabinet de l'Expert

Cette première réunion avait pour objet de préparer la mission technique et de lister les documents nécessaires à la compréhension des processus par les intervenants (FPM, FPA, Agents assermentés...).

Les acteurs :

- ALPA (Gaumont, Pathé..) pour l'audiovisuel et utilisation du système d'empreintes FPA
- SCPP
  SACEM
  SPPF
  } Pour la musique et utilisation du système d'empreintes « FPM »

4.5.1 Informations sur les œuvres Audiovisuelles

Métadonnées de l'œuvre

- Titre,
- Titre original,
- Dates de sorties (salle, DVD) ; France, USA, ...
- Ayants droit (producteurs, distributeurs, ...).

Données d'identification :

FPA a fourni à TMG sous forme de licences les logiciels de calcul de l'empreinte d'une œuvre en binaire (pas de code sources) ; FPA a aussi fourni à des laboratoires les codes de l'algorithme de calcul de l'empreinte.

En pratique, ce sont les laboratoires qui fournissent ces données après validation par les ayants droit.

Pour collecter les données, un compte FTP existe entre tous les labos (TMG ↔ Labo) ou via FPA.

TMG a développé des interfaces hommes machines (IHM) pour les ayants droit français.
TMG peut aussi se connecter pour collecter des données au format XML.

Dans certains cas, les ayants droit peuvent envoyer une BETACAM et/ou un DVD.

4.5.2 Informations sur les œuvres Musicales

C'est la société Kantar Media qui adresse les fichiers Audio ou Vidéo (clip) en format AVI pour les clips et MP3 pour la musique ; fichiers MP3 ou AVI plus les métadonnées en XML.

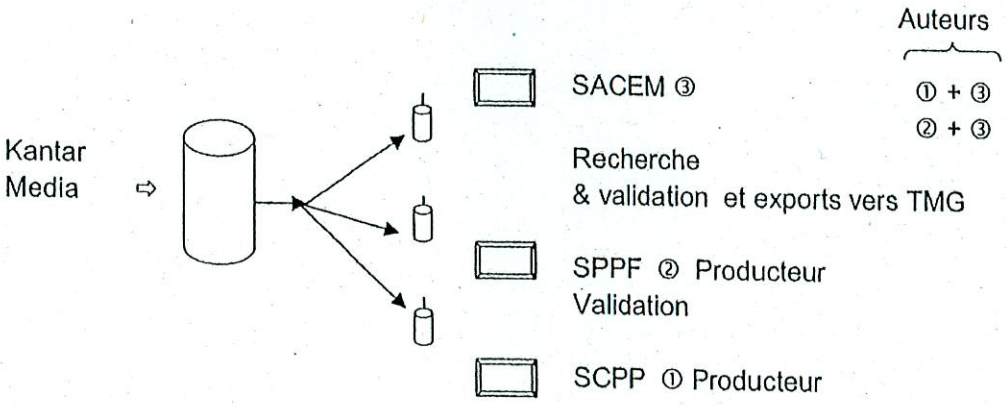
4.5.3 Rôle des agents assermentés dans l'attribution des droits (titularisation d'une œuvre)

- Ce rôle n'est joué que pour les ayants droit dans le domaine de la musique, SCPP, SACEM et SPPF ; L'ALPA n'a pas besoin de procéder à cette vérification.

- TMG reçoit les données directement de Kantar Media et adresse tous les fichiers reçus aux ayants droit (SPPF, SACEM, SCPP) qui, à travers les informations figurant dans les métadonnées adressées par Kantar, valident à travers un IHM construit par TMG les titulaires des droits.

Ex : un fichier qui contient 3 œuvres arrive de Kantar ; ces 3 œuvres seront vues par les 3 entités et chacune s'attribue les œuvres par la connaissance de ses propres ayants droit.

SCPP ⇨ Titulaire des œuvres
SPPF pour les producteurs
SACEM ⇨ Titulaire des œuvres pour les auteurs



Les Agents assermentés accèdent au système TMG via une clé (Token) sécurisée (cf. Annexes 3.3 et 3.4 du présent rapport).

4.5.4 Constitution de la base de données des œuvres qui sont sur les réseaux Pair à Pair (PaP)

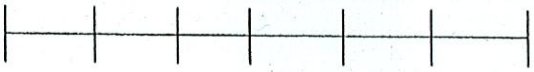
⇨ Recherche par mots clés dans chaque réseau PaP à travers les outils fournis par les protocoles.

⇨ Méthode des jeux de mots clés.

4.5.5 Méthode de comparaison des empreintes (voir réunion du 28 Septembre)

Pour calculer les empreintes sur les œuvres ayant pour origine un réseau PaP, on recherche des séquences ordonnées sur lesquelles on calcule des empreintes.

Ex : Œuvre J



E^1 E^n

$\left\{ E^J_1 \ldots E^J_w \right\}$ Si on trouve dans les fichiers/œuvres une séquence ordonnée de « n » empreintes alors il y a reconnaissance de l'œuvre (Ex : pour ALPA, ~ 35') ; pour une œuvre musicale, on prend 80 % de la durée.

A ce stade, TMG a constitué son référentiel de comparaison et collecté les œuvres reconnues sur ce principe dans les réseaux PaP (cf. Annexe 2 du présent rapport).

4.5.6 Phase de collecte des « incidents »

PHASE 1	PHASE 2
1.0 Reconnaissance de l'œuvre par le biais des empreintes + 1.1 Validation par les Agents assermentés de tout (si une œuvre n'est pas valldée elle reste en attente et ne rentre pas dans le process)	2.0 Collecte des fichiers (œuvres mises à disposition sur les réseaux PaP) sélectionnés sur la base de multicritères (métadonnées) ; le fichier est « préparé » (savoir-faire TMG) afin de pouvoir appliquer l'algorithme de génération d'empreinte. Si l'empreinte du fichier existe dans la base de référence, le système relève le hashcode du fichier aux fins qu'il soit validé une seconde fois par l'agent assermenté. 3.0 Collecte des adresses IP

4.5.7 Mode de collecte de l'adresse IP

- TMG capte l'adresse IP au moment de l'établissement de la connexion (Mode connecté).
- L'adresse est prise dans le socket et dans une même session (cf. Annexe 2.2 du présent rapport).

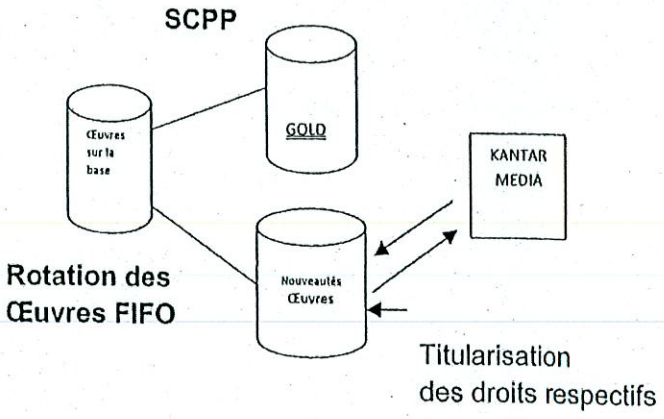
4.6 Réunion du 29 juin 2011 à la SCPP

Durant cette réunion la SCPP expose à l'Expert la façon dont est organisé le travail de ses agents assermentés chargés d'établir les constats transmis à l'Hadopi.

« … Les agents assermentés ont pour mission de constater la matérialité des atteintes aux droits, Producteurs et Auteurs.

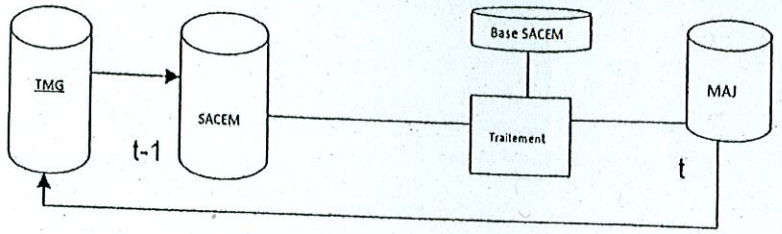
- *Les agents assermentés font du téléchargement et vérifient qu'il correspond bien aux répertoires,*
- *Plaintes sur les PV des agents assermentés,*
- *Un agent assermenté est un salarié des entités ayant reçu un agrément.*

- *Dans le cadre de notre mission générale, on surveille toutes les œuvres de nos producteurs ;*
- *Il y a une base de données qui n'est pas exhaustive ; cette base contient X œuvres... »*



Planning d'envoi des métadonnées

- En fin de matinée, la SACEM interroge les œuvres revendiquées par un de ses auteurs ; ce travail est fait dans un délai de 48 heures et le fichier est complété et envoyé.



- Après l'envoi des œuvres qui mettent à jour la base TMG, la base est à jour pour les producteurs (SCPP, SPPF) et sous 48 heures pour les auteurs (la SACEM).

⇩

- Phase de traitement TMG
- MP3 pour la musique, MPEG pour les clips et XML pour les métadonnées.

⇩

La titularisation est complète à ce stade

C'est TMG qui a, dans le cadre de son appel d'offres, proposé le système FPM ; ce système a été recetté.

- C'est TMG qui calcule l'empreinte sur les enregistrements adressés par Kantar Media.

⇩

L'ensemble du système a été recetté incluant le système FPM (les PV de recettes qui m'ont été adressés figurent à l'Annexe 3.2 du présent rapport).

4.7 Réunion du 1er juillet 2011 à la SACEM

Lors de cette réunion nous avons observé le process d'intégration des nouveautés (cf. infra)
Ce process est hebdomadaire, Kantar Media met à disposition les œuvres et les envoie à TMG.

