Image and Research
14th International Conference
Girona. 16-18 November 2016

Les métadonnées des images fixes et les médias sociaux

Résumé
Après un bref rappel sur les différents standards de métadonnées embarquées dans les images numériques fixes, nous examinerons la façon dont certaines plates-formes sociales se comportent avec ces informations. Nous terminerons en examinant les (rares) possibilités d'exploitation des métadonnées sur les réseaux sociaux à travers l'expérience du projet *PhotosNormandie* sur la plate-forme *Flickr*.

1. Rappel sur les différents standards de métadonnées embarquées dans les images

Afin de comprendre les deux modes opératoires permettant de décrire des images numériques fixes à l'aide de métadonnées, souvenons-nous des albums photos de jadis.
Pour décrire une photo, pour se rappeler des dates, des lieux, des personnes qui y figurent, de courtes légendes sont écrites sur l'album, à côté de la photo. Mais si les photos se détachent de l'album d'une façon ou d'une autre et deviennent totalement déclassées, l'association des légendes aux images est perdue et il peut être difficile de reconstituer ce lien.
Tout le monde bien sûr connaît la solution à ce problème. Il suffit d'écrire la légende au verso de la photo. De cette manière, la légende est réellement liée à la photo.

Avec l'écriture d'une légende sur un album et le griffonnage de quelques mots au dos d'une photo, les deux modes opératoires essentiels qui permettent de décrire l'image photographique sont représentés depuis fort longtemps. Nous retrouvons en effet les mêmes principes avec nos objets numériques, transposés en langage moderne; on reconnaît d'une part l'affectation de métadonnées ***externes*** à l'image [la légende dans un catalogue ou sur un album] et d'autre part l'ajout de métadonnée ***internes*** à l'image, c'est-à-dire l'insertion de métadonnées embarquées dans le fichier numérique qu'est l'image [l'écriture au verso de la photo].

Nous allons nous intéresser maintenant à trois techniques spécifiques de métadonnées des images numériques fixes:

- EXIF
- IPTC/IIM
- XMP

EXIF est une abréviation de *EXchangeable Image File*.
Le format EXIF a été développé en octobre 1995 par le consortium japonais JEIDA (*Japan Electronic Industry Development Association*) qui regroupait les industries électroniques avant l'année 2000. Ce n'est pas un standard, mais il est supporté par tous les fabricants d'appareils photographiques numériques (APN), avec néanmoins des variantes propriétaires.

Ce format définit les paramètres de prise de vue et les réglages de l'appareil au moment de la capture numérique.
Ce sont des métadonnées de type *interne*, parmi lesquelles on peut citer: le nom du fabricant et le modèle de l'appareil, les hauteur et largeur de l'image, les date et heure de la prise de vue, l'orientation, la résolution, le temps d'exposition, l'ouverture, la présence ou non d'un flash, les coordonnées GPS, etc.

L'IPTC (*International Press and Telecommunications Council*) est un consortium qui réunit les principales agences de presses du monde.
L'IPTC compte actuellement actuellement une cinquantaine d'agences (dont l'AFP et Agencia EFE, la principale agence de presse en langue espagnole au monde).
L'IPTC développe des standards techniques d'échange de données pour la presse.
Ces standards sont employés par la quasi-totalité des agences de presse du monde.

L'IPTC et la NAA (*Newspaper Association of America*) ont créé en 1991 le modèle global de données appelé *Information Interchange Model* [*IIM*]
Dès 1994, la société Adobe a utilisé un sous-ensemble simplifié de l'IIM pour définir dans son logiciel Photoshop les informations associées à une image.
Les métadonnées IPTC/IIM sont des métadonnées de type interne. Il s'agit d'un ensemble de champs textuels stockés dans le fichier image: Titre, Légende, Mots-clés, Copyright, etc.
Ce standard est toujours très utilisé dans la presse et l'édition bien qu'il soit considéré comme obsolète et remplacé par XMP.

XMP (*eXtensible Metadata Platform*) a été créé par la société Adobe en septembre 2001
Cette technique utilise une version simplifiée de RDF (*Resource Description Framework*) qui est un standard développé par le W3C, base du Web sémantique; RDF permet d'encoder, échanger et réutiliser des métadonnées structurées et peut s'exprimer en XML.
Les images munies de métadonnées XMP constituent probablement la plus grande collection d'objets décrits en RDF sur le Web, mais la technique n''est pas réservée aux images...
XMP utilise le schéma Dublin Core comme fondation.
C'est un standard ISO depuis mars 2012 (ISO 16684-1:2012), ce n'est donc plus une technologie spécifiquement Adobe.