4.8 Réunion du 8 juillet 2011 à la SCPP

Lors de cette réunion nous avons suivi le processus de validation des Agents assermentés qui n'a pas pu être fait le 29 juin ; lors de cette réunion on a pu constater les phases de titularisation, la validation des œuvres collectées sur les réseaux PaP et la validation par échantillonnage de la conformité des PV de constatation (cf. Annexes 3.3 et 3.4).

A ma demande, une cinématique commentée des process a été faite (cf. Annexe 3.4).

4.9 Réunion du 9 septembre 2011 chez Kantar Media

Lors de cette réunion j'ai pu valider les process vus antérieurement.

Kantar Média envoie les œuvres à TMG qui va générer les empreintes.

L'ensemble du process est décrit dans les documents confidentiels qui m'ont été transmis par Kantar Média (cf. Annexe 4 du présent rapport).

4.10 Réunion du 28 septembre 2011 chez TMG à Nantes

Sur demande de l'Expert, il a été procédé à l'analyse de la génération des empreintes des œuvres avec la méthode FPM (phonogramme et clip) et la méthode FPA (audiovisuel).

Définition

L'empreinte (fingerprint en anglais) d'une œuvre est un identifiant d'une œuvre musicale ou audiovisuelle. Cette empreinte est indépendante du support ou de l'encodage de l'œuvre. Un système d'empreinte est dit robuste si ce système assure l'unicité de l'empreinte de chaque œuvre.

Une note blanche FPM (article) m'a été remise dans laquelle est exposée la méthode (cf. Annexe 2).

Découpage du fichier audio en morceaux de 3 secondes et calcul du « BER » (Bit Error Rate) qui doit être inférieur à 0,35 pour FPM mais TMG l'a diminué à 0,20 pour renforcer encore la fiabilité du système.

TMG a fixé la durée maximum à 120 secondes et si le fichier adressé par Kantar Media est inférieur à 120 s, on prend 80 % de la longueur du temps.

Les informations générées par le système FPM passe par une API (interface logiciel) communiquée à TMG par FPM.

En synthèse, TMG traite la sortie FPM en ajoutant une couche de contrôle TMG (voir algorithme) (cf. Annexe 2.1).

Ce principe est le même pour le système FPA (cf. Annexe 2.2).

L'objectif du traitement pour générer les empreintes est d'éviter les négatifs et faux positifs ; une fois le fingerprint calculé, on soumet aux agents assermentés les œuvres pour validation.

Mise à jour de la base et alimentation dans la base de production d'une référence (numéro séquentiel) lié au fingerprint.

Le processus de collecte des œuvres mises à disposition sur les réseaux PaP se fait bien par recherche à partir des Métadonnées (sélection de mots clés) ; chargement de l'ensemble des fichiers (unique ZIP ou autres) et génération par une procédure spécifique à TMG des fichiers INPUT aux systèmes FPM ou FPA.

Collecte des hashcodes de ces fichiers et demande de validation par les Agents assermentés (cf. Annexe 3.4).

> **Définition**
> Le « hashcode » d'un fichier est un identifiant d'un fichier. Ce « hashcode » ne présume pas ce que représentent les données contenues dans le fichier et son calcul est uniquement basé sur la suite numérique qui constitue ce fichier. Un système de hashcode est dit robuste si ce système assure l'unicité du hashcode de chaque fichier.
> Les hashcode sont utilisés dans les protocoles PaP pour identifier un fichier.

Puis le système collecte l'adresse IP et le segment de 16 ko associé à l'œuvre mise à disposition (time stamp et position du segment dans le fichier) (cf. Annexe 2). TMG calcule son propre hashcode du segment (SHA1). Le segment est un sous ensemble du fichier complet de l'œuvre et constitue une preuve démontrable.

Spontanément et sans aucune préparation, j'ai demandé à interroger la base de référence afin de vérifier s'il y avait des doublons ; nous avons constaté 23 doublons sur la base Musique qui ont pu être expliqués par le fait que Kantar Média a transmis 2 masters (un master est le support original qui contient les métadonnées et le fichier de l'œuvre intégrale) pour la même œuvre ; le master est fourni par l'éditeur ; par la suite j'ai demandé à faire le même contrôle pour l'audiovisuel ; il n'y avait pas de doublons (cf. Annexe 5).

Ce même jour et pour éliminer tout aléa pour l'avenir, j'ai demandé à TMG, qui a accepté, d'introduire un test sur les bases des œuvres de référence de non existence d'un doublon ; et en cas d'apparition, de vérifier que l'empreinte ne concerne pas 2 œuvres différentes (faux positif) ; ce test garantit qu'un faux positif dans la base de référence serait immédiatement découvert.

Après les procédures de vérification, j'ai poursuivi ma réunion en allant constater le data center où se trouve une des plateformes TMG.

4.11 Conférence téléphonique du 10 octobre 2011 avec FPM chez L'Expert

Lors de cette conférence téléphonique, FPM a réexpliqué la méthodologie exposée dans le White Paper remis par TMG (cf. Annexe 2.2.1).

Après discussion, j'ai suggéré à FPM de tester la solidité des paramètres utilisés par TMG, à savoir : < à 0,20 pour le BER alors que le système FPM est paramétré à 0,35, auquel il faut ajouter le processus de non retenu d'une œuvre si on constate un doublon ou une quelconque erreur sur 10 % des traces communiquées par le logiciel FPM (cf. Note Fingerprint Audio/Vidéo TMG) (cf. Annexe 2.2.1)

FPM a passé sa batterie de tests sur la base des paramètres qui ont été communiqués par TMG le mercredi 12 octobre ; FPM estimant à une semaine le délai de réalisation de ce test.

Le résultat m'a été communiqué par Mail du 25 Octobre 2011 et confirme la « force » des paramètres standard FPM et ceux de TMG (cf. Supra) ce mail est confidentiel.

4.12 Réunion du 11 octobre 2011 à FPA

FPA n'a pas souhaité, tout comme FPM, exposer son savoir faire. Cependant la présentation qui a été faite et les exemples donnés, notamment pour la Télévision où le risque de faux positifs pouvait exister (il est beaucoup plus faible pour les œuvres Audiovisuelles), me permet de dire que le risque d'avoir une même empreinte pour deux œuvres différentes est quasi nul, d'autant plus que ce risque est réduit par le test qui est fait dans la base par un contrôle des doublons (cf. Annexe 2.2.2).

5. REPONSE A LA MISSION

Rappel de définitions

L'empreinte (fingerprint en anglais) d'une œuvre est un identifiant d'une œuvre musicale ou audiovisuelle. Cette empreinte est indépendante du support ou de l'encodage de l'œuvre. Un système d'empreinte est dit robuste si ce système assure l'unicité de l'empreinte de chaque œuvre.

Définition
Le « hashcode » d'un fichier est un identifiant d'un fichier. Ce « hashcode » ne présume pas ce que représentent les données contenues dans le fichier et son calcul est uniquement basé sur la suite numérique qui constitue ce fichier. Un système de hashcode est dit robuste si ce système assure l'unicité du hashcode de chaque fichier.
Les hashcode sont utilisés dans les protocoles PaP pour identifier un fichier.

Pour répondre à la mission, il fallait que je garantisse la fiabilité/robustesse :

- des méthodes de génération et de comparaison des empreintes (fingerprint) (voir 5.1 et 5.2)
  - pour la méthode utilisée pour les œuvres musicales (technologie développée par FPM)
  - pour la méthode utilisée pour les œuvres audiovisuelles (technologie développée par FPA)
- du processus de collecte des adresses IP (voir 5.3)
- de l'utilisation du système par les Agents Assermentés (voir 5.4)

Le schéma suivant résume les différentes actions effectuées par les différents acteurs du système.

3. **Constitution de la base des fichiers (et de leur Hashcode) à surveiller sur le PaP**
   - Recherche sur les réseaux PaP de fichiers contenant potentiellement des œuvres de référence. Cette recherche se fait à partir de mots clés en rapport avec les titres des œuvres de référence.
   - Validation du fait que les fichiers résultant de la recherche contiennent bien une œuvre de référence. Cette validation se fait d'abord par un système de comparaison automatique à partir des empreintes, ensuite manuellement par les agents assermentés.
   - Pour chacun des fichiers résultant de la recherche contenant bien une œuvre de référence, conservation du fichier et du Hashcode de celui-ci.
   - Le résultat est une base de fichiers dont on a la certitude qu'ils contiennent une œuvre de référence. Ces fichiers sont associés à leur Hashcode, et ce sont ces fichiers qui seront surveillés sur les réseaux PaP.
4. **Constatation de la mise à disposition des fichiers surveillés sur le PaP**
   - Collecte des adresses IP des accès à Internet à partir desquels sont été mis à disposition les fichiers surveillés sur le PaP (voir étape précédente).
   - Pour chacune des adresses IP collectées, téléchargement à partir de cette IP d'un segment du fichier mis à disposition.
   - Le résultat est un ensemble de constats de mise à disposition illicite de fichiers. Un constat contient notamment une adresse IP à partir de laquelle un segment de fichier a été mis à disposition, le segment de fichier ainsi qu'un horodatage de la mise à disposition.
5. **Génération des PVs**
   - Génération automatique des PV à partir des constats avec en pièce jointe le segment (signé) du fichier mis à disposition sur une adresse IP (ce fichier contenant une œuvre de référence) ainsi que le moment de cette mise à disposition.

Il est important de comprendre que l'empreinte ne sert qu'au traitement interne du système pour identifier une œuvre de référence (étape 2) et pour la comparer à l'empreinte qui est générée par les mêmes algorithmes sur les fichiers mis à disposition sur les réseaux PaP et qui sont sélectionnés par mots clés (étape 3).

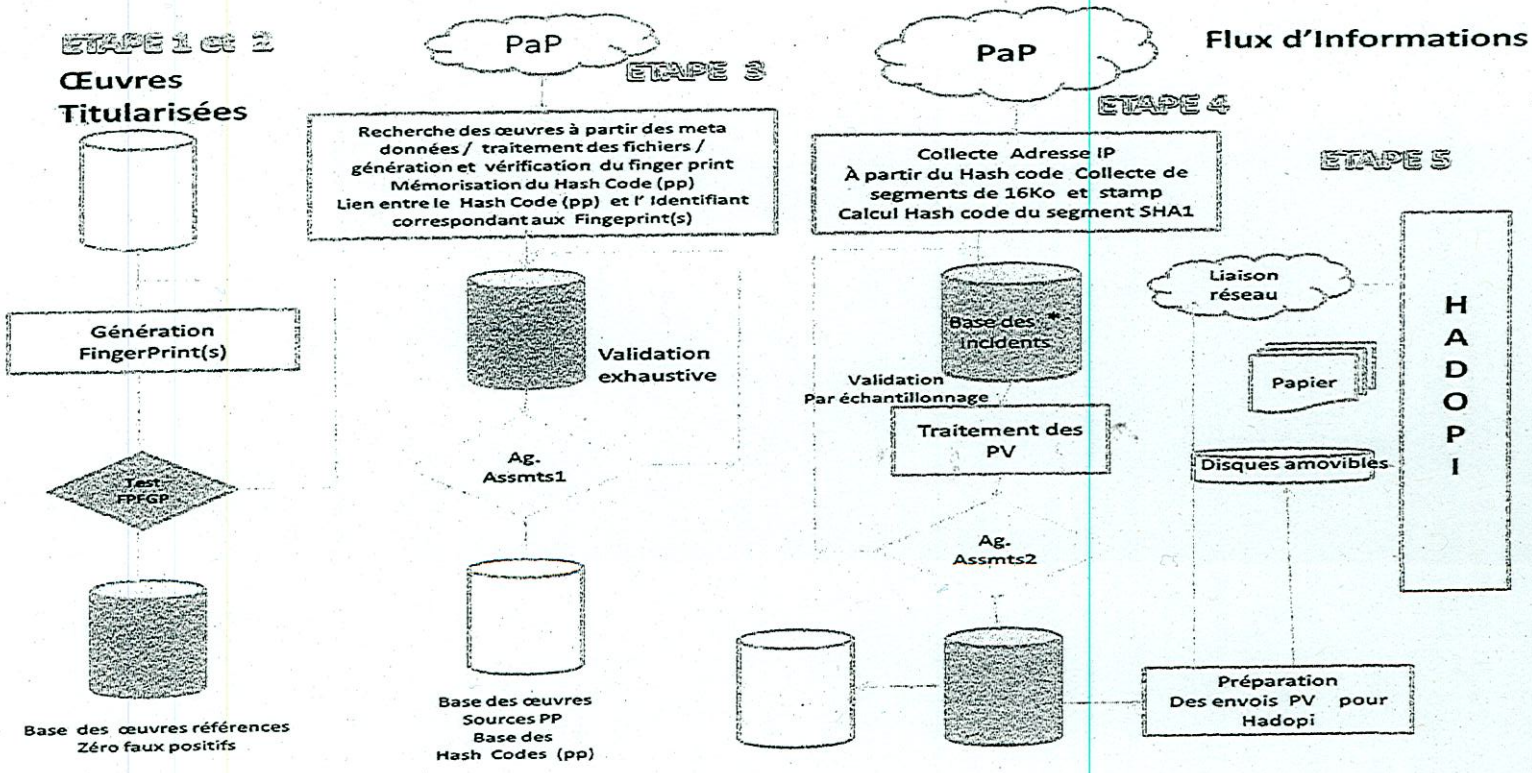
TMG a développé un savoir faire dans le traitement des fichiers mis à disposition sur les réseaux PaP pour pouvoir, quel que soit le format d'un fichier, appliquer les algorithmes de calcul d'empreintes sur le contenu de ces fichiers.

Cette précision est nécessaire pour ne pas faire de confusion entre le Hashcode du fichier et l'empreinte d'une œuvre/représentation du contenu d'un fichier.

A l'Annexe 7, figurent les PV de recettes de l'application (SACEM/SDRM, SCPP, SPPF).

A l'Annexe 8, j'ai joint la liste des personnes qui ont participé et/ou répondu à mes questions pour accomplir ma mission.

Le schéma ci-dessous synthétise le flux général du système de génération des PV de constatation.



*** Base d'incidents : éléments d'informations recueillies et qui apparaissent dans les PV.**

5.1 Analyser avec précision la méthode utilisée pour créer l'empreinte numérique d'une œuvre :

- Vérifier que cette méthode permet la création d'une empreinte unique de l'œuvre,
- Calculer la probabilité pour deux œuvres différentes de donner lieu à la création d'une même empreinte.

Pour répondre à cette partie de la mission et compte tenu du fait que FPM et FPA ont refusé de dévoiler leur savoir faire, je me suis attaché à vérifier la robustesse de leurs méthodes par les exposés qui m'ont été faits (cas pour FPA) et par un complément de test dans le cas de FPM. (cf. Annexes 2.2.1 et 2.2.2).

J'ai ensuite analysé le processus de génération des empreintes afin de voir s'il était possible d'insérer un contrôle pour garantir la non présence d'une même empreinte pour deux œuvres différentes dans la base de référence (cf. Annexe 6) (Analyse du protocole). Ce test garantit une remontée d'alerte en cas de modification des algorithmes de génération d'empreintes.

FPM

A ma demande FPM a testé les paramètres utilisés par TMG. La réponse ci-dessous confirme des paramètres TMG :

"..When we run the <FPM> test set for audio with the default settings, the technology has a false positive ratio of 0.000%. If we apply the TMG parameters -which are more stringent than the default settings- to this test set, the false positive ratio will remain 0.000%. .."

Traduction

"...Quand nous exécutons le test avec les paramètres standard <FPM>, nous obtenons un ratio de faux positifs de 0,000 % ; Si nous appliquons à ce test les paramètres TMG qui sont plus contraignants que nos paramètres standard, le ratio de faux positifs ne change pas et reste à 0,000 % ...».

J'ai validé ce résultat par un test que j'ai effectué in situ chez TMG à ma demande et sans préparation sur la base de données de références.

Suite à la mise en évidence 23 doublons d'empreintes dans la base, j'ai constaté que ces doublons correspondent tous au cas où les Ayants droit (Kantar) ont transmis 2 masters pour la même œuvre (un master est un support qui contient les métadonnées et le fichier de l'œuvre intégrale). Ces doublons n'étaient donc pas des cas de faux positifs.

Définition : Faux positif = même empreinte pour deux œuvres différentes.

Conclusion : les algorithmes de calcul d'empreintes de FPM sont robustes et garantissent la non existence de faux positifs.

FPA
Pour FPA l'approche anti faux positifs sur la conception algorithmique est la suivante :

- *L'objectif est : « aucun faux positif sur les vérités terrain » pour permettre un bon niveau d'automatisation en production*
- *En cas d'apparition de faux positif en production, le cas serait traité prioritairement en évolution de l'algorithme*
- *Des vérités terrain à grande échelle sont conçues pour tester l'absence de faux positif ; elles sont utilisées en validation de non régression :*
  3 000h réf. contre 4 jours de TV
  10 000h ref. contre 1 000h
- *Dans le cadre d'analyses spécifiques effectuées sur le filtrage par les sites de partage (en production depuis 01/2008), aucun faux positif n'a été identifié*

Les tests qui m'ont été présentés montrent la robustesse de leur algorithme.

Une recherche de doublons a été faite sur la base de référence ; il n'y a que des « fingerprints » uniques. La description de la base référence Vidéo figure en Annexe 2.2.2.

Conclusion : les algorithmes de calcul d'empreintes de FPA sont robustes et garantissent la non existence de faux positifs.

5.2 Analyser la méthode utilisée pour comparer les œuvres mises à disposition et les empreintes :

- Déterminer si les critères de comparaison utilisés entre deux empreintes sont suffisants pour authentifier une même œuvre,
- Evaluer dans quelle mesure le processus de comparaison des œuvres ne génère pas de faux-positifs.

La méthode utilisée pour comparer les œuvres mises à disposition sur les réseaux PaP et leurs empreintes est strictement identique à la génération d'empreinte des œuvres de référence.

Avant de générer les empreintes (cf. Annexe 2 ; document CEI V1.0 Fingerprint Audio/Vidéo), TMG doit « préparer » les fichiers collectés par mots clés sur les réseaux pair à pair (PaP) :

Si l'empreinte générée n'est pas trouvée dans la « base de référence », elle ne pourra pas être identifiée (pour mémoire la base de référence ne doit pas contenir d'empreintes identiques pour deux œuvres différentes, cf. supra)

Conclusion : les algorithmes de comparaison d'empreintes de FPA et les algorithmes de comparaison d'empreintes de FPM sont robustes et garantissent la non existence de faux positifs.

Aussi compte tenu de la problématique du secret, et bien que les approches de FPA et FPM soient robustes, j'ai proposé à TMG, pour éviter la présence de faux positifs dans « les bases de références », et par conséquent dans la « base source », de sécuriser le processus par l'adjonction d'un test. Ce test consiste à vérifier qu'il n'y a pas, dans la base, un même « fingerprint » pour des œuvres différentes (cf. schéma Flux Informations : Test FPFGP) afin de les détecter et les analyser en semi automatique.

TMG a accepté d'intégrer ce test dans son exploitation (cf. Annexe 7).

Ce test, une fois mis en place, garantira à l'Hadopi que dans la base de référence deux œuvres différentes ne pourront pas avoir une même empreinte sans que cela ne soit remarqué.