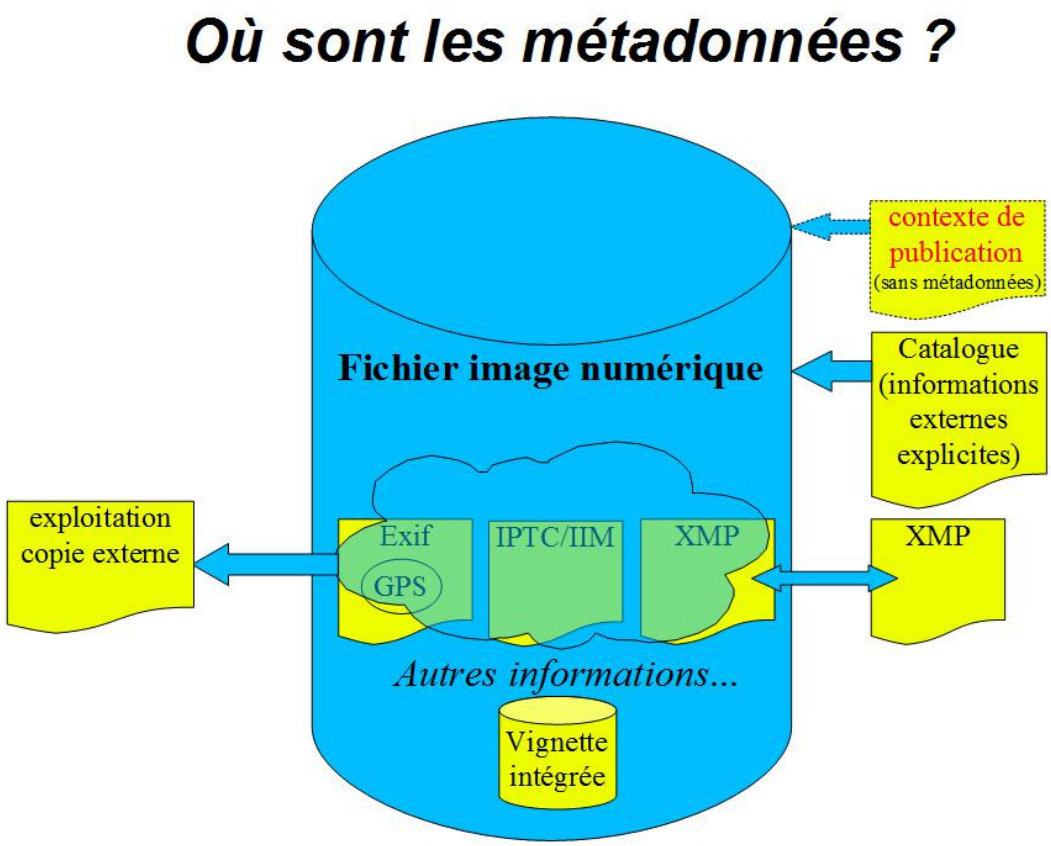

Image 1 - Où sont les métadonnées ?
Schéma des standards de métadonnées des images numériques fixes.

En résumé, les informations EXIF (et son sous-ensemble GPS) et les informations IPTC/IIM des images numériques fixes sont des métadonnées internes. Les métadonnées XMP quant à elles peuvent être soit internes, soit externes.
Une remarque: les métadonnées ne sont pas toujours sous une forme textuelle. Très souvent par exemple, une image peut contenir une vignette, une image de faible résolution, que l'on peut considérer comme une métadonnée.
Nous reviendrons plus loin sur ce qui est appelé ici le *contexte de publication*.

Les métadonnées internes présentent de nombreux avantages. En utilisant cette technique, l'échange est facilité car la ressource numérique transporte avec elle ses propres métadonnées lorsqu'elle est téléchargée, copiée, renommée, compactée, etc.

Mais les métadonnées internes présentent aussi certains inconvénients : il est nécessaire d'extraire les informations et de les copier dans une base de données pour exploiter une grande collection de ressources numériques. De la même façon qu'avec une légende inscrite au dos d'une photo, l'accès aux informations n'est pas direct, l'utilisateur doit effectuer une opération pour prendre connaissance de la légende.