5.3 Analyser le processus de collecte des adresses IP :

- Déterminer si le processus de collecte des adresses IP permet d'attester que les adresses IP enregistrées mettent effectivement à disposition les œuvres visées dans le procès verbal de constatation,
- Evaluer dans quelle mesure le processus de collecte protège contre les usurpations d'IP.

Le processus de collecte des adresses IP est décrit dans la documentation qui figure en Annexe 2.2 (CEI & Version 1.2 Spécification des Nodes de collecte).

Le principe même de fonctionnement des protocoles pair à pair (PaP) fait que la connexion doit être permanente lors du téléchargement d'un segment à partir d'une adresse IP et la méthode utilisée dans le système est conforme au fait que le téléchargement est fait dans une seule session (Socket au sens BSD version UNIX).

La modification d'une adresse IP en PaP n'a pas de sens si on veut garder un téléchargement intègre.

TMG collecte aussi d'autres informations, log du début et fin de chargement du segment, ce complément d'information (horodatage) augmente la preuve dans le contexte des réseaux pair à pair (PaP).

Tous les accès à la base de données des infractions sont réalisés par des procédures stockées, de plus l'adresse IP est cryptée dans la base des incidents.

En l'état on peut dire que la méthode de capture des adresses IP avec leurs compléments d'informations (horodatage) est robuste pour toutes les attaques externes, et comme pour tous les systèmes on ne peut pas exclure l'attaque interne.

TMG, à ma demande, a produit un document justifiant le dépôt des logiciels. J'ai recommandé le dépôt des codes sources des logiciels, des exécutables associés ainsi que de l'ensemble de la documentation, et ce au fur et à mesure des modifications. La date de dépôt doit correspondre aux versions en cours d'exploitation (cf. Annexe 7).

Conclusion : L'analyse de l'architecture et de la méthodologie de collecte des adresses IP me permettent de dire que le système est cohérent et FIABLE.

5.4 Analyser les processus automatiques et/ou semi automatiques et/ou manuels entre les différents acteurs (ayant droit, collecte des données IP, agents assermentés des ayants droit) qui entrent dans le mode opératoire d'identification d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre.

J'ai analysé l'ensemble des processus automatiques et semi automatiques entre les différents acteurs. Les personnes qui ont un rôle clé dans ce processus sont les Agents Assermentés ; ils ont pour tâche de valider les œuvres de référence qui sont intégrées dans la base tout comme les œuvres reconnues sur les réseaux pair à pair (PaP) ; cette validation est exhaustive (test Assmts1).

Une préparation est faite pour structurer les fichiers venant du réseau pair à pair (PaP) pour permettre le calcul de l'empreinte des contenus de ces fichiers et ainsi permettre la comparaison avec celles de la base de référence.

Si une œuvre n'est pas validée, elle ne rentre pas dans le processus et reste indéfiniment en suspens jusqu'à sa validation par l'Agent.

Les Agents assermentés interviennent dans une deuxième validation qui est la conformité du PV, cette validation est faite par échantillonnage ; l'échantillon doit être validé dans sa totalité (test Assmts2).

J'ai pu constater le fonctionnement de ces deux tests lors de ma visite chez l'ALPA et la SCPP (cf. Annexes 3.3. et 3.4).

Conclusion : Les processus automatiques et/ou semi automatiques et/ou manuels entre les différents acteurs qui entrent dans le mode opératoire d'identification d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre est FIABLE.

6. CONCLUSION

Pour répondre positivement à la mission, il fallait que je garantisse la fiabilité/robustesse :

- de l'utilisation du système par les Agents Assermentés
- des méthodes de génération et de comparaison des empreintes (fingerprint)
    - pour la méthode utilisée pour les œuvres musicales
    - pour la méthode utilisée pour les œuvres audiovisuelles
- du processus de capture des adresses IP

En l'état :

Concernant l'utilisation du système par les Agents assermentés :

Les processus automatiques et/ou semi automatiques et/ou manuels entre les différents acteurs qui entrent dans le mode opératoire d'identification d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre est FIABLE.

Concernant les méthodes de génération et de comparaison des empreintes (fingerprint) :

Les algorithmes de calcul d'empreintes utilisés pour les œuvres musicales sont ROBUSTES et garantissent la non existence de faux positifs.

Les algorithmes de calcul d'empreintes utilisés pour les œuvres audiovisuelles sont ROBUSTES et garantissent la non existence de faux positifs.

Les algorithmes de comparaison d'empreintes utilisés pour les œuvres musicales sont ROBUSTES et garantissent la non existence de faux positifs.

Les algorithmes de comparaison d'empreintes utilisés pour les œuvres audiovisuelles sont ROBUSTES et garantissent la non existence de faux positifs.

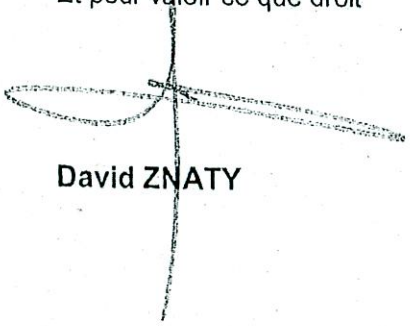
Concernant le processus de capture des adresses IP :

L'analyse de l'architecture et de la méthodologie de collecte des adresses IP me permettent de dire que le système est cohérent et FIABLE.

Conclusion : en l'état, le processus actuel autour du système TMG est FIABLE. Les documents constitués du procès verbal (saisine), et si nécessaire du fichier complet de l'œuvre (stocké chez TMG) associé au segment de 16 Ko constituent une preuve ROBUSTE.

Le mode opératoire utilisé permet donc l'identification sans équivoque d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre.

Et pour valoir ce que droit



David ZNATY

David ZNATY

MASTER IN SCIENCE OF MANAGEMENT
MASSACHUSETS INSTITUTE OF TECHNOLOGY

CHARGE DE COURS A L'ECOLE CENTRALE DE PARIS
PROFESSEUR ASSOCIE, PANTHEON - ASSAS
EXPERT AGREE PAR LA COUR D'APPEL DE PARIS
EN INFORMATIQUE ET TECHNIQUES ASSOCIEES

EXPERT AGREE PAR LA COUR DE CASSATION
PRESIDENT D'HONNEUR DE LA COMPAGNIE DES EXPERTS
AGREES PAR LA COUR DE CASSATION

2 Bis, Avenue de Ségur
75007 PARIS

TEL. : 01 40 65 04 06
FAX. : 01 45 55 84 91
Email : dznaty@alum.mit.edu

Mission d'expertise HADOPI
Décision du 10 juin 2011

RAPPORT D'EXPERTISE

MONSIEUR DAVID ZNATY

A

HADOPI

PARIS, LE 16 FEVRIER 2012

SOMMAIRE

1. PREAMBULE Page 4

2. INTRODUCTION Page 4

3. RAPPEL DE LA MISSION Page 5 - 6

4. REUNIONS D'EXPERTISE Page 7 - 17

5. REPONSE A LA MISSION Page 18 - 27

6. CONCLUSION Page 28 - 29

ANNEXE 1 *

Liste des Documents Reçus

ANNEXE 2 *

Documents TMG – FPM – FPA

ANNEXE 2.1
CEI & Version 1.0 – Présentation du Système CEI

ANNEXE 2.2
CEI & Version 1.2 – Spécification des Nodes de collecte

ANNEXE 2.3
Algorithmes Fingerprints

ANNEXE 2.3.1
FPM

ANNEXE 2.3.2
FPA

ANNEXE 3 *

Documents SACEM-SDRM, SCPP, SPPF

ANNEXE 3.1
Annexe technique au Contrat n° 1200910

ANNEXE 3.2
Document de recettes SACEM-SDRM, SCPP, SPPF
avec KANTAR MEDIA et TMG

ANNEXE 3.3
Copies d'écrans EXTRANET TMG

ANNEXE 3.4
Captures d'écrans agents assermentés

ANNEXE 4 *

Cahier des charges KANTAR MEDIA

ANNEXE 5 *

Contrôle de la base de référence

ANNEXE 6 *

Dépôt des codes TMG

ANNEXE 7 *

Tests à ajouter dans la base de référence

ANNEXE 8 *

Liste des personnes rencontrées

* *Documents strictement confidentiels placés sous scellés par mes soins et sont conservés dans un coffre fort par l'Hadopi*

1. PREAMBULE

Cette mission n'a pu être accomplie que sous réserve de la stricte confidentialité des documents communiqués par les entités listées au paragraphe 3.1 du présent rapport.

2. INTRODUCTION

La Haute Autorité pour la diffusion des œuvres et la protection des droits sur internet (Hadopi) m'a consulté aux fins de réaliser une mission d'expertise sur le système de traitement automatisé mis en œuvre par les sociétés et associations de perception et de répartition des droits saisissant l'Hadopi pour rechercher les mises à disposition d'œuvres protégées par un droit d'auteur sur les réseaux pair à pair et collecter les adresses IP concernées.

Les ayants droit ont mis en place un système de traitement ayant pour finalité la constatation de faits de contrefaçon commis sur des réseaux pair à pair (PaP) et la collecte des adresses IP, à partir desquelles ces faits ont été commis, qui fonctionne de la manière suivante :

- Le système de traitement calcule pour chaque œuvre choisie une empreinte unique et identifie les fichiers illicites dont le contenu correspond aux œuvres originales,
- Le système recherche ensuite les fichiers illicites identifiés en faisant des requêtes sur des réseaux pair à pair (PaP) et enregistre les adresses IP des utilisateurs ayant mis le fichier à disposition,
- Les agents assermentés des ayants droit valident ces constatations et signent les saisines transmises à la commission de protection des droits.

Dans ce contexte, l'objet de la présente mission d'expertise consiste à déterminer si le mode opératoire utilisé permet l'identification sans équivoque d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre.

Le 23 mai 2011, j'ai signé un engagement de confidentialité avec l'Hadopi.

Le 10 juin 2011 j'ai été confirmé pour traiter de la mission.

3. RAPPEL DE LA MISSION

3.1 Analyser avec précision la méthode utilisée pour créer l'empreinte numérique d'une œuvre :

- Vérifier que cette méthode permet la création d'une empreinte unique de l'œuvre,
- Calculer la probabilité pour deux œuvres différentes de donner lieu à la création d'une même empreinte.

3.2 Analyser la méthode utilisée pour comparer les œuvres mises à disposition et les empreintes :

- Déterminer si les critères de comparaison utilisés entre deux empreintes sont suffisants pour authentifier une même œuvre,
- Evaluer dans quelle mesure le processus de comparaison des œuvres ne génère pas de faux-positifs.

3.3 Analyser le processus de collecte des adresses IP :

- Déterminer si le processus de collecte des adresses IP permet d'attester que les adresses IP enregistrées mettent effectivement à disposition les œuvres visées dans le procès verbal de constatation,
- Evaluer dans quelle mesure le processus de collecte protège contre les usurpations d'IP.

3.4 Analyser les processus automatiques et/ou semi automatiques et/ou manuels entre les différents acteurs (ayant droit, collecte des données IP, agents assermentés des ayants droit) qui entrent dans le mode opératoire d'identification d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre.

L'Expert pourra se faire assister d'un sachant ayant au préalable signé un accord de confidentialité et pourra utiliser tous moyens matériels pour accomplir sa mission ci-dessus,

Le délai d'exécution de la mission n'a pas été respecté du fait des agendas des différents acteurs.

4. REUNIONS D'EXPERTISE

4.1 Réunion du 23 mai 2011 chez HADOPI

Cette réunion avait pour objet de communiquer à l'Expert le projet de mission et la liste des contacts :

- Les Ayants- droit qui saisissent l'Hadopi :

SCPP : Société Civile des Producteurs Phonographiques

SPPF : Société Civile des Producteurs Phonographiques de France

ALPA : Association de Lutte contre la Piraterie Audiovisuelle.

SACEM / SDRM

- Leurs prestataires :

  TMG, Trident Media gard, qui gère la plateforme informatique faisant objet de la présente expertise

KANTAR MEDIA, qui transmet les masters des œuvres phonographiques, il s'agit des supports qui contiennent les métadonnées et le fichier de l'œuvre intégrale

FPM et FPA : Pour des raisons de confidentialité nous ne révèlerons pas quels sont les fournisseurs de solution de calcul d'empreintes. Dans la suite du document nous appellerons le fournisseur de la solution pour le calcul d'empreinte d'œuvres musicales FPM et nous appellerons le fournisseur de la solution pour le calcul d'empreinte d'œuvres audiovisuelles FPA.

4.2 Réunion du 10 juin 2011 au Cabinet de l'Expert

Lors de cette réunion, il m'a été exposé la perception technique d'Hadopi du système (approche en 3 étapes) permettant d'aboutir au PV de saisine auquel est joint un « Chunk » (une partie du fichier représentatif de l'œuvre, dont la mise à disposition a été constatée sur un réseau PaP).

4.3 Réunion du 15 juin 2011 chez Hadopi

Lors de cette réunion nous avons poursuivi les discussions techniques ; un certain nombre de documents m'ont été remis (cf. Annexe 1) ainsi qu'une clé USB contenant 3 saisines à titre d'exemple (ALPA, SACEM, SPPF) (format XML) ; une saisine est composée de 3 fichiers XML.
A l'annexe 3.4 figure un exemple de saisine.

4.4 Réunion du 27 juin 2011 chez ALPA

Alpa n'est pas titulaire des droits, mais dispose d'une délégation de pouvoir des ayants droit pour constater les faits de contrefaçon en matière audiovisuelle et pour saisir l'Hadopi.

Lors de cette réunion nous avons visualisé les écrans utilisés par l'ALPA et développés par TMG pour les Agents assermentés qui établissent et signent les constats ainsi que les process (cf. supra). La cinématique de ces écrans a fait l'objet d'une édition avec commentaires et figure à l'annexe 3.3 et 3.4 du présent rapport.

4.5 Réunion avec TMG du 28 juin 2011 au Cabinet de l'Expert

Cette première réunion avait pour objet de préparer la mission technique et de lister les documents nécessaires à la compréhension des processus par les intervenants (FPM, FPA, Agents assermentés...).

Les acteurs :

- ALPA (Gaumont, Pathé..) pour l'audiovisuel et utilisation du système d'empreintes FPA
- SCPP
  SACEM
  SPPF
  } Pour la musique et utilisation du système d'empreintes « FPM »

4.5.1 Informations sur les œuvres Audiovisuelles

Métadonnées de l'œuvre

- Titre,
- Titre original,
- Dates de sorties (salle, DVD) ; France, USA, ...
- Ayants droit (producteurs, distributeurs, ...).

Données d'identification :

FPA a fourni à TMG sous forme de licences les logiciels de calcul de l'empreinte d'une œuvre en binaire (pas de code sources) ; FPA a aussi fourni à des laboratoires les codes de l'algorithme de calcul de l'empreinte.

En pratique, ce sont les laboratoires qui fournissent ces données après validation par les ayants droit.

Pour collecter les données, un compte FTP existe entre tous les labos (TMG ↔ Labo) ou via FPA.

TMG a développé des interfaces hommes machines (IHM) pour les ayants droit français.
TMG peut aussi se connecter pour collecter des données au format XML.

Dans certains cas, les ayants droit peuvent envoyer une BETACAM et/ou un DVD.

4.5.2 Informations sur les œuvres Musicales

C'est la société Kantar Media qui adresse les fichiers Audio ou Vidéo (clip) en format AVI pour les clips et MP3 pour la musique ; fichiers MP3 ou AVI plus les métadonnées en XML.

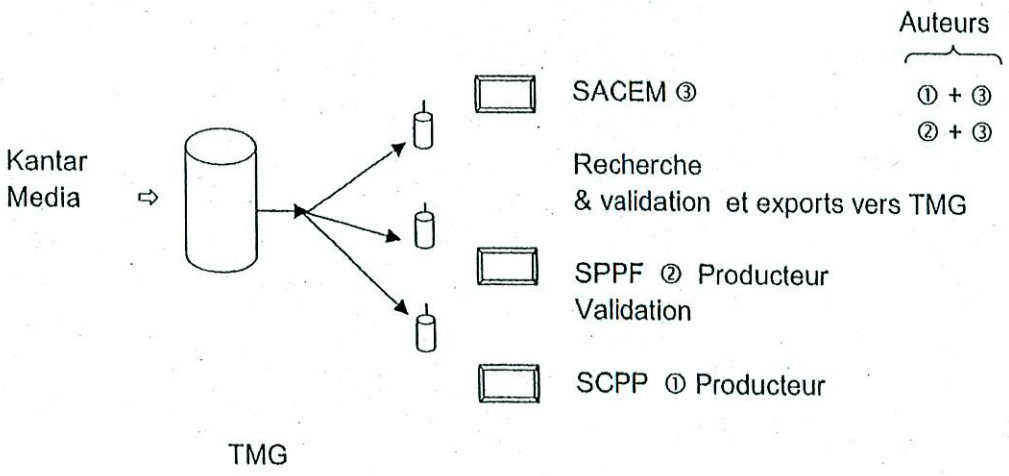
4.5.3 Rôle des agents assermentés dans l'attribution des droits (titularisation d'une œuvre)

- Ce rôle n'est joué que pour les ayants droit dans le domaine de la musique, SCPP, SACEM et SPPF ; L'ALPA n'a pas besoin de procéder à cette vérification.

- TMG reçoit les données directement de Kantar Media et adresse tous les fichiers reçus aux ayants droit (SPPF, SACEM, SCPP) qui, à travers les informations figurant dans les métadonnées adressées par Kantar, valident à travers un IHM construit par TMG les titulaires des droits.

Ex : un fichier qui contient 3 œuvres arrive de Kantar ; ces 3 œuvres seront vues par les 3 entités et chacune s'attribue les œuvres par la connaissance de ses propres ayants droit.

SCPP ⇨ Titulaire des œuvres
SPPF pour les producteurs
SACEM ⇨ Titulaire des œuvres pour les auteurs



Les Agents assermentés accèdent au système TMG via une clé (Token) sécurisée (cf. Annexes 3.3 et 3.4 du présent rapport).

4.5.4 Constitution de la base de données des œuvres qui sont sur les réseaux Pair à Pair (PaP)

⇨ Recherche par mots clés dans chaque réseau PaP à travers les outils fournis par les protocoles.

⇨ Méthode des jeux de mots clés.

4.5.5 Méthode de comparaison des empreintes (voir réunion du 28 Septembre)

Pour calculer les empreintes sur les œuvres ayant pour origine un réseau PaP, on recherche des séquences ordonnées sur lesquelles on calcule des empreintes.

Ex : Œuvre J

E^1 E^n

$\left\{ E^J_1 . E^J_w \right\}$ Si on trouve dans les fichiers/œuvres une séquence ordonnée de « n » empreintes alors il y a reconnaissance de l'œuvre (Ex : pour ALPA, ~ 35') ; pour une œuvre musicale, on prend 80 % de la durée.

A ce stade, TMG a constitué son référentiel de comparaison et collecté les œuvres reconnues sur ce principe dans les réseaux PaP (cf. Annexe 2 du présent rapport).