Revenons sur le *contexte de publication*.
Google Images indexe seulement le nom du fichier image et le texte qui encadre l'image dans la page où elle apparaît, c'est ce qui est appelé ici le *contexte de publication* d'une image.
Pour un moteur de recherche généraliste en effet, les images qui contiennent des métadonnées constituent une partie insignifiante des images du web.
Google Images n'indexe donc pas les métadonnées internes des images (IPTC/IIM ou XMP).
Il suffit pour s'en assurer d'effectuer un test en indexant une image avec un mot-clé "hapax", c'est-à-dire un pseudo mot-clé unique, à la fois en IPTC/IIM et en XMP.

On peut se demander cependant si la présence de métadonnées internes aux images améliore le positionnement dans les résultats de recherche de *Google Images* ? Il s'agit là d'une interrogation récurrente en *SEO* (*Search Engine Optimization / Optimisation pour les moteurs de recherche*).
Les tests effectués à différentes périodes montrent que *non*, les métadonnées n'améliorent pas le positionnement des images lors d'une recherche.
Mais *Google* a probablement des projets internes qui exploitent les métadonnées internes aux images (nous n'en savons rien...)

2. Les médias sociaux et les métadonnées internes aux images fixes

Le Manifeste "*Embedded Metadata*" (métadonnées embarquées/intégrées) de l'*IPTC PhotoMetadata Working Group* définit cinq principes directeurs pour la création et le stockage des métadonnées, afin qu'elles soient transportées avec le fichier chaque fois que c'est possible.
Le Manifeste affirme que les métadonnées associées à une image doivent être persistantes dans toutes les étapes du flux des informations (*workflow*) - y compris dans tout affichage public.
Le Manifeste s'adresse à tous les organismes qui ajoutent et gèrent des métadonnées ainsi qu'aux fournisseurs de matériels et de logiciels dont les systèmes exploitent des flux de données.

Le Manifeste a conduit récemment une enquête publique intitulée *How Social Media sites manage metadata ?*
La méthodologie et les résultats de l'enquête sont décrits sur le site du Manifeste à l'adresse suivante:
<http://www.embeddedmetadata.org/social-media-test-results.php>

La méthode de l'enquête est simple. Une image de test possédant un jeu complet de métadonnées (EXIF, IPTC/IIM, XMP) est téléchargée sur différents réseaux sociaux. On observe ensuite les métadonnées de ces images stockées sur ces plate-formes.
Le détail du protocole est décrit sur le site du Manifeste à l'adresse suivante:
<http://www.embeddedmetadata.org/social-media-test-procedure.php>

Deux séries de tests ont été effectués en 2013 puis en 2015, afin de prendre en compte les nouveaux réseaux sociaux et observer les évolutions éventuelles entre les deux dates.

Les deux objectifs principaux étaient d'analyser précisément quelles sont les métadonnées embarquées qui s'affichent sur chaque plate-forme de réseau social, et vérifier les métadonnées préservées et celles qui sont altérées ou supprimées.

Quatre tests précis ont été définis:

- Quelles sont les métadonnées embarquées qui s'affichent dans l'interface utilisateur ?
- Les informations de crédit sont-elles correctement affichées ?

(vérification des "4C" : *Caption, Creator, Copyright Notice, Creditline*)

- Quelles sont les métadonnées préservées lorsque l'on récupère l'image depuis un navigateur, à l'aide d'une commande du genre *Save As* [*Sauvegarder sous*] ?
- Quelles sont les métadonnées préservées lorsque le réseau social propose un téléchargement de l'image (à l'aide d'un bouton *Download* par exemple) ?

Les résultats ne sont pas très brillants!
Les réseaux sociaux les plus connus altèrent les métadonnées embarquées d'une manière ou d'une autre. On peut même estimer que la situation se détériore globalement car les résultats étaient un peu meilleurs en 2013.
Examinons maintenant quelques résultats pour différentes plate-formes connues.

Dropbox

- Aucune métadonnée n'est affichée
- Les métadonnées sont préservées uniquement lors d'un *download*, elles ne sont pas préservées avec un *Save As*
- Dégradation: en 2013, elles étaient préservées avec un *Save As*

Facebook

- Aucune métadonnée n'est affichée
- Seules les métadonnées *Copyright Notice* et *Creator* de l'IPTC/IIM sont préservées avec un *Save As*. Toutes les autres sont supprimées.
- Légère amélioration depuis 2013: toutes les métadonnées étaient alors supprimées avec un *Save As*

On constate aussi sur *Facebook* une curiosité intrigante. La plate-forme ajoute systématiquement deux métadonnées IPTC/IIM dans les champs *Special Instructions* et *Original Transmission Reference*.
Il est extrêmement difficile de comprendre à quoi correspondent ces codes générés lors du téléchargement d'une image sur *Facebook*. Il n'existe en effet aucune communication de la société sur ce sujet et le *reverse engineering* est impuissant ici pour déterminer la signification de ces informations. Par ailleurs, l'IPTC ignore tout de cette particularité des images ayant transité par *Facebook*.
Grâce à ce dispositif, *Facebook* est peut-être capable d'effectuer un suivi élémentaire des images qui ont transité sur la plate-forme.