4.5.6 Phase de collecte des « incidents »

PHASE 1	PHASE 2
1.0 Reconnaissance de l'œuvre par le biais des empreintes + 1.1 Validation par les Agents assermentés de tout (si une œuvre n'est pas validée elle reste en attente et ne rentre pas dans le process)	2.0 Collecte des fichiers (œuvres mises à disposition sur les réseaux PaP) sélectionnés sur la base de multicritères (métadonnées) ; le fichier est « préparé » (savoir-faire TMG) afin de pouvoir appliquer l'algorithme de génération d'empreinte. Si l'empreinte du fichier existe dans la base de référence, le système relève le hashcode du fichier aux fins qu'il soit validé une seconde fois par l'agent assermenté. 3.0 Collecte des adresses IP

4.5.7 Mode de collecte de l'adresse IP

- TMG capte l'adresse IP au moment de l'établissement de la connexion (Mode connecté).
- L'adresse est prise dans le socket et dans une même session (cf. Annexe 2.2 du présent rapport).

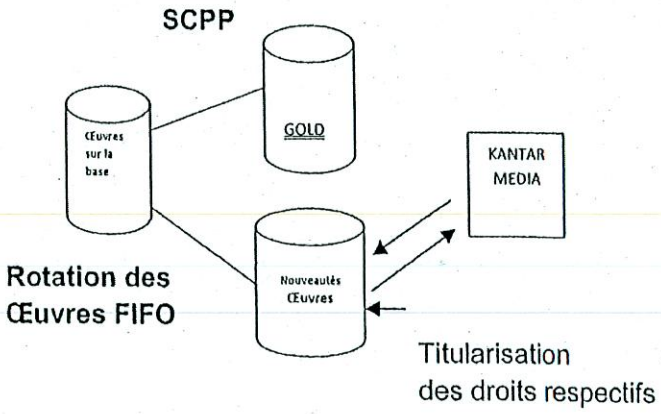
4.6 Réunion du 29 juin 2011 à la SCPP

Durant cette réunion la SCPP expose à l'Expert la façon dont est organisé le travail de ses agents assermentés chargés d'établir les constats transmis à l'Hadopi.

« ... Les agents assermentés ont pour mission de constater la matérialité des atteintes aux droits, Producteurs et Auteurs.

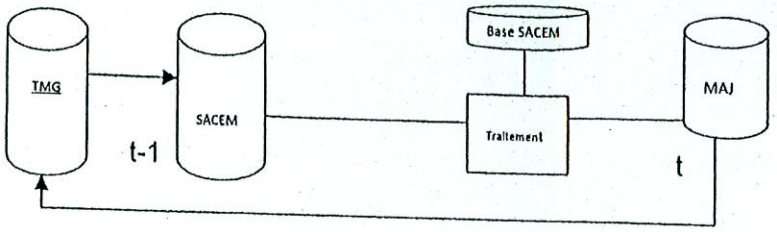
- *Les agents assermentés font du téléchargement et vérifient qu'il correspond bien aux répertoires,*
- *Plaintes sur les PV des agents assermentés,*
- *Un agent assermenté est un salarié des entités ayant reçu un agrément.*

- *Dans le cadre de notre mission générale, on surveille toutes les œuvres de nos producteurs ;*
- *Il y a une base de données qui n'est pas exhaustive ; cette base contient X œuvres... »*



Planning d'envoi des métadonnées

- En fin de matinée, la SACEM interroge les œuvres revendiquées par un de ses auteurs ; ce travail est fait dans un délai de 48 heures et le fichier est complété et envoyé.



- Après l'envoi des œuvres qui mettent à jour la base TMG, la base est à jour pour les producteurs (SCPP, SPPF) et sous 48 heures pour les auteurs (la SACEM).

⇩

- Phase de traitement TMG
- MP3 pour la musique, MPEG pour les clips et XML pour les métadonnées.

⇩

La titularisation est complète à ce stade

C'est TMG qui a, dans le cadre de son appel d'offres, proposé le système FPM ; ce système a été recetté.

- C'est TMG qui calcule l'empreinte sur les enregistrements adressés par Kantar Media.

⇩

L'ensemble du système a été recetté incluant le système FPM (les PV de recettes qui m'ont été adressés figurent à l'Annexe 3.2 du présent rapport).

4.7 Réunion du 1er juillet 2011 à la SACEM

Lors de cette réunion nous avons observé le process d'intégration des nouveautés (cf. infra)
Ce process est hebdomadaire, Kantar Media met à disposition les œuvres et les envoie à TMG.

4.8 Réunion du 8 juillet 2011 à la SCPP

Lors de cette réunion nous avons suivi le processus de validation des Agents assermentés qui n'a pas pu être fait le 29 juin ; lors de cette réunion on a pu constater les phases de titularisation, la validation des œuvres collectées sur les réseaux PaP et la validation par échantillonnage de la conformité des PV de constatation (cf. Annexes 3.3 et 3.4).

A ma demande, une cinématique commentée des process a été faite (cf. Annexe 3.4).

4.9 Réunion du 9 septembre 2011 chez Kantar Media

Lors de cette réunion j'ai pu valider les process vus antérieurement.

Kantar Média envoie les œuvres à TMG qui va générer les empreintes.

L'ensemble du process est décrit dans les documents confidentiels qui m'ont été transmis par Kantar Média (cf. Annexe 4 du présent rapport).

4.10 Réunion du 28 septembre 2011 chez TMG à Nantes

Sur demande de l'Expert, il a été procédé à l'analyse de la génération des empreintes des œuvres avec la méthode FPM (phonogramme et clip) et la méthode FPA (audiovisuel).

Définition

L'empreinte (fingerprint en anglais) d'une œuvre est un identifiant d'une œuvre musicale ou audiovisuelle. Cette empreinte est indépendante du support ou de l'encodage de l'œuvre. Un système d'empreinte est dit robuste si ce système assure l'unicité de l'empreinte de chaque œuvre.

Une note blanche FPM (article) m'a été remise dans laquelle est exposée la méthode (cf. Annexe 2).

Découpage du fichier audio en morceaux de 3 secondes et calcul du « BER » (Bit Error Rate) qui doit être inférieur à 0,35 pour FPM mais TMG l'a diminué à 0,20 pour renforcer encore la fiabilité du système.

TMG a fixé la durée maximum à 120 secondes et si le fichier adressé par Kantar Media est inférieur à 120 s, on prend 80 % de la longueur du temps.

Les informations générées par le système FPM passe par une API (interface logiciel) communiquée à TMG par FPM.

En synthèse, TMG traite la sortie FPM en ajoutant une couche de contrôle TMG (voir algorithme) (cf. Annexe 2.1).

Ce principe est le même pour le système FPA (cf. Annexe 2.2).

L'objectif du traitement pour générer les empreintes est d'éviter les négatifs et faux positifs ; une fois le fingerprint calculé, on soumet aux agents assermentés les œuvres pour validation.

Mise à jour de la base et alimentation dans la base de production d'une référence (numéro séquentiel) lié au fingerprint.

Le processus de collecte des œuvres mises à disposition sur les réseaux PaP se fait bien par recherche à partir des Métadonnées (sélection de mots clés) ; chargement de l'ensemble des fichiers (unique ZIP ou autres) et génération par une procédure spécifique à TMG des fichiers INPUT aux systèmes FPM ou FPA.

Collecte des hashcodes de ces fichiers et demande de validation par les Agents assermentés (cf. Annexe 3.4).

> **Définition**
> Le « hashcode » d'un fichier est un identifiant d'un fichier. Ce « hashcode » ne présume pas ce que représentent les données contenues dans le fichier et son calcul est uniquement basé sur la suite numérique qui constitue ce fichier. Un système de hashcode est dit robuste si ce système assure l'unicité du hashcode de chaque fichier.
> Les hashcode sont utilisés dans les protocoles PaP pour identifier un fichier.

Puis le système collecte l'adresse IP et le segment de 16 ko associé à l'œuvre mise à disposition (time stamp et position du segment dans le fichier) (cf. Annexe 2). TMG calcule son propre hashcode du segment (SHA1). Le segment est un sous ensemble du fichier complet de l'œuvre et constitue une preuve démontrable.

Spontanément et sans aucune préparation, j'ai demandé à interroger la base de référence afin de vérifier s'il y avait des doublons ; nous avons constaté 23 doublons sur la base Musique qui ont pu être expliqués par le fait que Kantar Média a transmis 2 masters (un master est le support original qui contient les métadonnées et le fichier de l'œuvre intégrale) pour la même œuvre ; le master est fourni par l'éditeur ; par la suite j'ai demandé à faire le même contrôle pour l'audiovisuel ; il n'y avait pas de doublons (cf. Annexe 5).

Ce même jour et pour éliminer tout aléa pour l'avenir, j'ai demandé à TMG, qui a accepté, d'introduire un test sur les bases des œuvres de référence de non existence d'un doublon ; et en cas d'apparition, de vérifier que l'empreinte ne concerne pas 2 œuvres différentes (faux positif) ; ce test garantit qu'un faux positif dans la base de référence serait immédiatement découvert.

Après les procédures de vérification, j'ai poursuivi ma réunion en allant constater le data center où se trouve une des plateformes TMG.

4.11 Conférence téléphonique du 10 octobre 2011 avec FPM chez L'Expert

Lors de cette conférence téléphonique, FPM a réexpliqué la méthodologie exposée dans le White Paper remis par TMG (cf. Annexe 2.2.1).

Après discussion, j'ai suggéré à FPM de tester la solidité des paramètres utilisés par TMG, à savoir : < à 0,20 pour le BER alors que le système FPM est paramétré à 0,35, auquel il faut ajouter le processus de non retenu d'une œuvre si on constate un doublon ou une quelconque erreur sur 10 % des traces communiquées par le logiciel FPM (cf. Note Fingerprint Audio/Vidéo TMG) (cf. Annexe 2.2.1)

FPM a passé sa batterie de tests sur la base des paramètres qui ont été communiqués par TMG le mercredi 12 octobre ; FPM estimant à une semaine le délai de réalisation de ce test.

Le résultat m'a été communiqué par Mail du 25 Octobre 2011 et confirme la « force » des paramètres standard FPM et ceux de TMG (cf. Supra) ce mail est confidentiel.

4.12 Réunion du 11 octobre 2011 à FPA

FPA n'a pas souhaité, tout comme FPM, exposer son savoir faire. Cependant la présentation qui a été faite et les exemples donnés, notamment pour la Télévision où le risque de faux positifs pouvait exister (il est beaucoup plus faible pour les œuvres Audiovisuelles), me permet de dire que le risque d'avoir une même empreinte pour deux œuvres différentes est quasi nul, d'autant plus que ce risque est réduit par le test qui est fait dans la base par un contrôle des doublons (cf. Annexe 2.2.2).

5. REPONSE A LA MISSION

Rappel de définitions

L'empreinte (fingerprint en anglais) d'une œuvre est un identifiant d'une œuvre musicale ou audiovisuelle. Cette empreinte est indépendante du support ou de l'encodage de l'œuvre. Un système d'empreinte est dit robuste si ce système assure l'unicité de l'empreinte de chaque œuvre.

Définition
Le « hashcode » d'un fichier est un identifiant d'un fichier. Ce « hashcode » ne présume pas ce que représentent les données contenues dans le fichier et son calcul est uniquement basé sur la suite numérique qui constitue ce fichier. Un système de hashcode est dit robuste si ce système assure l'unicité du hashcode de chaque fichier.
Les hashcode sont utilisés dans les protocoles PaP pour identifier un fichier.

Pour répondre à la mission, il fallait que je garantisse la fiabilité/robustesse :

- des méthodes de génération et de comparaison des empreintes (fingerprint) (voir 5.1 et 5.2)
  - pour la méthode utilisée pour les œuvres musicales (technologie développée par FPM)
  - pour la méthode utilisée pour les œuvres audiovisuelles (technologie développée par FPA)
- du processus de collecte des adresses IP (voir 5.3)
- de l'utilisation du système par les Agents Assermentés (voir 5.4)

Le schéma suivant résume les différentes actions effectuées par les différents acteurs du système.

3. **Constitution de la base des fichiers (et de leur Hashcode) à surveiller sur le PaP**
    - Recherche sur les réseaux PaP de fichiers contenant potentiellement des œuvres de référence. Cette recherche se fait à partir de mots clés en rapport avec les titres des œuvres de référence.
    - Validation du fait que les fichiers résultant de la recherche contiennent bien une œuvre de référence. Cette validation se fait d'abord par un système de comparaison automatique à partir des empreintes, ensuite manuellement par les agents assermentés.
    - Pour chacun des fichiers résultant de la recherche contenant bien une œuvre de référence, conservation du fichier et du Hashcode de celui-ci.
    - Le résultat est une base de fichiers dont on a la certitude qu'ils contiennent une œuvre de référence. Ces fichiers sont associés à leur Hashcode, et ce sont ces fichiers qui seront surveillés sur les réseaux PaP.
4. **Constatation de la mise à disposition des fichiers surveillés sur le PaP**
    - Collecte des adresses IP des accès à Internet à partir desquels sont été mis à disposition les fichiers surveillés sur le PaP (voir étape précédente).
    - Pour chacune des adresses IP collectées, téléchargement à partir de cette IP d'un segment du fichier mis à disposition.
    - Le résultat est un ensemble de constats de mise à disposition illicite de fichiers. Un constat contient notamment une adresse IP à partir de laquelle un segment de fichier a été mis à disposition, le segment de fichier ainsi qu'un horodatage de la mise à disposition.
5. **Génération des PVs**
    - Génération automatique des PV à partir des constats avec en pièce jointe le segment (signé) du fichier mis à disposition sur une adresse IP (ce fichier contenant une œuvre de référence) ainsi que le moment de cette mise à disposition.

Il est important de comprendre que l'empreinte ne sert qu'au traitement interne du système pour identifier une œuvre de référence (étape 2) et pour la comparer à l'empreinte qui est générée par les mêmes algorithmes sur les fichiers mis à disposition sur les réseaux PaP et qui sont sélectionnés par mots clés (étape 3).

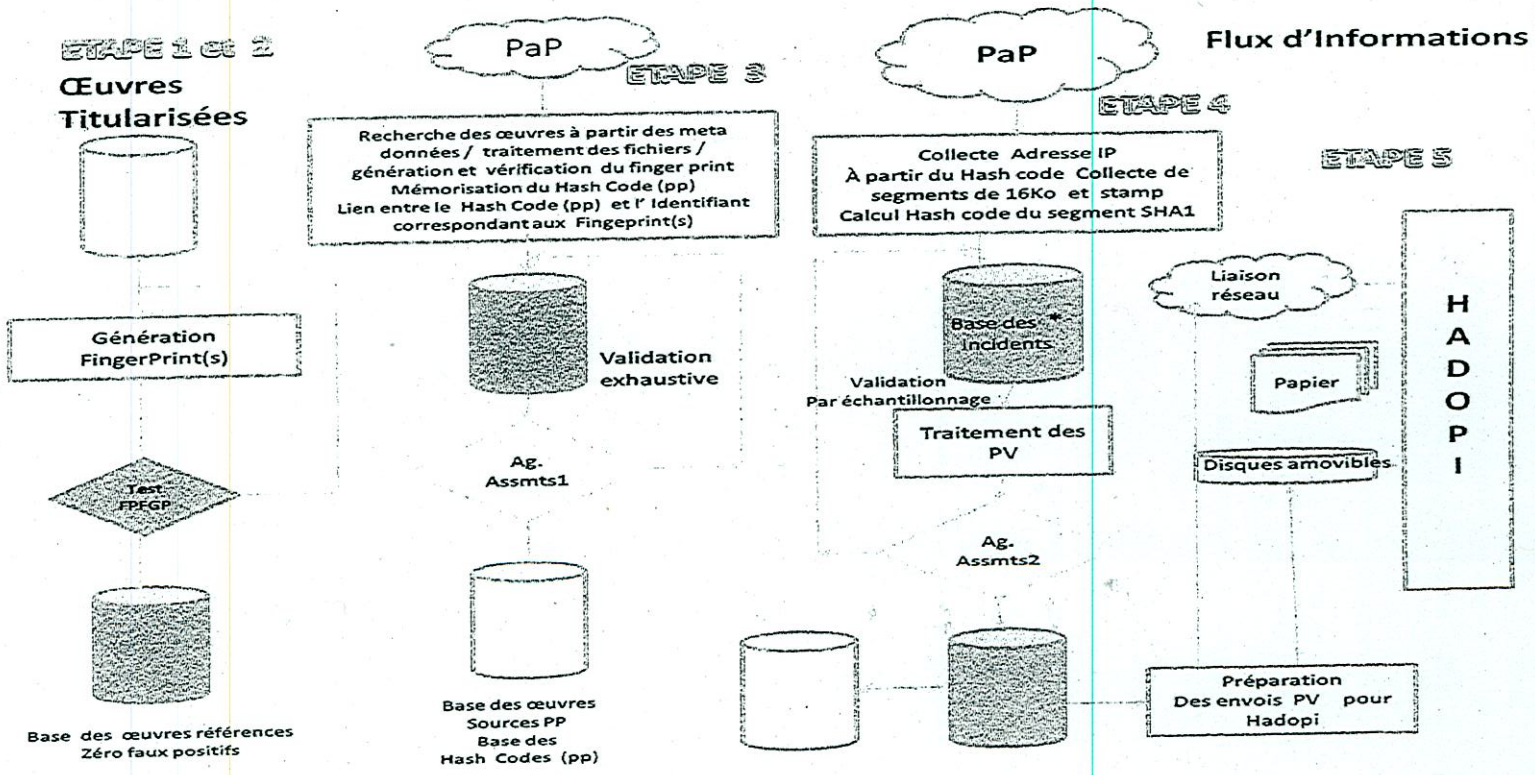
TMG a développé un savoir faire dans le traitement des fichiers mis à disposition sur les réseaux PaP pour pouvoir, quel que soit le format d'un fichier, appliquer les algorithmes de calcul d'empreintes sur le contenu de ces fichiers.

Cette précision est nécessaire pour ne pas faire de confusion entre le Hashcode du fichier et l'empreinte d'une œuvre/représentation du contenu d'un fichier.

A l'Annexe 7, figurent les PV de recettes de l'application (SACEM/SDRM, SCPP, SPPF).

A l'Annexe 8, j'ai joint la liste des personnes qui ont participé et/ou répondu à mes questions pour accomplir ma mission.

Le schéma ci-dessous synthétise le flux général du système de génération des PV de constatation.



*** Base d'incidents : éléments d'informations recueillies et qui apparaissent dans les PV.**

5.1 Analyser avec précision la méthode utilisée pour créer l'empreinte numérique d'une œuvre :

- Vérifier que cette méthode permet la création d'une empreinte unique de l'œuvre,
- Calculer la probabilité pour deux œuvres différentes de donner lieu à la création d'une même empreinte.

Pour répondre à cette partie de la mission et compte tenu du fait que FPM et FPA ont refusé de dévoiler leur savoir faire, je me suis attaché à vérifier la robustesse de leurs méthodes par les exposés qui m'ont été faits (cas pour FPA) et par un complément de test dans le cas de FPM. (cf. Annexes 2.2.1 et 2.2.2).

J'ai ensuite analysé le processus de génération des empreintes afin de voir s'il était possible d'insérer un contrôle pour garantir la non présence d'une même empreinte pour deux œuvres différentes dans la base de référence (cf. Annexe 6) (Analyse du protocole). Ce test garantit une remontée d'alerte en cas de modification des algorithmes de génération d'empreintes.

FPM

A ma demande FPM a testé les paramètres utilisés par TMG. La réponse ci-dessous confirme des paramètres TMG :

"..When we run the <FPM> test set for audio with the default settings, the technology has a false positive ratio of 0.000%. If we apply the TMG parameters -which are more stringent than the default settings- to this test set, the false positive ratio will remain 0.000%. .."