Flickr

- Quelques métadonnées sont affichées correctement, mais pas toutes les "4C"
- Toutes les métadonnées sont préservées lors d'un *download* ou un *Save As* de l'image

dans sa définition originale, mais elles sont supprimées dans les autres définitions
- Dégradation: vers 2010, toutes les résolutions intermédiaires proposées par la plate-forme possédaient les métadonnées de l'image originale

Google Photos
- Quelques métadonnées sont affichées correctement, mais pas toutes les "4C"
- Les métadonnées sont préservées lors d'un *download* de l'image originale
- Seules les métadonnées Exif sont préservées avec un *Save A*s sur les images en résolution réduite
- Dégradation: en 2013, toutes les métadonnées étaient préservées avec un *Save As* sur les images en résolution réduite

Instagram
- Aucune métadonnée n'est affichée
- Aucune sauvegarde n'est possible
- En 2013, le *Save As* était possible mais supprimait les métadonnées
- *Instagram* est l'un des pires réseaux sociaux au regard des métadonnées internes des images fixes

Pinterest
- Aucune métadonnée n'est affichée
- Les métadonnées sont préservées avec un *Save As* de l'image dans sa définition originale, mais pas lors d'un *download*
- Non testé en 2013

Tumblr
- Aucune métadonnée n'est affichée
- Seules les métadonnées Exif sont préservées avec un *Save As*, toutes les autres sont supprimées
- Dégradation: en 2013, toutes les métadonnées embarquées étaient préservées avec un *Save As*

Twitter
- Aucune métadonnée n'est affichée
- Seules les images en résolution réduite sont disponibles avec un Save As, sans aucune métadonnées
- Inchangé depuis 2013
- *Twitter* est lanterne rouge avec *Instagram*

Ces mauvais résultats ne sont pas une fatalité.
Il est tout à fait possible de concevoir des réseaux sociaux respectueux des métadonnées embarquées, à l'exemple de *Behance*, plate-forme appartenant à Adobe qui regroupe des porte-folios d'artistes (lien <www.behance.net>).

Behance
- Toutes les métadonnées 4C sont correctement affichées

- Plusieurs autres métadonnées (mais pas toutes) sont également affichées
- Toutes les métadonnées sont préservées lors d'un *download* et avec un *Save As*

3. Travailler avec des métadonnées internes sur un média social - le projet *PhotosNormandie* sur la plate-forme *Flickr*

Malgré ces résultats peu encourageants, il est possible de travailler efficacement sur un réseau social "médiocre" avec les métadonnées internes des images.
Le projet *PhotosNormandie* sur Flickr a pour objectif d'améliorer la description documentaire d'un fonds de plus de 3400 photographies historiques sur la bataille de Normandie (du Débarquement du 6 juin 1944 à fin août 1944).
C'est un projet de *crowdsourcing* (contenus générés par les utilisateurs), actif sur la plate-forme grand public Flickr depuis janvier 2007 à l'adresse suivante:
<www.flickr.com/photos/photosnormandie/>.
Il est ouvert à tous. Une soixantaine de contributeurs au total ont travaillé sur *PhotosNormandie* et une dizaine d'intervenants participent régulièrement au projet.

Quelle est l'origine des photos du projet *PhotosNormandie* ?
- 2760 photos proviennent du site *Archives Normandie 1939-1945* qui n'existe plus aujourd'hui; c'était un Service public du *Conseil Régional de Basse-Normandie* mis en place en 2004 ; toutes ces photos sont issues des Archives Nationales des États-Unis et du Canada et sont "libres de droit" (au sens américain)
- 296 photos de *The Allison Collection*; il s'agit de photos transmises par radio en 1944
- 322 photos proviennent de la bibliothèque de la ville de Cherbourg-Octeville
- 163 photos proviennent de la Médiathèque de Lisieux

Les légendes des photos sont écrites selon les standards de métadonnées IPTC/IIM et XMP. Plusieurs champs textuels de base (Titre, Légende, Mots-clés, Copyright, etc.) sont stockés dans le fichier image. La méthode s'appuie sur une fonctionnalité peu connue de la plate-forme Flickr: le renseignement automatique de champs Flickr à partir des champs IPTC lors du téléchargement d'une photo.