Traduction

"...Quand nous exécutons le test avec les paramètres standard <FPM>, nous obtenons un ratio de faux positifs de 0,000 % ; Si nous appliquons à ce test les paramètres TMG qui sont plus contraignants que nos paramètres standard, le ratio de faux positifs ne change pas et reste à 0,000 % ...».

J'ai validé ce résultat par un test que j'ai effectué in situ chez TMG à ma demande et sans préparation sur la base de données de références.

Suite à la mise en évidence 23 doublons d'empreintes dans la base, j'ai constaté que ces doublons correspondent tous au cas où les Ayants droit (Kantar) ont transmis 2 masters pour la même œuvre (un master est un support qui contient les métadonnées et le fichier de l'œuvre intégrale). Ces doublons n'étaient donc pas des cas de faux positifs.

Définition : Faux positif = même empreinte pour deux œuvres différentes.

Conclusion : les algorithmes de calcul d'empreintes de FPM sont robustes et garantissent la non existence de faux positifs.

FPA
Pour FPA l'approche anti faux positifs sur la conception algorithmique est la suivante :

- *L'objectif est : « aucun faux positif sur les vérités terrain » pour permettre un bon niveau d'automatisation en production*
- *En cas d'apparition de faux positif en production, le cas serait traité prioritairement en évolution de l'algorithme*
- *Des vérités terrain à grande échelle sont conçues pour tester l'absence de faux positif ; elles sont utilisées en validation de non régression :*
  3 000h réf. contre 4 jours de TV
  10 000h ref. contre 1 000h
- *Dans le cadre d'analyses spécifiques effectuées sur le filtrage par les sites de partage (en production depuis 01/2008), aucun faux positif n'a été identifié*

Les tests qui m'ont été présentés montrent la robustesse de leur algorithme.

Une recherche de doublons a été faite sur la base de référence ; il n'y a que des « fingerprints » uniques. La description de la base référence Vidéo figure en Annexe 2.2.2.

Conclusion : les algorithmes de calcul d'empreintes de FPA sont robustes et garantissent la non existence de faux positifs.

5.2 Analyser la méthode utilisée pour comparer les œuvres mises à disposition et les empreintes :

- Déterminer si les critères de comparaison utilisés entre deux empreintes sont suffisants pour authentifier une même œuvre,
- Evaluer dans quelle mesure le processus de comparaison des œuvres ne génère pas de faux-positifs.

La méthode utilisée pour comparer les œuvres mises à disposition sur les réseaux PaP et leurs empreintes est strictement identique à la génération d'empreinte des œuvres de référence.

Avant de générer les empreintes (cf. Annexe 2 ; document CEI V1.0 Fingerprint Audio/Vidéo), TMG doit « préparer » les fichiers collectés par mots clés sur les réseaux pair à pair (PaP) :

Si l'empreinte générée n'est pas trouvée dans la « base de référence », elle ne pourra pas être identifiée (pour mémoire la base de référence ne doit pas contenir d'empreintes identiques pour deux œuvres différentes, cf. supra)

Conclusion : les algorithmes de comparaison d'empreintes de FPA et les algorithmes de comparaison d'empreintes de FPM sont robustes et garantissent la non existence de faux positifs.

Aussi compte tenu de la problématique du secret, et bien que les approches de FPA et FPM soient robustes, j'ai proposé à TMG, pour éviter la présence de faux positifs dans « les bases de références », et par conséquent dans la « base source », de sécuriser le processus par l'adjonction d'un test. Ce test consiste à vérifier qu'il n'y a pas, dans la base, un même « fingerprint » pour des œuvres différentes (cf. schéma Flux Informations : Test FPFGP) afin de les détecter et les analyser en semi automatique.

TMG a accepté d'intégrer ce test dans son exploitation (cf. Annexe 7).

Ce test, une fois mis en place, garantira à l'Hadopi que dans la base de référence deux œuvres différentes ne pourront pas avoir une même empreinte sans que cela ne soit remarqué.

5.3 Analyser le processus de collecte des adresses IP :

- Déterminer si le processus de collecte des adresses IP permet d'attester que les adresses IP enregistrées mettent effectivement à disposition les œuvres visées dans le procès verbal de constatation,
- Evaluer dans quelle mesure le processus de collecte protège contre les usurpations d'IP.

Le processus de collecte des adresses IP est décrit dans la documentation qui figure en Annexe 2.2 (CEI & Version 1.2 Spécification des Nodes de collecte).

Le principe même de fonctionnement des protocoles pair à pair (PaP) fait que la connexion doit être permanente lors du téléchargement d'un segment à partir d'une adresse IP et la méthode utilisée dans le système est conforme au fait que le téléchargement est fait dans une seule session (Socket au sens BSD version UNIX).

La modification d'une adresse IP en PaP n'a pas de sens si on veut garder un téléchargement intègre.

TMG collecte aussi d'autres informations, log du début et fin de chargement du segment, ce complément d'information (horodatage) augmente la preuve dans le contexte des réseaux pair à pair (PaP).

Tous les accès à la base de données des infractions sont réalisés par des procédures stockées, de plus l'adresse IP est cryptée dans la base des incidents.

En l'état on peut dire que la méthode de capture des adresses IP avec leurs compléments d'informations (horodatage) est robuste pour toutes les attaques externes, et comme pour tous les systèmes on ne peut pas exclure l'attaque interne.

TMG, à ma demande, a produit un document justifiant le dépôt des logiciels. J'ai recommandé le dépôt des codes sources des logiciels, des exécutables associés ainsi que de l'ensemble de la documentation, et ce au fur et à mesure des modifications. La date de dépôt doit correspondre aux versions en cours d'exploitation (cf. Annexe 7).

Conclusion : L'analyse de l'architecture et de la méthodologie de collecte des adresses IP me permettent de dire que le système est cohérent et FIABLE.

5.4 Analyser les processus automatiques et/ou semi automatiques et/ou manuels entre les différents acteurs (ayant droit, collecte des données IP, agents assermentés des ayants droit) qui entrent dans le mode opératoire d'identification d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre.

J'ai analysé l'ensemble des processus automatiques et semi automatiques entre les différents acteurs. Les personnes qui ont un rôle clé dans ce processus sont les Agents Assermentés ; ils ont pour tâche de valider les œuvres de référence qui sont intégrées dans la base tout comme les œuvres reconnues sur les réseaux pair à pair (PaP) ; cette validation est exhaustive (test Assmts1).

Une préparation est faite pour structurer les fichiers venant du réseau pair à pair (PaP) pour permettre le calcul de l'empreinte des contenus de ces fichiers et ainsi permettre la comparaison avec celles de la base de référence.

Si une œuvre n'est pas validée, elle ne rentre pas dans le processus et reste indéfiniment en suspens jusqu'à sa validation par l'Agent.

Les Agents assermentés interviennent dans une deuxième validation qui est la conformité du PV, cette validation est faite par échantillonnage ; l'échantillon doit être validé dans sa totalité (test Assmts2).

J'ai pu constater le fonctionnement de ces deux tests lors de ma visite chez l'ALPA et la SCPP (cf. Annexes 3.3. et 3.4).

Conclusion : Les processus automatiques et/ou semi automatiques et/ou manuels entre les différents acteurs qui entrent dans le mode opératoire d'identification d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre est FIABLE.

6. CONCLUSION

Pour répondre positivement à la mission, il fallait que je garantisse la fiabilité/robustesse :

- de l'utilisation du système par les Agents Assermentés
- des méthodes de génération et de comparaison des empreintes (fingerprint)
  - pour la méthode utilisée pour les œuvres musicales
  - pour la méthode utilisée pour les œuvres audiovisuelles
- du processus de capture des adresses IP

En l'état :

Concernant l'utilisation du système par les Agents assermentés :

Les processus automatiques et/ou semi automatiques et/ou manuels entre les différents acteurs qui entrent dans le mode opératoire d'identification d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre est FIABLE.

Concernant les méthodes de génération et de comparaison des empreintes (fingerprint) :

Les algorithmes de calcul d'empreintes utilisés pour les œuvres musicales sont ROBUSTES et garantissent la non existence de faux positifs.

Les algorithmes de calcul d'empreintes utilisés pour les œuvres audiovisuelles sont ROBUSTES et garantissent la non existence de faux positifs.

Les algorithmes de comparaison d'empreintes utilisés pour les œuvres musicales sont ROBUSTES et garantissent la non existence de faux positifs.

Les algorithmes de comparaison d'empreintes utilisés pour les œuvres audiovisuelles sont ROBUSTES et garantissent la non existence de faux positifs.

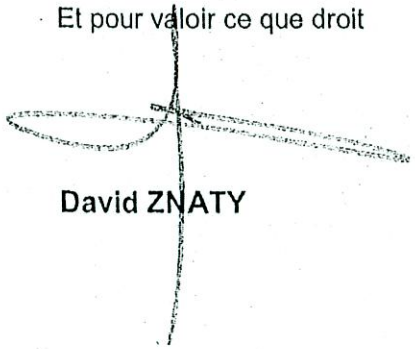
Concernant le processus de capture des adresses IP :

L'analyse de l'architecture et de la méthodologie de collecte des adresses IP me permettent de dire que le système est cohérent et FIABLE.

Conclusion : en l'état, le processus actuel autour du système TMG est FIABLE. Les documents constitués du procès verbal (saisine), et si nécessaire du fichier complet de l'œuvre (stocké chez TMG) associé au segment de 16 Ko constituent une preuve ROBUSTE.

Le mode opératoire utilisé permet donc l'identification sans équivoque d'une œuvre et de l'adresse IP ayant mis à disposition cette œuvre.

Et pour valoir ce que droit



David ZNATY