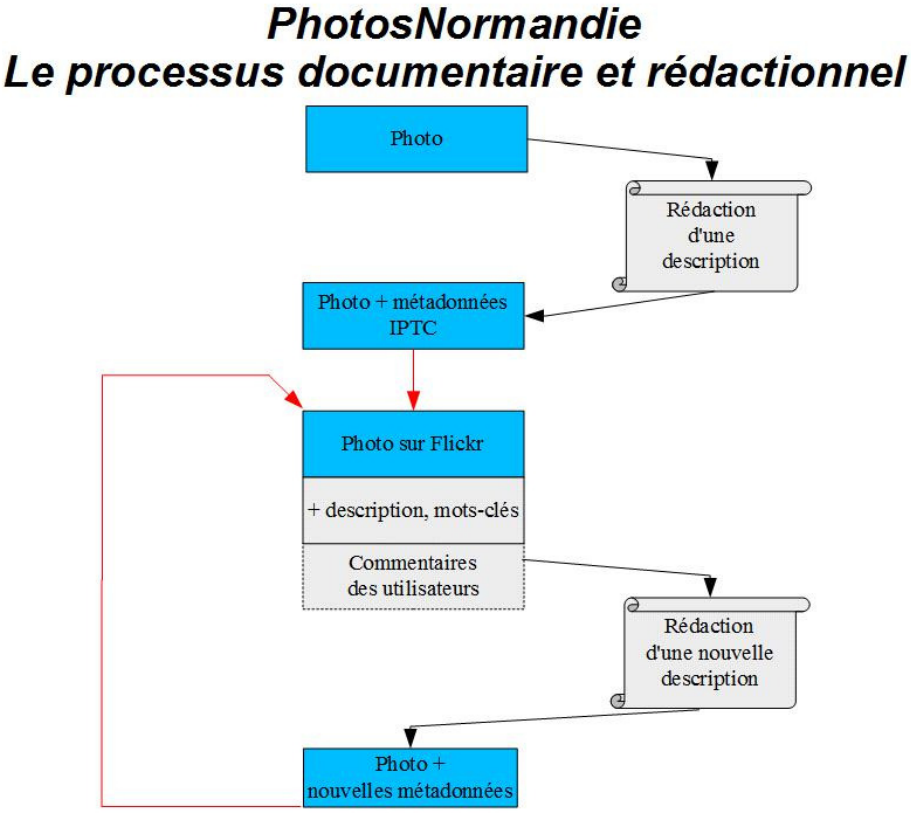

Image 2 - Description du processus documentaire et rédactionnel dans le projet PhotosNormandie

L'utilisation systématique des standards IPTC/IIM et XMP dans le projet *PhotosNormandie* concrétise les avantages des métadonnées internes. La description textuelle de l'image est toujours disponible avec l'image et facilement réutilisable.L'utilisateur reste libre de la technologie de base de données utilisée pour l'exploitation de son corpus d'images.
On peut résumer ce point de vue ainsi: les métadonnées sont comme les images, elles vous appartiennent, elles n'appartiennent pas à votre prestataire de service.
Pour le projet *PhotosNormandie* par exemple, nous pouvons facilement quitter Flickr et poursuivre le projet sur une autre plate-forme supportant un jeu réduit de métadonnées embarquées (comme 500px, fotki, ipernity [clone de Flickr], joomeo, smugmug).

La méthode utilisée sur Flickr présente néanmoins quelques inconvénients. Ainsi, la mise à jour d'une description est lourde puisqu'il est nécessaire de télécharger à nouveau chaque photo contenant une nouvelle description. Mais surtout, une URL Flickr pointant sur une photo ne peut être considérée comme stable car le numéro d'identification Flickr change lorsque l'on télécharge à nouveau l'image. Cependant, l'accès direct à chaque photo demeure possible car il est possible de travailler avec les références (ID) des photos, et les avantages de la méthode compensent très largement ses inconvénients.

Le bilan documentaire du projet *PhotosNormandie* est extrêmement positif:

- depuis fin janvier 2007, la galerie et les photos ont été vues plus de 36 millions de fois (soit plus de 7700 visites par jour; le 6 juin 2014, près de 200 000 visites ont été comptabilisées)
- on constate une grande progression depuis quatre ans environ
(4500 visites quotidiennes en 2012)
- 9222 descriptions corrigées et mises à jour (certaines descriptions ont été corrigées plusieurs fois)
- 417 photos correspondent à des séquences filmées retrouvées (plus de 1 sur 10; il doit en exister davantage)